AIMS Mathematics, 8(10): 22984–23002.
DOI: 10.3934/math.20231169
Received: 27 April 2023
Revised: 28 June 2023
Accepted: 03 July 2023
Published: 19 July 2023

http://www.aimspress.com/journal/Math

Research article

Electroencephalogram based face emotion recognition using multimodal fusion and 1-D convolution neural network (ID-CNN) classifier

Youseef Alotaibi[1,*] and Veera Ankalu. Vuyyuru[2]

[1] Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Makkah, 21955, Saudi Arabia
[2] Department of computer science and engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, 522502, A.P, India

* **Correspondence:** Email: yaotaibi@uqu.edu.sa.

Abstract: Recently, there has been increased interest in emotion recognition. It is widely utilised in many industries, including healthcare, education and human-computer interaction (HCI). Different emotions are frequently recognised using characteristics of human emotion. Multimodal emotion identification based on the fusion of several features is currently the subject of increasing amounts of research. In order to obtain a superior classification performance, this work offers a deep learning model for multimodal emotion identification based on the fusion of electroencephalogram (EEG) signals and facial expressions. First, the face features from the facial expressions are extracted using a pre-trained convolution neural network (CNN). In this article, we employ CNNs to acquire spatial features from the original EEG signals. These CNNs use both regional and global convolution kernels to learn the characteristics of the left and right hemisphere channels as well as all EEG channels. Exponential canonical correlation analysis (ECCA) is used to combine highly correlated data from facial video frames and EEG after extraction. The 1-D CNN classifier uses these combined features to identify emotions. In order to assess the effectiveness of the suggested model, this research ran tests on the DEAP dataset. It is found that Multi_Modal_1D-CNN achieves 98.9% of accuracy, 93.2% of precision, 89.3% of recall, 94.23% of F1-score and 7sec of processing time.

Keywords: electroencephalogram (EEG); emotion recognition; CNN; feature fusion; pre-processing; canonical correlation

1. Introduction

Daily lives depend heavily on emotion detection which enables software applications to respond in ways that consider the user's emotional states [1]. The use of emotion recognition can be seen in many different fields including communication skills, health monitoring and the monitoring and prediction of fatigue condition. Different levels of modalities are influenced by emotion recognition. The most common ways that emotions are displayed externally are through the visual, verbal, gestural, biological, electroencephalogram (EEG) and temperature of the body [2,3]. Due to the ease with which their datasets may be created, voice and visual are utilised most frequently in emotion recognition [4].

Recent research has focused on emotion recognition in unimodal modalities like text, audio and images. Even though unimodal emotion identification has achieved several ground-breaking advances over the years, it still has certain issues [5]. Poor accuracy results from the usage of unimodality which is unable to accurately express a particular emotion that the user is experiencing at the time. Therefore, it will be more thorough and detailed to use multimodal qualities to characterize a certain emotion. Emotion recognition accuracy is improved by multimodality [6,7]. However, as most models call for the use of stored information for offsite implementation, the results indicate some difficulties in integrating video, audio and EEG emotion recognition at the same time. There are three types of affect modelling: category, dimensional and component. Six fundamental emotions including joy, sadness, fear, anger, disgust and surprise are categorized into separate categories using categorized models which are simple to explain. Hierarchical models have been widely used in impact research due to their low complexity. The elucidation of additional complicated and difficult feelings is possible with volumetric models which portray emotion as a point in multidimensional space with the dimension's valence, induction and control [8].

Researchers have mainly relied on the streamlined two-dimensional concept of valence and arousal that was suggested in [9] where arousal varies in terms of feelings from calm to thrilled and valence-type fluctuates between uncomfortable to appealing [10]. Yet, such a multidimensional environment may present an important obstacle to effortless emotion recognition systems. A pair of fundamental emotions can be combined to create complicated feelings according to the component model of emotions which also arranges emotions in a hierarchical form. The eight fundamental bipolar emotions in Plutchik's [11] most well-known component model is based on evolutionary theory. Facial expression, body language, vocal tonality, gesticulation and a variety of physiological signals including heart rate, perspiration production, pupil size and brain activity to name a few can all be used to convey affect.

Numerous researchers have looked at the issue of identifying emotions from facial expressions in movies and still photos. Deep learning methodology advancements have generated a great deal of interest in using these techniques for facial emotion recognition (FER) [12,13], the majority of which are based on supervised learning. The techniques are unsuitable for online applications because they do not permit progressive, adaptive learning on new data. The reader is referred to [14] and the references therein for a fantastic review of the use of deep learning and superficial methods of learning to FER. Given this, the following are the outcomes of this work:

- By deleting unnecessary waves from EEG data and pixels in video clips, we suggested an effective and compact multimodal emotion identification model that utilises 2 methods i.e., facial footage and EEG data.
- A selection of footage that correlate to four distinct valance-arousal emotional spaces

happiness, neutrality, sorrow and fear serve as the basis for the stimulants. A neural network classifier uses 4 fundamental emotion levels to identify face expressions.

- Employing convolution neural networks (CNNs) with both global and local convolution kernels, we retrieve spatial features of left and right cerebral avenues as well as all EEG channels using the initial EEG signals. To combine closely related data gathered from EEG and face video recordings, researchers apply exponential canonical correlation analysis (ECCA).
- The 1D-CNN model was created primarily to simplify the complexity while transforming its characteristics into more understandable terms for analysis of correlations. Numerous tests were run on one public dataset to validate the suggested strategy.

The rest of this paper is ordered as follows: Section 2 mentions a few existing research works. Section 3 shows the proposed approach and methodologies. Section 4 exhibits the experimental outcomes and discussion. Finally, Section 5 includes the conclusion and speculation on future work.

2. Related works

In general, most work on electroencephalogram-based face emotion recognition using multimodal features deep learning method are similar to our work. In [15], the authors suggested a multi-modal technique that uses face video data and EEG recording to depict the reaction to emotional cues. The experimental finding demonstrates 97.5% accuracy in recognising facial emotions and classifying them into excitation (class 0) and expressiveness (class 1), exceeding state-of-the-art for the DEAP dataset. In order to enhance the model's generalizability for many themes, the authors in [16] offered a unique attention mechanism-based multi-scale feature fusion network (AM-MSFFN) that takes high-level features into account at various scales.

To extract sequential temporal and geographical information from EEG signals, we first use a spatial-temporal convolutional block. The signify-value approach was used in [17] to identify individual differences and derive each participant's categorization threshold using electroencephalogram and periphery physiologic information. An approach for learning shared cross-domain latent representations of the multi-modal data was proposed in [18] as the multi-modal domain adaptive variational autoencoder (MMDA-VAE).

The differential entropy (DE) features based on EEG data were extracted in [19] and then transformed into EEG topographic maps (ETM). The multichannel fusion approach was then utilised to combine the ETM and facial expressions. For the categorization of subject-dependent emotions, the deep learning classifier CBAM_ResNet34 utilised residual network (ResNet) and convolutional block attention module (CBAM). A novel method of multi-modal emotion identification was presented in [20]. The method creates a multi-level convolutional neural network (CNN) model for facial expression emotion recognition based on the modal information of facial expression. A stacked bidirectional LSTM (Bi-LSTM) model for emotion recognition is created using electroencephalography (EEG) information modes.

