DOI: 10.4149/BLL_2023_133

Bratisl Med J 2023; 124 (11)
862 – 869

REVIEW

Artificial neural networks based multimodal device for autism spectrum disorder

Haixia YAN

Shaanxi University of Technology, Hanzhong, Shaanxi, China. **haiyansut@gmail.com**

ABSTRACT
The neuro developmental condition known as Autism Spectrum Disorder (ASD) affects people on a lifetime basis and exhibits itself in a wide range of ways. In this research work a brand-new semi-supervised training method for the recognition of discrete multi-modal autism spectrum disorder is proposed. At the coarse-grained level, we consider that various methodologies are anticipated to explore equivalent information about child autism. To build DC AlexNet, this combines two small network branches and a large network (trunk network). The network trunk is programmed just to become familiar with the distinguishing characteristics shared by face images at different resolutions. It is built using recently suggested residential components. To project images to a place where their ranges are as little as possible, two branch networks are programmed to learn coupled-mappings (CMs) that are particular to a given resolution. The suggested technique is properly assessed utilizing the databases for the OMEGE and DIAEMO datasets by evaluating it to state-of-the-art techniques in terms of many parameters. Deep Coupled AlexNet is developed to obtain 98.13 % of accuracy, 95.1 % of precision, 94.3 % of recall and 95.4 of F1-score for OMEGE dataset. Moreover, 98.6 % of accuracy, 97.2 % of precision, 98.5 of recall and 97.5 % of F1-score for DIAEMO dataset *(Tab. 8, Fig. 10, Ref. 16)*. Text in PDF *www.elis.sk*
KEY WORDS: autism spectrum disorder, artificial neural networks, emotion recognition, interaction design, multimodal factors.

Introduction

People with Autism Spectrum Disorder (ASD) exhibit difficulties in their socio-emotional reciprocity, that can take various forms, including a strange social attitude and incapability to participate in typical front and back conversation to a complete lack of initiating social interaction, based on the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM V) (1, 2). Autism-related social communication problems are partly caused by autism spectrum disorder (ER) deficiencies, which highlight the challenges with performing and comprehending socio-emotional signs (3). Autism sufferers exhibit issues with emotional processing, which include the capacity to recognize faces, distinguish between them, identify facial expressions, and remember faces.

As a result, people with autism frequently exhibit elevated tension and anxiety, aberrant face recognition, and disrupted emotion processing. Emotional competence, or the capacity to recognize and categories internally experienced emotions, helps in understanding our inner feelings and control when they influence us and others, which enhances our sense of wellbeing. Building emotional intelligence, that has been shown to help people in many areas of their lives, including their professional and educational careers, is facilitated by increasing our emotional awareness (4, 5). Humans acquire and enhance emotional intelligence throughout their lifetimes, which includes the capacity to recognize one's own feelings as well as emotional awareness, or the capacity to identify other people's emotions. But not everyone has the same capacity to develop this talent.

For illustrations, individuals with autism spectrum disorder (ASD) struggle to comprehend and manage their own feelings as well as the emotions of others. One of the most effective means of communication is the expression of human emotion on the face. Faces are incredibly expressive. It has been discovered that the language aspect of a communication only makes up a meagre 7 % of its overall relevance and influence, whereas the tone of the message accounts for around 38 % of the whole signal and the leftover signifies or portrays 55 % of the message (6).

Researchers have been drawn to this field that since pioneering work of Charles Darwin on the study of human emotion. There are seven fundamental emotions that all people experience. These basic emotions-neutral, angry, disgusted, scared, pleased, sad, and surprised – can be recognized from a person's facial expression. Since every person's face is unique, finding a solution to the issue of facial feature recognition is not an easy task. Numerous variables, including physical attributes, sex, genes, and age, influence the features (7). The challenge is very difficult due to the high level of unpredictability. When creating an emotion identification sys-

Shaanxi University of Technology, Hanzhong, Shaanxi, China.

Address for correspondence: Haixia YAN, Shaanxi University of Technology, Hanzhong 723001, Shaanxi, China.

tem, numerous considerations must be made. Any face processing system's primary step is to accurately recognize and categorize faces. The facial expression recognition system must function in a variety of environments, including those with fluctuating illumination conditions and other lighting issues, the wearing of eyeglasses, the presence of beards, etc. (8). To construct an ideal system, the system must be able to solve some of these issues. Consequently, the following are the contributions of this work:

The layout of the present paper is structured as follows: A relevant collection of research for a neural network model of emotional interaction is included in section 2 of the presentation. Section 3 provides more details on the proposed multimodal feature extractor and classifier. In section 4, the performance of the suggested model is shown along with a comparison. Section 5 presents the general conclusion for the suggested model.

Literature overview

The quality of multi-modal characteristics is critical in emotion recognition. Therefore, effective elements in audio, visual, and lexical modalities have been studied in earlier research for autism spectrum disorder.

The effect of the categorization approach, choosing the best feature combination, and data augmentation on the accuracy of speech autism spectrum disorder is described in (9). Reducing computation complexity is mostly dependent on selecting the proper handcrafted feature and classifier combination. In terms of classification, the proposed model, a 1D convolutional neural network (1D CNN), performs better than conventional machine learning methods. In (10), the authors offer a unique recurrent neural network model for identifying students' personal and collective feelings as depicted in camera footage. The model is used to recognize students' emotions based on footage pictures of their faces. In (11), the author introduces a method called the Gated Bidirectional Alignment Network (GBAN) that combines a novel group gated fusion (GGF) layer to incorporate the portrayals of various methodologies with an attention-based bidirectional alignment network with hidden states LSTM to expressly acquire the alignment relationship with both text and speech. By building a facial motion speech autism spectrum disorder (FM-SER) model, it is possible to discuss feature extraction of speaking, video, and object tracking in (12). As a result, the decision level integrating scheme is created with a higher degree of accuracy and a greater positive computational efficiency for emotion recognition. As the economy has grown and technological advancements have continued to evolve, many electronic items have been incorporated into all facets of people's life with benefits like speed and convenience (14). Electronic mobile terminal gadgets have arguably become a part of everyday life. It is inextricably linked and interdependent.

The APP interface vision is utilized as the preliminary step to investigate and examine the development of a student-friendly human-computer interface helpful for developing a more kid-friendly APP user experience. This is done with help of the survey of children's cognitive habits, cognitive science, and application preferences. Keeping the therapy more effective to users, (15) improves the robot's capacity to sense user emotions using face-reading facial expressions and to generate light sensory effects. A serious game in which the youngster must discern the feelings of the robot was designed as a therapy session. In order to enable social engagement with pupils, (16) developed a system that is capable of identifying emotions by facial expressions and interacting with a robotic system (Zeno R50 Robokind robotic platform, dubbed ZECA). Activities involving social communication employed ZECA as a mediator.

Proposed system framework

We presume that a video database generally contains data from key representatives (acoustic, visual and lexical). Select a labeled video database $\{X^L,Y\} = (x_i^a, x_i^v, x_i^l, y_i)\}\hat{}\, n^L$ and an unlabeled video database $\{X'^{uL},Y\} = (x'^a_i, x'^v_i, x'^l_i)\}\hat{}\, n^{uL}$, where x^a, x^v, x^l symbolize, the features are represented using the verbal, visual, and aural modes, accordingly. L and uL are employed to differentiate data with labels and without labels, and n^L and n^{uL} represent the size of the tagged and untagged data's, our objective is to use untagged data in testing to improve the recognition performance. Prior research on the timing of human emotion expression indicated that speech emotions can typically be reliably identified after 4 seconds. Based on this finding, we assume that even though emotion expression varies between modalities at the frame level, the aggregate emotional status should be consistent at the coarse-grained assertion level. This presumption can be used to extract supervision from data which is without label. While training, we enhance classification performance on labelled data while simultaneously reducing the difference in inter-modality distribution on labelled and unlabeled data, as depicted in the equation (1).

$$objective = classify\ (X^L,Y) + Reconstruct\left(X^L,X'^{uL}\right) + match\,(X^L,X'^{uL}) \quad (1)$$

Design model for autism children

The design analysis stage is where children's smart toys are prepared. This step demands study and investigation based on the data acquired to determine useful information that will be employed.