In [21], a multimodal attention network using bilinear pooling based on low-rank decomposition is suggested to determine the attention weights of facial video characteristics and the related EEG features in the fusion. Finally, attention weights and the outputs from the two-modality network are used to calculate the continuous sphere polarity values. The 3D-CNN is employed in [22] to obtain the EEG signal's final predictions. For the face approach, the precise facial pixels with emotional information are first extracted using the mask-RCNN object detection technique in conjunction with

openCV modules. The 3D-CNN output characteristics of the face chunks are then classified using the support vector machine (SVM) classifier.

Based on a global to local feature aggregation network (GLFANet), the authors of [23] suggested an EEG emotion identification method. This program creates an undirected topological network to represent the spatial connection link between channels by first using the spatial location of the EEG signal channels and the frequency domain properties of each channel. The emotion recognition of EEG and face is made possible in [24] by the spatio-temporal neural network and the separable residual network developed by fusion. In order to identify different emotions more accurately, the authors in [25] offered an EEG emotion detection model based on the attention mechanism and a pre-trained convolution capsule network. This concept uses coordinate attention to add relative spatial information to the input signal before mapping the EEG signal to a higher-dimensional space which enhances the EEG's ability to record emotional information. A unique approach to multi-task learning with a capsule network (CapsNet) and an attention mechanism is proposed in [26] as a basis for EEG-based emotion identification. Finally, in order to extract crucial information, the attention mechanism might alter the weight of various channels [27].

The techniques covered in the literature supported the use of facial video clips and EEG for emotion identification. Nevertheless, EEG data were quite delicate and are dampened by inferior implants. If the individual's inner state of mind differs from its gestures, the subject's gestures were inadequate for a fair judgement. Exterior achievement, on the other hand, is merely one way to portray feeling and is unable to capture the full range of a person's feelings [28–30]. The neurological system of the body has an impact on physiological differences which can better reflect an individual's emotional state. As a result, a distinctive study trend is emerging among scientists all over the world: the integration of physiologic and quasi-physiological data for the detection of emotion. It is effective to combine facial video clips and EEG signals for bimodal emotion recognition because they have both been extensively investigated in non-physiological and physiological contexts. As a result, the matching information can improve the objectivity and accuracy of emotion recognition because to this cooperative connection.

In most of their trials, the researchers used full-channel EEG signals to gather EEG signal data. The precision of experimental results is impacted by the whole channel signal, which is not favorable to experimentation. Channel selection technology is a current research hotspot since it is unknown which channels will be able to represent changes in mood the best. Researchers have made some advances in using deep understanding to emotion recognition based on EEG signals in recent years with the rapid development of deep learning. There are countless models for mixed neural network emotion recognition.

3. Proposed methodology

This section outlines the general model architecture and examines the suggested methods for continual multimodal emotion recognition. Our methodology revolves around the execution of continuous synchronized multimodal emotion recognition for face and EEG. Figure 1 depicts the system model in broad strokes. The four stages of the system's model include pre-processing, feature extraction, fusion and emotion recognition model (A–D). The overall architecture's components are individually explained as below.

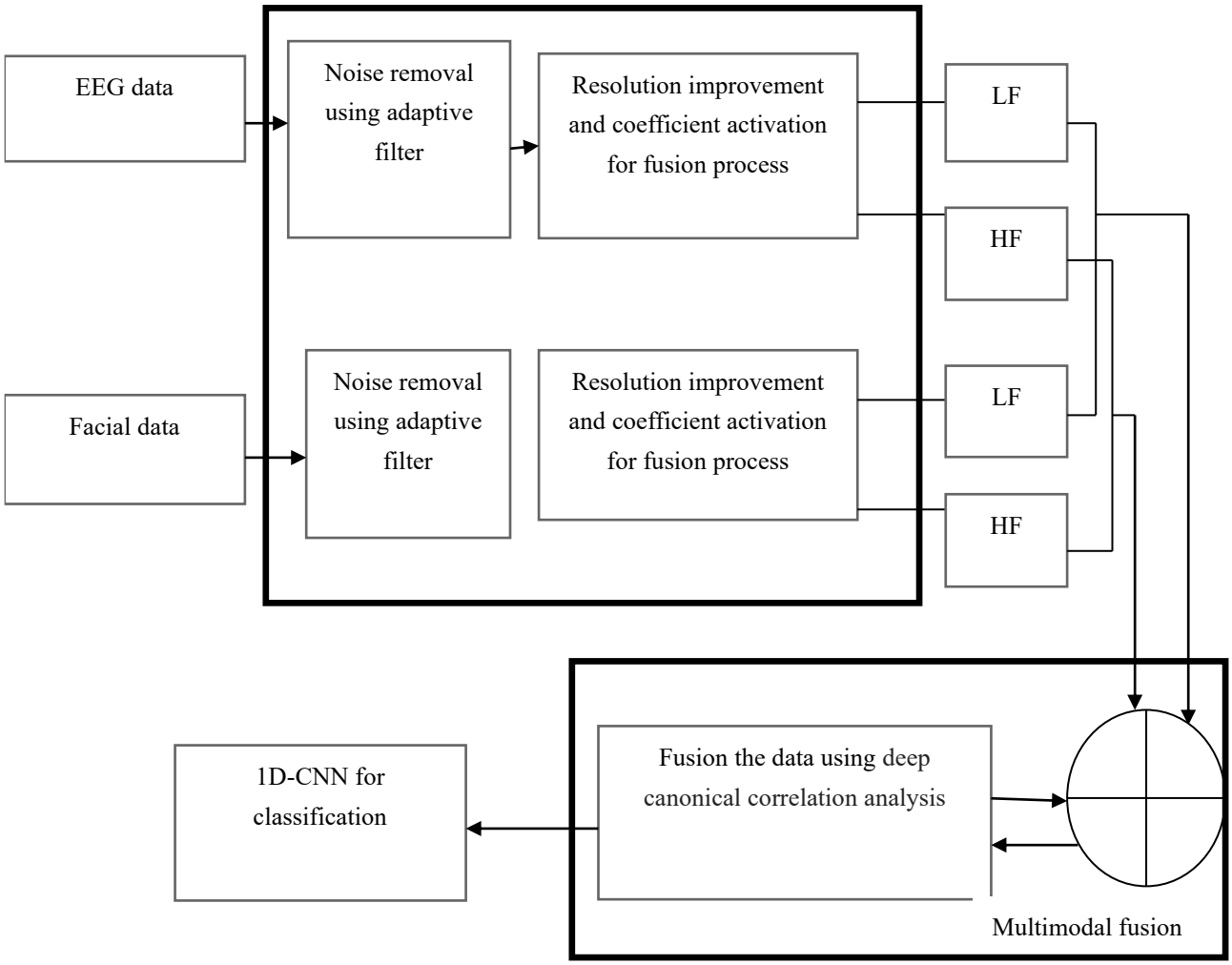

Figure 1. System architecture for multimodal fusion based facial emotion recognition.