Input: The quantity of smart toys P with dataset $Q = \{q_1, q_2, \ldots q_n\}$ having n objects.

Output: P number $\{R_1, R_2, \ldots R_k\}$ to reduce the objective function

Choose the number of toys P .

Randomly choose k objects based on data of the toy center $C_1, C_2, \ldots C_k$

Allot the object $q_i (i = 1,2,3 \ldots n)$ to the nearby smart toy center $C_j, 1 < j < p$ where m is the total amount of data variables as shown in equation (2).

$$\left|q_i - C_j\right| = \min_{i<j<q} \sqrt{\sum_l^m \left(q_{il} - C_{jl}\right)^2} \quad (2)$$

Estimate the new center C_j of every intelligent toy, where N_j is the number of objects in the j-th intelligent toy S_j as shown in equation (3).

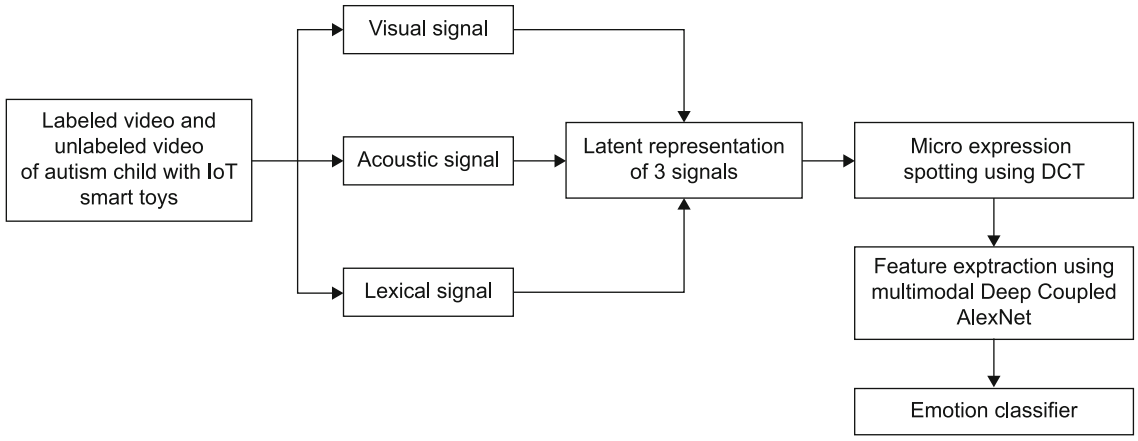

Fig. 1. Framework of interaction design based on emotion recognition.

$$C_j = \frac{1}{N_i}\sum_{q_j} q_i, \quad j = 1,2,\ldots p \tag{3}$$

If the objective function is the minimum and the smart toy center remains unchanged, the algorithm terminates.

Design of Framework model

The purpose of emotional interaction is to get the software to make logical modifications based on the assumption that it recognizes when people feel and can react to the passage of user emotions.

The goal of designing emotional interactions is to identify, understand, and react to human emotions using a range of perceptive techniques. The framework for interface design focused on autism spectrum disorders is depicted in Figure 1. The sensing, implementation, identification, emotion calculation, and optimizing modules make up the majority of the interactive design system.

Representation of multimodal features

The sound files were first made mono and re sampled to 16.000 Hz. The voice waveform samples were run through a bank of 31 gamma tone filters that were ranged 80 to 8000 Hz apart. The Hilbert transform was then used to estimate the envelope in each gamma tone sub band. Then, a modulation filter bank with a set of 25 regularly spaced first-order Butterworth band pass filters with a frequency of 0.75 Hz and a distance of 0.5 Hz was used to transmit the envelopes in each sub band. The average of every envelope sub band was again calculated using the gamma tone filters, it was resembled to 25 Hz to reflect the video frame rate, after which it was normalized to have zero-unit variance and mean for each footage. The correlation between the time-continuous analogue signal $x\,(t)$ and the frequency spectrum $X\,(f)$ can be shown in equation (4) and (5)

$$x\,(t) = \int_{-\infty}^{+\infty} X(f)e^{i2\pi ft}\,df \tag{4}$$

$$X(f) = \int_{-\infty}^{+\infty} x(t)e^{i2\pi ft}\,dt \tag{5}$$

The continuous signal is recovered by the discrete sampling pulse signal $x\,(nTs)$. The Nyquist sampling theorem states that the sampling interval of a discrete pulse signal $x\,(nTs)$ should fulfill the criteria $T_s \le \frac{1}{2f}$ when the frequency of the sampled signal, f_s, is greater than double that of the signal's maximum frequency, f_{max}.