3.1. Pre-processing using adaptive filters

The pre-processing step is an important element in the image and signal analysis schema [31,32]. It can enhance the original image and reduce noise or unwanted details.

i. Histogram equalization—this technique increases the pixel's brightness spectrum from the basic scale to 0 to 255. Therefore, the upgraded images have a broader variety of brightness and somewhat stronger contrast.

ii. The filter kernel size was empirically adjusted at 5 × 5 size. Hence Hist. eq. + Gaussian blur—this filter lowers certain noise and undesired details that can be confusing for the neural network.

iii. Hist. eq. + bilateral filter—this filter preserves edges while also reducing some noise and extraneous details that could confuse the neural network. The experimentally determined parameters of the filter are: width = 5, color = 1 and height = 75.

iv. Adaptive masking—In this suggested pre-processing technique, we first determined the maximum (max) and minimum (min) pixel intensities before applying binary thresholding with the threshold given by Eq 1. After that, morphologic closure was

applied. By doing this, the adaptive mask is created, and following bitwise operation, the starting image's aperture is removed.

v. Adaptive masking + hist. eq. + Gaussian blur—this method joins adaptive masking with histogram equalization and Gaussian blur.

$$threshold = min + 0.9.(max - \min). \tag{1}$$

3.2. Resolution improvement

The process of combining many low-resolution points of view (POVs) iteratively to create a higher-resolution image is referred to here as extremely-resolution. The initial estimate of the high-resolution image, $f^{(0)}$, in the Irani and Peleg definition of a high-resolution method can be based on the average of the amplified observations relocated to a common starting frame.

$$f^{(0)} = \frac{1}{K}\sum_{k=1}^{K} T_k^{-1}(gk \uparrow s) \tag{2}$$

where $\uparrow s$ is the up-sampling operator from the low-resolution to the high-resolution representation, T_k^{-1} is the geometry conversion to a common standard structure and gk is one of K captures. If the process of acquisition were sufficiently modelled, it would be possible to extract the low-resolution data collected gk from the "true" image f. The procedure would involve moving the image to the kth point of view, blurring it to compensate for the system's low resolution, down sampling it to that rate and adding noise. The low-resolution data is modelled for a certain estimate of the image, $f^{(n)}$:

$$g_k^{(n)} = (T_k\left(f^{(n)}\right) * \mathrm{h}) \downarrow \mathrm{s} \tag{3}$$

where $s \downarrow$ is the down sampling operator, which aggregates pixels belonging to the lesser resolution and $*$ h is the distortion operator with the gaussian h. The phrase "noise" is dropped. *Tk*. is the *k*th acquisition's initial geometric transformation from the common reference frame. The imager and the object have often moved physically from their initial positions. The distinction in the low-resolution data *gk* and the term g˜(n)k , which indicates whatever the low-resolution information should have been had the estimate $f^{(n)}$ been accurate, is used to correct the prior estimate of the high-resolution picture $f^{(n)}$in order to get a better estimate of the image f. The subsequent high-resolution estimation phase $f(n+1)$ is as follows:

$$f^{(n+1)} = f^{(n)} + \frac{1}{K}\sum_{k=1}^{K} T_k^{-1}(((gk - \tilde{\mathrm{g}}(\mathrm{n})\mathrm{k}))). \tag{4}$$

Here, the disparities among *gk* and g˜(n)k are summed over K captures, relocated to a common reference frame,T_k^{-1} and up-sampled to create the smaller super-resolution pixel size$\uparrow s$. Assume that $I_1(x,y)$ and $I_2(x,y)$ are two input resolution-improved pictures and that $W_1(x,y)$ and $W_1(x,y)$ are, respectively, the maximum, minimum and mean-maximum DTCWT coefficients.

Maximum selection: For high pass values, the highest possible coefficients are chosen and the mean of the low pass values is used as

$$W_{HF}(x,y) = \max\left(W_1(x,y), W_2(x,y)\right). \tag{5}$$

$$W_{LF}(x,y) = 0.5\ \left(W_1(x,y), W_2(x,y)\right). \tag{6}$$

Here, $W_{HF}$and W_{LF}indicates high pass and low pass filter. Mean selection-average is calculated using both the high pass and low pass factors.

$$W_{HF}(x,y) = 0.5\ \left(W_1(x,y), W_2(x,y)\right). \tag{7}$$

$$W_{LF}(x,y) = 0.5\ \left(W_1(x,y), W_2(x,y)\right). \tag{8}$$

High pass values are averaged while low pass values are used to choose the highest coefficients.

$$W_{HF}(x,y) = 0.5\ \left(W_1(x,y), W_2(x,y)\right). \tag{9}$$

$$W_{LF}(x,y) = \max\left(W_1(x,y), W_2(x,y)\right). \tag{10}$$

Utilizing fusion principles, the low-pass and high-pass coefficients for each slice have been blended. Each image simply chooses the absolute wavelet coefficients with location from the input images to serve as the coefficients at chosen place in the combined image. Before discussing feature extraction and classification, the concern notations and abbreviations are given in Table 1.

Table 1. Description of notations.

Notation	Abbreviation
H^{input}	Input vector
H^{output}	Output vector
H	Feature map
$\boldsymbol{\delta}(.)$	Activation function
l	Scaling level
msl	Maximum scaling level
tsl	Total scaling level
u_x and v_x	Variables u and v in x and y direction
f_g	Filter
m	Length of filter
$af(s')$	Activating function
c_q^d	Offset variable
h_d	Hidden unit

3.3. Feature extraction of EG and facial clips using CNN

The CNN architecture was picked for the suggested method's feature collection and emotion identification functions. The approach comprises of two unique models for various tasks. The initial CNN model is made to gather characteristics from facial video clips and EEG data. After the features are extracted, ECCA is used to perform feature-level fusion and the SoftMax layer is supplied by strongly associated characteristics for categorization. The scaling layer, a building element used to

adaptively extract useful data-driven spectrogram-like features from raw EEG signals, will be discussed first. The scaling layer-based convolutional neural network will then be introduced. We examine a multi-kernel convolutional layer that receives a one-dimensional input with shape sampling points as inputs and generates a two-dimensional spectrogram-like feature map as output while scaling levels using the following layer-wise transmission method.

$$H^{output}(l) = \delta(bias\ (l) + downsample(weignt, l) \times H^{input}). \tag{11}$$

The one-dimensional signal and $H^{input}$is the input vector with shape time steps. The output of H is a feature map that resembles a spectrogram in that it is a matrix of activations with shape time steps and scaling levels which is given by H^{output}. Scaling a fundamental kernel produces biases for several kernels. An activation function is indicated by $\boldsymbol{\delta}(.)$ where weight is the fundamental kernel from which all other kernels are scaled. The scaling level is controlled by the hyperparameter l. Down sample, a pooling operator, down samples the weight l time using an average filter and a window size of 2. This scales the data-driven pattern weight to a specific period in order to capture specific frequency-like representations from H^{input}.

Assume we wish to extract features for signal H^{input}at the l th scaling level. We first generate the l th scaling level kernel scaled from $weight$ by down sample $weight$. Then, we perform the cross-correlation operator of the scaled kernel and H^{input}. Then, we add the previous result and the $bias\ (l)$, and then feed the sum to the activation function $\boldsymbol{\delta}(.)$. We repeat the above process total scaling level tsl times with different setups of hyper-parameter l on a range of 0 to maximum scaling level msl where the maximum scaling level msl is the $l\,th$ level that makes the length of vector $downSample\ (weight, l)$equal to 1 and the total scaling level $tsl = msl + 1$. Finally, we stack all extracted feature vectors into a 2D tensor to obtain the data driven spectrogram-like feature map.

3.4. Feature level fusion using ECCA

In this study, we combined highly correlated variables from face video clips and EEG data using exponential canonical correlation analysis (ECCA). Initially, ECCA was designed to compute representations of various modalities by subjecting them to a number of nonlinear transformations in stacked layers. The first step in traditional CCA is typically a matrix to vector conversion which alters the space structure of the original data and leads to dimension disaster. Exponential canonical correlation analysis which processes the matrix directly was presented as a solution to these issues.

$I_x \times I_y \times I_z$are the dimensions of the x, y and z directions. Let $F \in r^{e^x} = R^{I_x \times I_y \times I_z}$signify a three-dimensional source data. Where $1 \le i \le I_x, 1 \le j \le I_y, 1 \le k \le I_z$, an element of F is denoted by $F_{i,j,k}$. The definition of the dot product of two three-dimensional variables is S $< S, T \ge \sum_{i,j,k}^{g} S_{i,j,k} T_{i,j,k}$. $\|F\| = \sqrt{< F, F >}$ is the formula that defines the norm of F . Three-dimensional data in each direction can be flattened into its corresponding vector space, just like a

matrix. The x direction flattened matrix is denoted as $F_{(x)} \in R^{I_x(I_y \times I_z)}$ I_x and $(I_y \times I_z)$ are the rows and columns respectively of $F_{(x)}$. The product of $F_{(x)}$ and matrix T is defined as $T^y F_{(x)}$ or $F_{(x)}T$. We construct objective function as follows:

$$Obj(1) = \log_{I_x \times I_y \times I_z} cov(F1_x u_{xy} u_{yz} u_{zx}, F2_y v_{xy} v_{yz} v_{zx}) \tag{12}$$

$$such\ that\ \left(F1_x u_{xy} u_{yz} u_{zx}\right) = 1\ and\ var\ \left(F2_y v_{xy} v_{yz} v_{zx}\right) = 1. \tag{13}$$

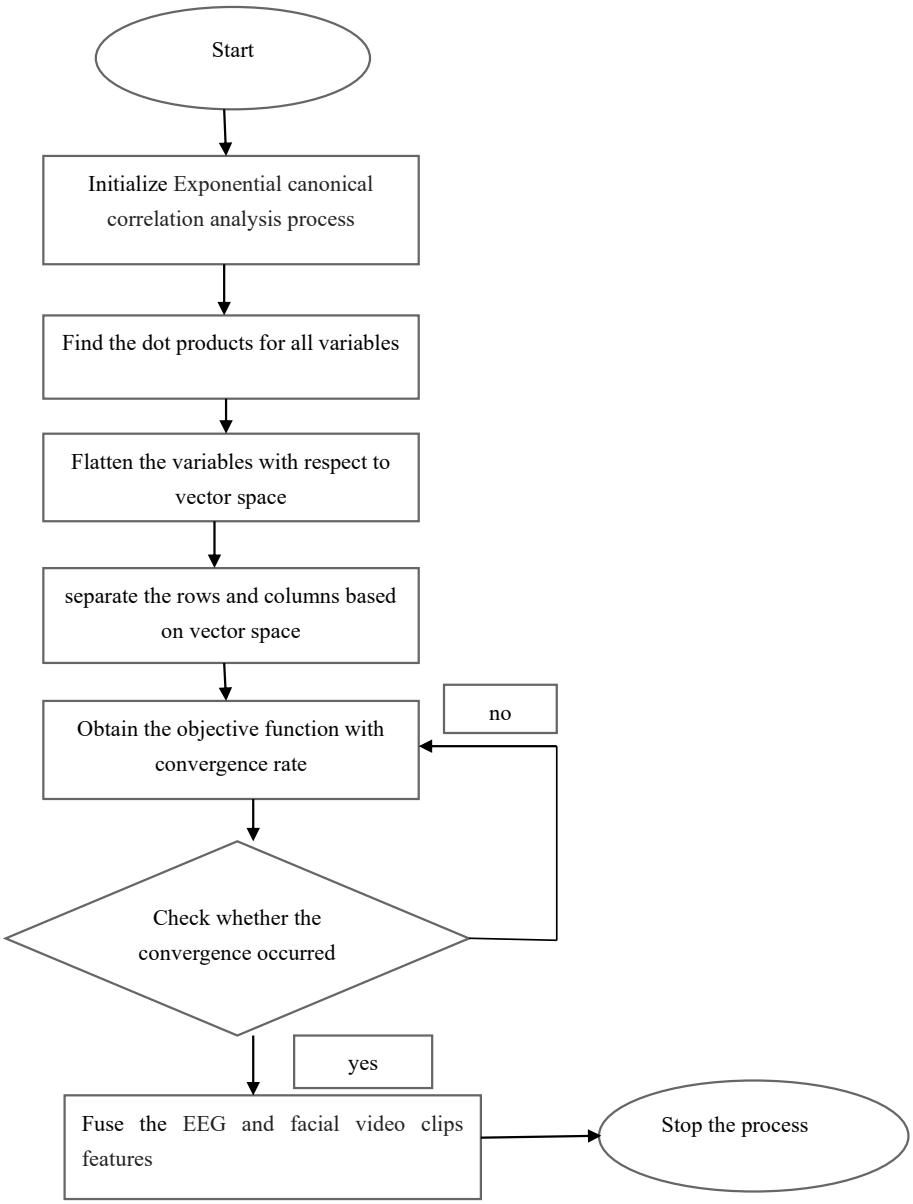

Figure 2. Flow chart for multimodal fusion process.

With the high-dimensional nonlinear optimization issue, the objective function (12) has nonlinear restrictions. The illustration of fusion process is shown in Figure 2.

Directly locating the closed form solution is challenging. To get the solution in this study, we employ an alternate numerical iterative method. We only talk about the x-direction-specific solution of the transforms u_x and v_x Additionally, the y, z transforms are same. In order to create the deep learning model that would be utilized in the ECCA method, we employed a grid search methodology to determine the best hyperparameters. We chose the regulation parameter to be $1e^5$, cross-entropy loss and a stochastic gradient descent optimizer after a few time-consuming experiments.

3.5. 1D-CNN classifier

To enhance the ECCA, the 1D model is created using CNN architecture. The model consists of three convolutional layers with a max-pooling layer coupled to each of them, three input layers, one dropout layer and ultimately a SoftMax layer. The convolutional layers are created using, respectively, 128, 256 and 512 convolution kernels with kernel sizes of 3, 5 and 3 with strides of 1. As an activation function, ReLU is a nonlinear function. The pool size and stride for each max-pooling layer are both 2. Designing the dropout layer uses the value 0.4. A group of neurons make up the layers. Every layer in this has connections to every layer's neuron. The prediction is represented by the last completely connected layers which also create the result layer.

Convolutional neural networks are regarded as the go-to solution because they are built to effectively map picture data to an output variable for any prediction issue utilizing attributes as an input. Each neuron's input, which is coupled to the local receptive field of the previous layer and tends to retrieve the local feature, is included in the categorization of the CNN structure. Since each convolutional layer reduces the number of input features to fully connected layers, the output improves as we add more convolutional layers, as seen in Figure 3.