This is illustrated in equation (6) and (7).

$$x\,(nTs) = \int_{-1/2T}^{1/2T} X(f)e^{i2\pi ft},df \tag{6}$$

$$X(f) = T_s\sum_{n=-\infty}^{+\infty} x\,(nTs)e^{-i2\pi fnT} \tag{7}$$

$x\,(nTs)$ s) Can completely indicate $X(f)$; $x(t)$ can be determined as follow in equation (8)

$$x(t) = T_s\sum_{n=-\infty}^{+\infty} x\,(nTs)\int_{-1/2T}^{1/2T} X(f)e^{i2\pi ft},df \tag{8}$$

The visual signals that can obtain landmark time series were first displaced by one sample and low-pass filtered at 8 Hz to reduce jitter from the frame-to-frame estimation. Energy above this frequency range is not likely to be generated by speaker motion that can be picked up at a 25 Hz video sample rate. Canonical correlation analysis (CCA) identifies linear transformations which project each data set to a common space where their correlation is optimum for multidimensional signals. Let $X_A \in R^{T\times I_A}$ and $X_A \in R^{T\times I_v}$ be two zero-mean datasets, where T denotes time, and J_A, J_V are the number of features in the two datasets. Since the projections of the centered data $X_A w_{Aj}$ and $X_V w_{Vj}$ are optimally correlated, CCA predicts pairs of vectors w_{Aj} and w_{vj} as follows in equation (9) and (10).

$$\rho = \max\frac{(X_A w_{Aj})^T(X_V w_{Vj})}{|X_A w_{Ai}||X_V w_{Vi}|} \tag{9}$$

$$= max\frac{w_{Aj}\,\Sigma_{AV} w_{Vj}}{\sqrt{|w_{Aj}\,\Sigma_A w_{Vj}||w_{Vj}\,\Sigma_v w_{Vj}|}} \tag{10}$$

Where $\Sigma_A = X_A{}'X_A$, and $\Sigma_v = X_v{}'X_v$ are the (un normalized) covariance matrices and $\Sigma_{Av} = X_A{}'X_v$ is the cross-covariance.

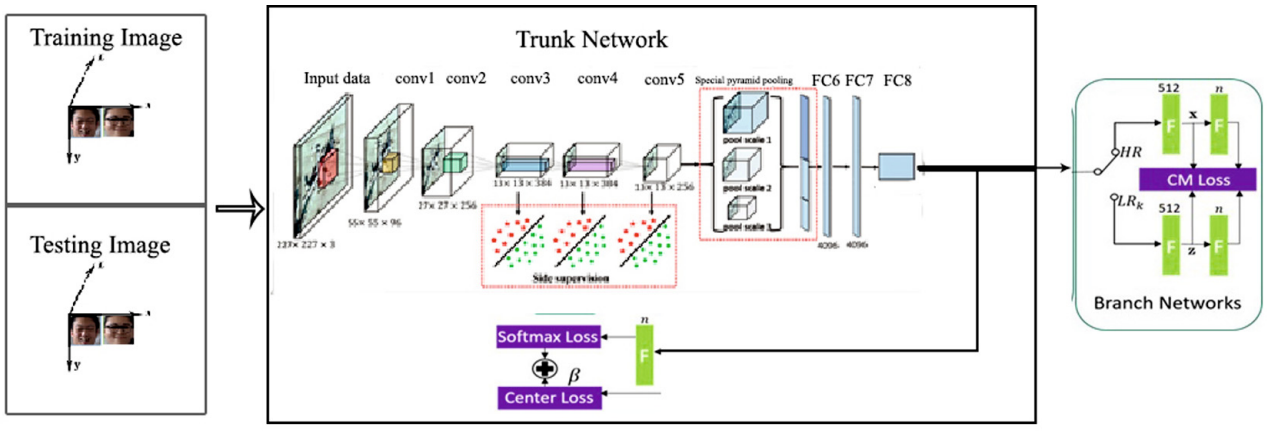

Fig. 2. Architecture of the proposed DCAlexNet model.

The canonical variates or canonical constituents are mappings of the data $X_A w_{Aj}$ and $X_v w_{vj}$.

Micro expression spotting using DCT

A common handcrafted element in the frequency domain is a one-dimensional vector known as upper-left discrete cosine coefficient. There are eight conventional DCT variants. In our work, we employ the most popular DCT-II variation, which is specified as follows.

$$F_{u,v} = 4\sum_{x=0}^{N_1-1}\sum_{y=0}^{N_2-1} s(x,y)[\cos\frac{2x+1}{2N1} + \cos\frac{2y+1}{2N2}] \quad (11)$$

Where, $s, F \in R^{(N1,N2)}$ are the images with $N1 \times N2$ pixels are in spatial domain and frequency domain. These kernels are designed to reconnect as much as possible and can be used independently to apply to input images to provide estimations of the gradient component for each orientation i.e, $V(i)$ and $H(j)$. To get the exact gradient magnitude at every site, these can then be aggregated. Rotation correction mainly aims to maintain two eyes on the horizontal line. The image is cropped and resized horizontally between two ears and vertically between the jaw and eyebrows. As a result, the preprocessed image has 128128 pixels and includes face organs that are significant to expression.

Feature extraction using multimodal Deep Coupled AlexNet

To select image characteristics at a finer degree of resolution, the system utilizes deep architectures with 11 to 18 convolution layer and only 7*7 filter layers. It expanded the proportion of feature map in the third, fourth, and fifth convolutional layers from (385,384,256) (385,384,256) to (512,1024,512) (512,1024,512), which also increased the number of attributes that the network can recognize. First, 3*3 Max Pooling and a dropout layer with a parameter of 0.8 was introduced; next, local response normalization was utilized in the outcome units; and finally, the batch size was adjusted to 1000 in training the network. The pooling layer came after the convolutional layer, which decreased the dimensionality of the feature and prevented fitting problem. The convolutional layer involved a 3 x 3 filters and exponential linear unit for the purpose of activation function. The dropout layer employed a value of 0.8, the maxpool layer a 4 x 4 filter, and the up-sampling layer a 7 x 7 filter. The entire network was flattened using convolutional layer aggregation. It was ultimately output using the sigmoid function and convolutional layer. Figure 2 depicts the suggested DCAlexNet model's design.

The gradient is a partial derivative of the loss for every variable that can be learned, and a specific variable updating is described as follows in equations (12) and (13),

$$x(l) = x(l) - \beta\frac{\delta J(x,a)}{\delta x(l)} = x(l) - \beta\sum_{i=1}^{n}\frac{\delta J(x,a;r(i),s(i)}{\delta x(l)} \quad (12)$$

$$y(l) = y(l) - \beta\frac{\delta J(x,a)}{\delta x(l)} = y(l) - \beta\sum_{i=1}^{n}\frac{\delta J(y,a;r(i),s(i)}{\delta y(l)} \quad (13)$$

β is the update rate of the parameters, x and y are the offset coordinates and weight matrix for each layer, $a(i), b(i), 1 \le i \le N$ is a given set of samples. The levels of activation produced by the inception model are adjusted using the feed-forward layer. It generates values that are precisely below 1 and closer to 0. These values result in less computational complexity and are simpler to deal with in the model. The width of the enhanced image produced by the inception model is decreased through pooling.