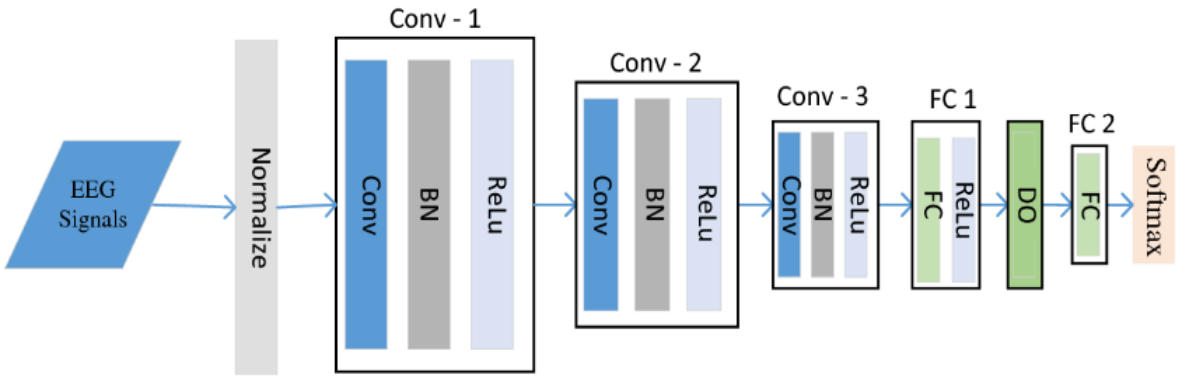

Figure 3. Architecture of ID-CNN for classification.

Consider the input signal sequence as s_i where $i = 1,2,3, \ldots . ni$ and the filter as f_g and $i = 1,2,3, \ldots . m$. Here, the length of the filter m must be greater than the length of the signal sequence n_i. Through a partial convolution operation, the filter is run based on the input features of the preceding layer. The 1d-CNN's convolved output, x_i, is expressed in

$$x_i = \sum_{g=1}^{m} f_g \times s_{i-g+1}. \tag{14}$$

The local connection network is established here by correlating every single neuron in the d^{th} layer with neurons in the $(d-1)^{th}$ layer of the local window. The activating function $af(s')$ in the layer of convolution performs the non-linear mapping. A revised linear unit in this 1D-CNN model utilises an activation function in order to speed up convergence with maximum of max (O, s').

$$af(s') = \max(0, s')\cdot \tag{15}$$

Moreover, the input of q^{th} neurons of the d^{th} layer is derived in Eqs 16 and 17. Furthermore, Eqs 16 and 17 derive the q^{th} input of the d^{th} layer of neuron

$$c_q^d = af(\sum_{r=1}^{m} f_r^d \times c_{q-r+m}^{d-1} + h_d) \tag{16}$$

$$= af(f^d \times c_{(q-m+1):q}^{d-1} + h_d). \tag{17}$$

The offset variables in the equation are known as c_q^d where $q = 1,2,3,\ldots.ni$ and the m^{th} dimension filter is known as $f_r^d \epsilon R^m$ which is constant across all neurons in the layer of convolution. Due to various benefits including easier learning and execution with fewer hidden layers, fewer computational hurdles with simple array operations and efficient operation in 1D EEG signals, 1D-cnns are widely used. Here, h_d indicates the hidden unit. The outputs of the first CNN layer or block's 1D convolution layer have been considered as inputs for the second layer or block of CNN. The result from the first CNN layer was subjected to BM in the subsequent CNN layer which also imposed the first Conv1D configuration. At the conclusion of the second CNN layer, a new sub-block called max pooling layer (MaxPooling1D) has been added. MaxPooling1D has been set up with a window that slides that has a height of three.

As a result, Conv1D layer has undergone various adjustments in the third CNN layer. An average pooling layer with the same sliding window height has been used in place of the max pooling layer where the Conv1D has been configured with 160 sliding windows (feature detection) with 5 kernel size or height. To prevent overfitting in the following stage, a dropout layer with a rate of 0.5 has been taken into consideration. In order to obtain the classification result, a dense layer that is fully linked and activated with a softmax has been employed to construct probabilistic distributions over the two classes.

3.6. Convolutional layer (Maxpool-1D)

The 1D max pooling layer decreases the feature map dimension by only keeping the maximum value of the feature map in a window patch with a predetermined pool size. The output feature maps (convolution outputs, c) created by the conv1D layers are given as an input to the layer. With the convolution process, the window is shifted and moved across the feature map as illustrated in the image. Max pooling's functionality c_h^l can be modelled as

$$c_h^l = max\ c_p^{l-1}.r_h. \tag{18}$$

$r_h$stands for the pooling area with index h in this case. The values of pool size and stride are used in this work as 2, respectively. The max pooling procedure is illustrated using the following parameters: $c_{m1} = \max(c_1, c_2)$; $c_{m2} = \max(c_3, c_4)$; $c_{m3} = \max(c_5, c_6)$.

3.7. Flatten layer and dropout

The input data are converted into a one-dimensional vector by the flatten layer and supplied to the dense/fully connected layer. After the flatten layer, a dropout parameter is added to help the design generalize better by minimizing over fitting during retraining. This is accomplished by varying some nodes' activations to zero, as determined by a dropout rate. A dropout rate of 0.25 has been employed in this study.

3.8. Dense layer for classification

The flattened output is fed into the dense/fully interconnected layer. Next layer creates the categorization result with dimension $M \times 1$; where M is the number of categories, as an inputs. The layering process is typically expressed as

$$output = \sigma(< input, w_d > + b_d) \tag{19}$$

where b_d stands for the bias vector for this layer, is the activation function and $< input, w_d >$denotes the dot product between the input and the weight vector $w_d$employed in this layer. For binary and multi-class classification in this study, we employ sigmoid and softmax activation, respectively. The sigmoid activation function is given by

$$\sigma(z) = \frac{1}{1+e^{-z}}. \tag{20}$$

The above function generates a binary output that represents the likelihood for a binary classification, depending on which category label is either "0" or "1". The function that activates softmax can additionally be expressed as

$$Softmax(z)_i = p_i = \frac{\exp(z)_i}{\sum_{j=1}^{m} \exp(z)_j} \tag{21}$$

where zi stands for the i-th member of the output vector from layer z before it. To place the value of pi between 0 and 1, the numerator is normalized by the total of all logarithmic terms from 1 to M. The categories labels for classes for multi-class classification are generated by this layer.

4. Experimental analysis

For both datasets, a three-class problem (happy, neutral and sad) is considered for testing the proposed model. We applied EEG and facial video clip data to detect three classes of emotions (happy,

neutral and sad) because these are the very basic categories of human emotions. A leave-one-subject-out strategy was used to conduct experiments for each dataset and the results were compared with three existing methods such as attention mechanism-based multi-scale feature fusion network (AM-MSFFN) [16], CBAM_ResNet34 [19], stacked bidirectional LSTM (Bi-LSTM) model [20], GLFANet [23] and CapsNet [26]. In an experiment where one participant is left out for testing, all subject's training data is included. Furthermore, from the data of N subjects, (N-1) subjects number of trials for every topic for train (90% training set, 10% validation) and 1 topic number of trails for each topic for test were done.

In this work, we executed offline experiments using DEAP datasets. 32 subjects' EEG, video and other ancillary physiological data are included in the DEAP [33] dataset. Participants were invited to watch 40 one-minute music videos while the information was being collected. Every video's frame size was set to 720 × 576 pixels at the frame rate of 50 frames per second during recording. Moreover, this dataset contains ratings of the levels of arousal and valence for each stimulus from each subject. Please take note that we only used 22 people for whom all 40 trials' worth of facial video clips and EEG data were accessible. Table 2 shows the results of fusion of EEG and facial video clips using ECCA in terms of accuracy.