Assume that the input sample $X_i \in R^n$ represents the raw input data and $Y_i \in \{1,2,3 \ldots k\}$ represents the corresponding ground truth label for sample X_i. There are M layers in total in the pre-trained AlexNet architecture, the weight combinations for the pre-trained DCAlexNet architecture are $W = W(1), W(2), \ldots W(M)$. Meanwhile, for every classifier in each hidden layer of the pre-trained AlexNet architecture, the associated weights are $w = w^{(1)}, .. w^{(m-1)}$. The interactions between the filtration and weight variables in the pre-trained AlexNet architecture are depicted in Equations (14) and (15):

$$Z^{(m)} = f(Q^{(m)} \text{ and } Z^{(0)} = X \quad (14)$$

$$Q^{(m)} = W^{(m)} - Z^{(m-1)} \quad (15)$$

In Equations (14) and (15), m stands for the amount of layers in the pre-trained AlexNet architecture, m stands for a particular layer in that infrastructure, $W^{(m)}$, m = 1...M are the network weight values to be trained, $Q^{(m)}$ stands for the convolved reactions on the preceding characteristics map, and f(.) is the pooling function on Q. Equation (16) displays the overall significant function of pre-trained AlexNet architecture,

$$F(W) = P(W) + Q(W) \tag{16}$$

where $P(W)$ and $Q(W)$ are the production desire and the totaled related goals, which are depicted in Equations (17) and (18),

$$P(W) = log_x w(out) + W(w(out)) \tag{17}$$

$$Q(W) = \sum_{m=1}^{M-1} log_x w(out) + W(w(out)) \tag{18}$$

Where, $w(out)$ pertains towards the final gradient classifier value.

The aforementioned formulations start making it subjectively clear that the pre-trained AlexNet architecture imposes a limitation at each hidden layer to effectively develop a better tag forecasting and provide a significant push for possessing inherently biased and reasonable features at every specific layer in addition to learning the convolutional kernels W^*.

Emotion classification and its loss

We can use multi-modal fusion by explicitly integrating the latent representation and then feeding it into the classifier since we presume that the emotion state is aligned between methods at the observation level. We calculate the cross-entropy loss for optimization for the labelled data as given in equation (19) and (20).

$$L_{cls} = \frac{1}{n^L} \sum_{i=1}^{n^L} \sum_{k=1}^{K} y_{i,k} \, log(p_{i,k}) \tag{19}$$

$$(p_{i,0}, p_{i,1}, \ldots p_{i,k}) = softmax\left(C\left(\left[z_i^a; z_i^v; z_i^l\right]\right)\right) \tag{20}$$

where $\boldsymbol{C}$ is an emotion classifier using neural networks, K is the classifications of emotion loss, $\boldsymbol{n^L}$ is the maximum number of supervised samples, $\boldsymbol{y_i}$ and $\boldsymbol{p_i}$ are the annotation and forecasted emotion class probabilities for input data $\boldsymbol{x_i}$ and $\left[\boldsymbol{z_i^a}; \boldsymbol{z_i^v}; \boldsymbol{z_i^l}\right]$ is the combination of all of the modality' hidden representations of $\boldsymbol{x_i}$. We create the combined loss function shown below by combining all of the losses. The loss function is set for the supervised portion as given in equations (21) and (22):

$$L^s = L_{cls} + \propto L_{rec} + \beta L_{pair} \tag{21}$$

$$L_{rec} = L_{DAE}(z^a) + L_{DAE}(z^v) + L_{DAE}(z^l) \tag{22}$$

The latent representation still collapses into zero space, we discover. We include unpaired samples in the training to prevent nonsensical matching of modal representation. The term "unpaired" denotes that the feature obtained from many modalities is not synchronized or from the same video. There should be a clear difference between the information's paired and unpaired residual dispersion. The features are moved around among the auditory, verbal, visual, and other modes so they are no longer arranged. This allows us to create unpaired samples. We anticipate that the unpaired samples will be mapped into distinct emotion spaces, increasing the training distribution distance.

Tab. 1. Accuracy for OMGE dataset.

Categories	1D CNN	RNN	GBAN	DC-AlexNet
Happy	89.7	85.6	91.3	98.6
Anger	87.6	88.0	91.6	97.0
sadness	87.0	87.5	92.0	97.6
neutral	88.4	88.3	92.3	96.0

Tab. 2. Analysis of Precision for OMGE dataset.

Categories	1D CNN	RNN	GBAN	DC-AlexNet
Happy	81.3	87.4	93.0	97.4
Anger	88.0	87.0	93.5	93.0
sadness	83.4	87.2	92.0	95.4
neutral	84.0	87.4	92.1	94.0

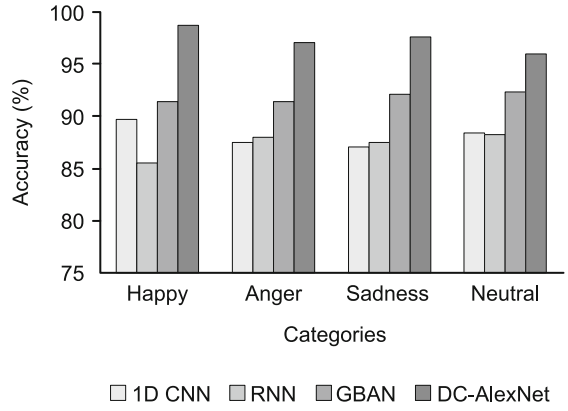

Fig. 3. Comparison of accuracy for OMGE dataset.

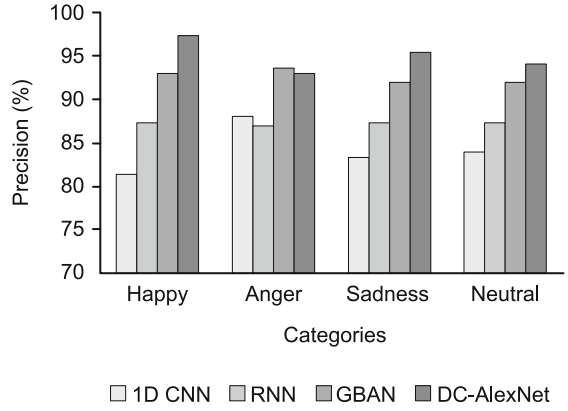

Fig. 4. Comparison of precision for OMGE dataset.

Results and discussion

We measure the effectiveness of our proposed Deep Coupled AlexNet (DC-AlexNet) utilizing measures including accuracy, precision, recall, F1-score, and AUC-score. Three cutting-edge techniques, including 1D convolutional neural network (1D CNN) (9), Recurrent neural network (10), and Gated Bidirectional Alignment Network (GBAN) (11), are evaluated with these parameters.

Dataset description: The following two datasets are used to thoroughly examine the suggested method:

OMGE Dataset (16) – The publicly accessible3 OMGE Dataset is based on monologues. Each utterance in the dataset's video examples is attributed to a single fundamental emotion. Video instances are broken up into many utterances. In order to create instances with compound emotions, we merged two to three utterances from a video.DIAEMO Dataset (17).

Quantitative performance on OMGE Dataset

Tables 1 and 2 show the evaluation of accuracy and precision for OMSE dataset for the proposed DC-AlexNet and the existing 1D CNN, RNN, and GBAN accordingly.

The accuracy for OMGE dataset comparison of the proposed DC-AlexNet and the existing 1D CNN, RNN, and GBAN is shown in Figure 3. The X and Y axes show different categories and percentage numbers, correspondingly. In contrast, the suggested DC-AlexNet technique achieves 98.13 % accuracy, which is 10.13 % superior to the existing 1D CNN, RNN, and GBAN methods. This is 9.13 % better than RNN and 5.53 % better than GBAN method. Figure 4 illustrates the comparison between the proposed DC-AlexNet technique and the existing 1D CNN, RNN, and GBAN methods. The proposed DC-AlexNet approach obtained 95.1 % of precision, which is 11.1 % better than the 1D CNN, 8.2 % better than the RNN method, and 2.3 % better than that of the GBAN method.

The accuracy for OMGE dataset comparison of the proposed DC-AlexNet and the existing 1D CNN, RNN, and GBAN is shown in Figure 3. The X and Y axes show different categories and percentage numbers, correspondingly. In contrast, the suggested DC-AlexNet technique achieves 98.13 % accuracy, which is 10.13 % superior to the existing 1D CNN, RNN, and GBAN methods. This is 9.13 % better than RNN and 5.53 % better than GBAN method. Figure 4 illustrates the comparison between the proposed DC-AlexNet technique and the existing 1D CNN, RNN, and GBAN methods. The proposed DC-AlexNet approach obtained 95.1 % of precision, which is 11.1 % better than the 1D CNN, 8.2 % better than the RNN method, and 2.3 % better than that of the GBAN method.