Table 2. Results of fusion of EEG and facial video clips using ECCA in terms of accuracy.

Number of users	EEG data	Facial data	EEG+facial
User-1	88.5	91.2	97.4
User-2	87.4	90	99
User-3	85.7	90.4	98
User-4	87	93.2	98.2
User-5	84.5	92	98.4
User-6	83.5	92.5	99
User-7	87.3	94	98.3
User-8	86.2	94.1	98.4
User-9	87.3	93.6	98.7
User-10	87	92	98.8
Average accuracy (%)	**87.8**	**95.1**	**98.9**

The suggested Multi_Modal_1D-CNN technique, the existing AM-MSFFN, CBAM_ResNet34, Bi-LSTM, GLFANet and CapsNet methods are compared in Figure 4 where the X-axis indicates the various classes and the Y-axis displays the accuracy achieved in %. When analyzing, the existing methods achieves 92.4%, 94%, 97.3%, 95.4% and 96.4% of accuracy and the suggested Multi_Modal_1D-CNN achieves 98.9% which is 5.5%, 4.9%, 2.6%, .3.3% and 2.4% better than forementioned existing methods.

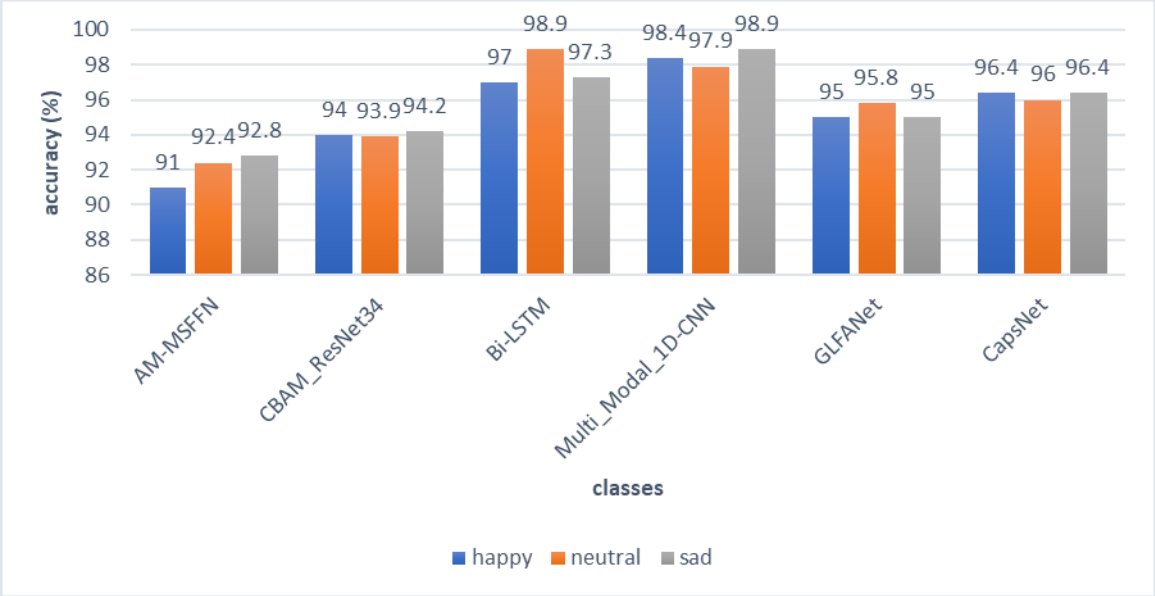

Figure 4. Comparison of accuracy.

The suggested Multi_Modal_1D-CNN technique, the existing AM-MSFFN, CBAM_ResNet34, Bi-LSTM, GLFANet and CapsNet methods are compared in Figure 5 where the X-axis indicates the various classes and the Y-axis displays the precision achieved in %. When analyzing, the existing methods achieves 89.3%, 87%, 81%, 82.2% and 85% of precision and the suggested Multi_Modal_1D-CNN achieves 93.2% which is 4.1%, 6.2%, 12.2%, 10.5% and 7.2% better than existing methods.

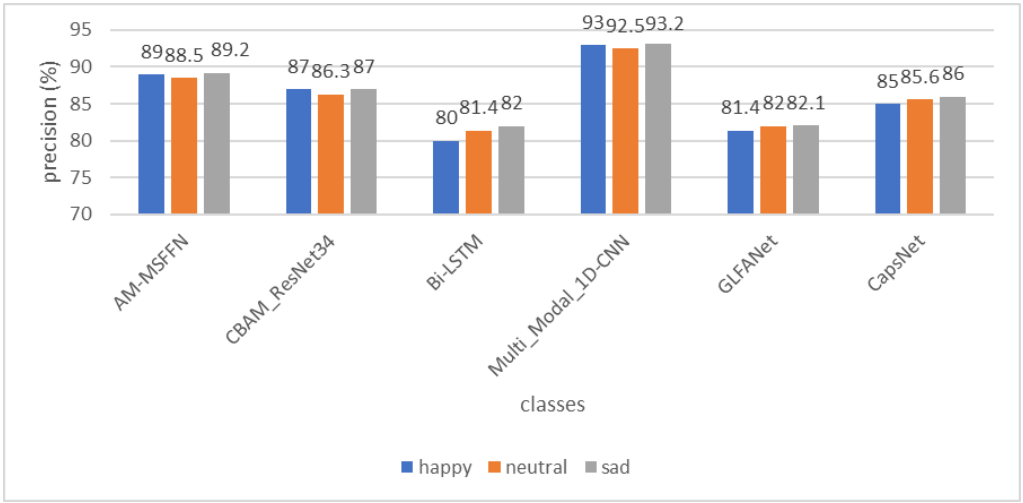

Figure 5. Comparison of precision.

The suggested Multi_Modal_1D-CNN technique, the existing AM-MSFFN, CBAM_ResNet34 Bi-LSTM, GLFANet and CapsNet are compared in Figure 6 where the X-axis indicates the various classes and the Y-axis displays the recall achieved in %. When analyzing, the existing methods achieves 78%, 77.3% ,78% ,77.5% and 78% of recall and the suggested Multi_Modal_1D-CNN achieves 89.3% which is 11.3%, 13%, 11.3% and 12.2% better than existing methods.

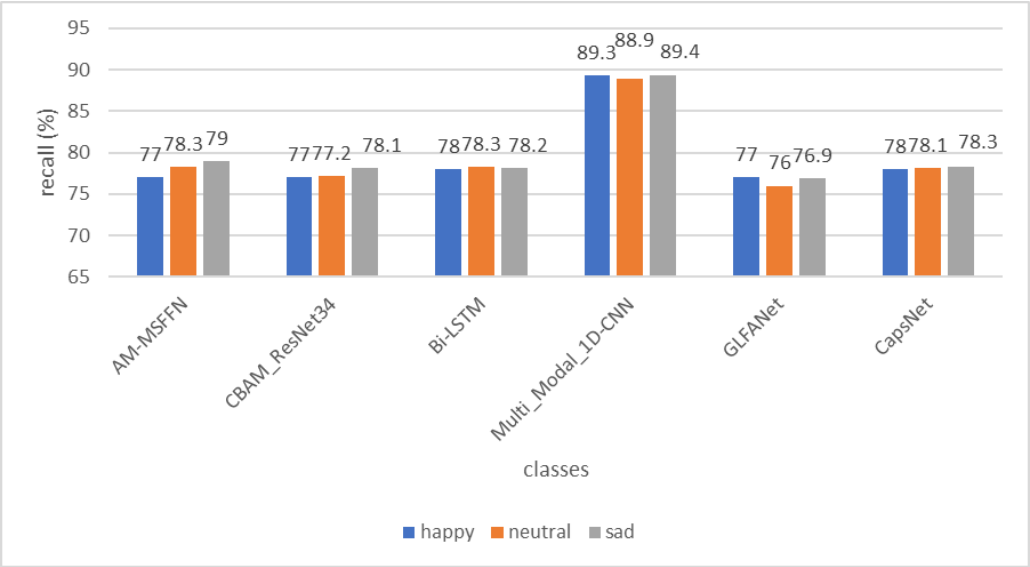

Figure 6. Comparison of recall.

The suggested Multi_Modal_1D-CNN technique, the existing AM-MSFFN, CBAM_ResNet34, Bi-LSTM, GLFANet and CapsNet methods are compared in Figure 7 where the X-axis indicates the various classes and the Y-axis displays the F1-score achieved in %. When analyzing, the existing methods achieves 82.4%, 81%, 87.5%, 91.2% and 92% of F1-score and the suggested Multi_Modal_1D-CNN achieves 94.23% which is 12.23%, 13.23%, 7.32%, 3.03% and 2.23% better than existing methods.

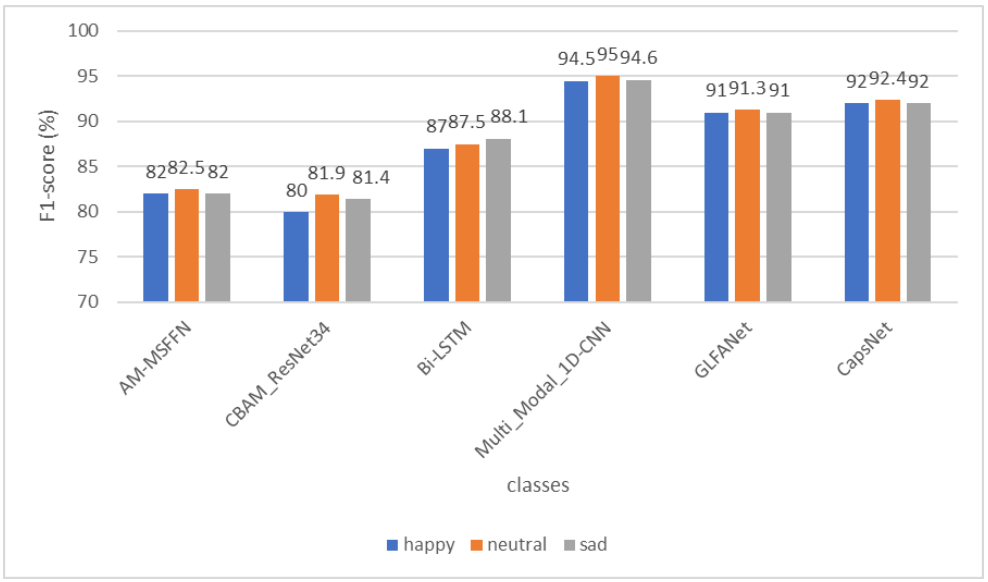

Figure 7. Comparison of F1-score.

The suggested Multi_Modal_1D-CNN technique, the existing AM-MSFFN, CBAM_ResNet34, Bi-LSTM, GLFANet and CapsNet methods are compared in Figure 8 where the X-axis indicates the various classes and the Y-axis displays the processing time achieved in %. When analyzing, the existing

methods achieves 12 sec, 14 sec, 11 sec, 13 sec and 12 sec of processing time and the suggested Multi_Modal_1D-CNN achieves 7 sec which is 5 sec, 7 sec, 4 sec, 6 sec and 5 sec better than existing methods. Table 3 summarizes all results.

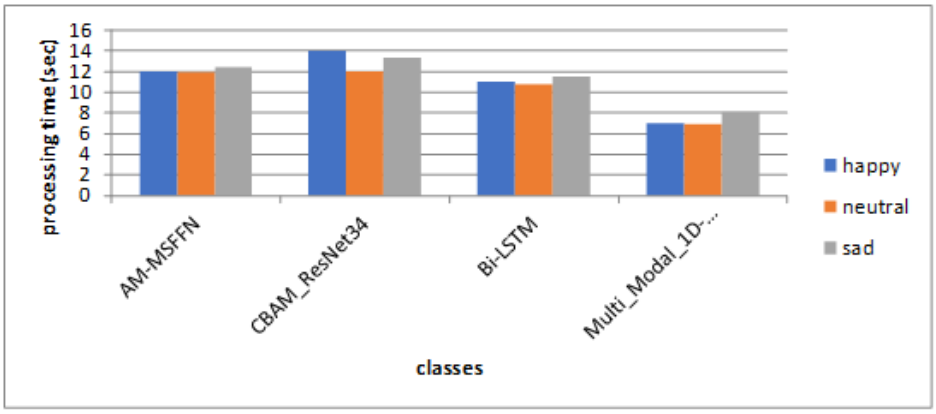

Figure 8. Comparison of processing time.

Table 3. Overall comparative analysis.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Processing time (sec)
AM-MSFFN	92.4	89.3	78	82.4	12
CBAM_ResNet34	94	87	77.3	81	14
Bi-LSTM	97.3	81	78	87.5	11
GLFANet	95.4	82.2	77.5	91.2	13
CapsNet	96	85	78	92	12
Multi_Modal_1D-CNN	98.9	93.2	89.3	94.23	7

5. Conclusions

This study suggests a multimodal emotion identification model based on the combination of facial expressions and EEG inputs. The features of EEG data and facial expressions can be directly extracted by the end-to-end model. Pre-trained CNN is utilized to extract facial features from facial expressions and exponential canonical correlation analysis is used to combine the features of key emotion images. The results of the experiments demonstrate that the proposed model can recognize emotions accurately and that utilizing EEG and facial expressions together has a stronger multimodal emotion detection effect than using either EEG or facial expressions separately. In order to extract facial expression features and cut down on the resources and operating time needed for the model, we will next investigate a more trustworthy pre-training model. To enhance the multimodal emotion recognition model, we will also aim to add other modalities such as non-physiological inputs.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors extend their appreciation to the Deanship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number: IFP22UQU4281768DSR159.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, *Biomed. Signal Process. Control*, **47** (2019), 312–323. https://doi.org/10.1016/j.bspc.2018.08.035
2. M. Liu, J. Tang, Audio and video bimodal emotion recognition in social networks based on improved alexnet network and attention mechanism, *J. Inf. Process. Syst.*, **17** (2021), 754–771. https://doi.org/10.3745/JIPS.02.0161
3. J. N. Njoku, A. C. Caliwag, W. Lim, S. Kim, H. Hwang, J. Jung, Deep learning based data fusion methods for multimodal emotion recognition, *J. Korean Inst. Commun. Inf. Sci.*, **47** (2022), 79–87. https://doi.org/10.7840/kics.2022.47.1.79
4. Q. Ji, Z. Zhu, P. Lan, Real-time nonintrusive monitoring and prediction of driver fatigue, *IEEE T. Veh. Techol.*, **53** (2004), 1052–1068. https://doi.org/10.1109/TVT.2004.830974
5. H. Zhao, Z. Wang, S. Qiu, J. Wang, F. Xu, Z. Wang, et al., Adaptive gait detection based on foot-mounted inertial sensors and multi-sensor fusion, *Inf. Fusion*, **52** (2019), 157–166. https://doi.org/10.1016/j.inffus.2019.03.002
6. J. Gratch, S. Marsella, Evaluating a computational model of emotion, *Auton. Agent. Multi-Agent Syst.*, **11** (2005), 23–43. https://doi.org/10.1007/s10458-005-1081-1
7. J. Edwards, H. J. Jackson, P. E. Pattison, Emotion recognition via facial expression and affective prosody in schizophrenia: A methodological review, *Clin. Psychol. Rev.*, **22** (2002), 789–832. https://doi.org/10.1016/S0272-7358(02)00130-7
8. T. Fong, I. Nourbakhsh, K. Dautenhahn, A survey of socially interactive robots, *Rob. Auton. Syst.*, **42** (2003), 143–166. https://doi.org/10.1016/S0921-8890(02)00372-X
9. J.A. Russell, A circumplex model of affect, *J. Per. Soc. Psychol.*, **39** (1980), 1161–1178. https://doi.org/10.1037/h0077714
10. H. Gunes, B. Schuller, M. Pantic, R. Cowie, Emotion representation, analysis and synthesis in continuous space: A survey, In: *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2011, 827–834. https://doi.org/10.1109/FG.2011.5771357

11. R. Plutchik, The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, *Am. Sci.*, **89** (2001), 344–350. http://www.jstor.org/stable/27857503
12. A. Gudi, H. E. Tasli, T. M. Den Uyl, A. Maroulis, Deep learning based facs action unit occurrence and intensity estimation, In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015, 1–5. https://doi.org/10.1109/FG.2015.7284873
13. R. T. Ionescu, M. Popescu, C. Grozea, Local learning to improve bag of visual words model for facial expression recognition, In: *ICML 2013 Workshop on Representation Learning*, 2013.
14. S. Li, W. Deng, Deep facial expression recognition: A survey, *IEEE T. Affect. Comput.*, **13** (2020), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446
15. S. Wang, J. Qu, Y. Zhang, Y. Zhang, Multimodal emotion recognition from EEG signals and facial expressions, *IEEE Access*, **11** (2023), 33061–33068. https://doi.org/10.1109/ACCESS.2023.3263670
16. Y. Jiang, S. Xie, X. Xie, Y. Cui, H. Tang, Emotion recognition via multi-scale feature fusion network and attention mechanism, *IEEE Sens. J.*, **10** (2023), 10790–10800. https://doi.org/10.1109/JSEN.2023.3265688
17. Q. Zhang, H. Zhang, K. Zhou, L. Zhang, Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (PMD) for emotion recognition, *Tsinghua. Sci. Techol.*, **28** (2023), 673–685. https://doi.org/10.26599/TST.2022.9010038
18. Y. Wang, S. Qiu, D. Li, C. Du, B. L. Lu, H. He, Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition, *IEEE/CAA J. Autom. Sinica*, **9** (2022), 1612–1626. https://doi.org/10.1109/JAS.2022.105515
19. D. Li, J. Liu, Y. Yang, F. Hou, H. Song, Y. Song, et al., Emotion recognition of subjects with hearing impairment based on fusion of facial expression and EEG topographic map, *IEEE T. Neur. Syst. Reh.*, **31** (2022), 437–445. https://doi.org/10.1109/TNSRE.2022.3225948
20. Y. Wu, J. Li, Multi-modal emotion identification fusing facial expression and EEG, *Multimed. Tools Appl.*, **82** (2023), 10901–10919. https://doi.org/10.1007/s11042-022-13711-4
21. D. Y. Choi, D. H. Kim, B. C. Song, Multimodal attention network for continuous-time emotion recognition using video and EEG signals, *IEEE Access*, **8** (2020), 203814–203826. https://doi.org/10.1109/ACCESS.2020.3036877
22. E. S. Salama, R. A. El-Khoribi, M. E. Shoman, M. A. W. Shalaby, A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition, *Egypt. Inf. J.*, **22** (2021), 167–176. https://doi.org/10.1016/j.eij.2020.07.005
23. S. Liu, Y. Zhao, Y. An, J. Zhao, S. H. Wang, J. Yan, GLFANet: A global to local feature aggregation network for EEG emotion recognition, *Bio. Signal. Process. Control*, **85** (2023), 104799. https://doi.org/10.1016/j.bspc.2023.104799
24. Y. Hu, F. Wang, Multi-modal emotion recognition combining face image and EEG signal, *J. Circuit. Syst. Comput.*, **32** (2022), 2350125. https://doi.org/10.1142/S0218126623501256
25. S. Liu, Z. Wang, Y. An, J. Zhao, Y. Zhao, Y. D. Zhang, EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network, *Knowl. Based Syst.*, **265** (2023), 110372. https://doi.org/10.1016/j.knosys.2023.110372

26. C. Li, B. Wang, S. Zhang, Y. Liu, R. Song, J. Cheng, et al., Emotion recognition from EEG based on multi-task learning with capsule network and attention mechanism, *Comput. Bio. Med.*, **143** (2022), 105303. https://doi.org/10.1016/j.compbiomed.2022.105303
27. S. J. Savitha, M. Paulraj, K. Saranya, Emotional classification using EEG signals and facial expression: A survey, In: *Deep Learning Approaches to Cloud Security*, Beverly: Scrivener Publishing, 2021, 27–42. https://doi.org/10.1002/9781119760542.ch3
28. Y. Alotaibi, A new meta-heuristics data clustering algorithm based on tabu search and adaptive search memory. *Symmetry*, **14** (2022), 623. https://doi.org/10.3390/sym14030623
29. H. S. Gill, O. I. Khalaf, Y. Alotaibi, S. Alghamdi, F. Alassery, Multi-model CNN-RNN-LSTM based fruit recognition and classification, *Intell. Autom. Soft Comput.*, **33** (2022), 637–650. https://doi.org/10.32604/iasc.2022.02258
30. Y. Alotaibi, M. N. Malik, H. H. Khan, A. Batool, S. U. Islam, A. Alsufyani, et al., Suggestion mining from opinionated text of big social media data, *CMC*, **68** (2021), 3323–3338. https://doi.org/10.32604/cmc.2021.016727
31. H. S. Gill, O. I. Khalaf, Y. Alotaibi, S. Alghamdi, F. Alassery, Fruit image classification using deep learning, *CMC*, **71** (2022), 5135–5150. https://doi.org/10.32604/cmc.2022.022809
32. T. Thanarajan, Y. Alotaibi, S. Rajendran, K. Nagappan, Improved wolf swarm optimization with deep-learning-based movement analysis and self-regulated human activity recognition, *AIMS Mathematics*, **8** (2023), 12520–12539. https://doi.org/10.3934/math.2023629
33. S. Koelstra, C. Mühl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, et al., DEAP: A database for emotion analysis; using physiological signals, *IEEE T. Affect. Comput.*, **3** (2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

AIMS Press