Table 3 and 4 show the evaluation of recall and F1_score for OMSE dataset for the proposed DC-AlexNet and the existing 1D CNN, RNN, and GBAN accordingly.

Figure 5 compares the recall for the OMGE dataset between the proposed DC-AlexNet and the existing 1D CNN, RNN, and GBAN. The X and Y axes show different categories and percentage numbers, respectively. Comparatively, the suggested DC-AlexNet approach obtains 94.3 % of recall, which is 18.5 % better than the existing 1D CNN, RNN, and GBAN methods that achieve 76.8 %, 88.3 %, and 93.4 % of recall, accordingly. According to Figure 6, when contrasted, the existing 1D CNN, RNN, and GBAN methods each attains 82.3 %, 86.3 %, and 87 % of the F1-score, whereas the suggested DC-AlexNet method attains 95.4 % of the F1-score, which is 13.1 % superior than the 1D CNN, 9.1 % higher than the RNN method, and 8.4 % better than the GBAN method.

Tab. 3. Analysis of recall for OMGE dataset.

Categories	1D CNN	RNN	GBAN	DC-AlexNet
Happy	76.4	89.0	89.5	91.4
Anger	74.5	86.5	90.2	91.6
sadness	77.0	88.0	91.4	92.4
neutral	76.4	87.1	90.0	94.0

Tab. 4. Analysis of F1-score for OMGE dataset.

Categories	1D CNN	RNN	GBAN	DC-AlexNet
Happy	78.4	90.0	90.2	91.6
Anger	76.5	88.2	92.0	93.8
sadness	79.0	89.0	93.5	94.0
neutral	78.4	88.7	92.0	95.0

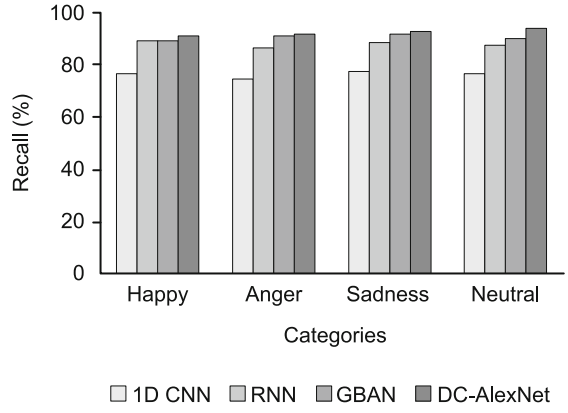

Fig. 5. Comparison of recall for OMGE dataset.

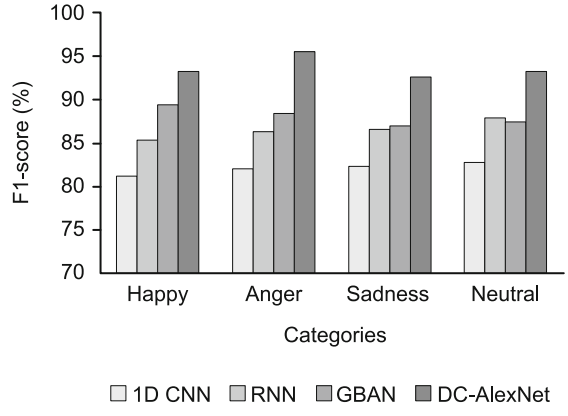

Fig. 6. Comparison of F1-score for OMGE dataset.

Tab. 5. Analysis of accuracy for DIAEMO Dataset.

Categories	1D CNN	RNN	GBAN	DC-AlexNet
happy	87.9	81.2	87.4	98.7
surprise	88.0	81.4	87.4	98.0
Happy surprise	87.4	81.0	88.2	98.5
awful	88.3	81.6	86.4	98.6
Surprise angry	88.2	81.6	86.3	97.0
sad	88.4	81.7	88.1	97.5
Surprise angry sad	88.1	82.4	86.0	97.5

Tab. 6. Analysis of precision for DIAEMO Dataset.

Categories	1D CNN	RNN	GBAN	DC-AlexNet
happy	83.4	71.2	88.1	97.3
surprise	83.2	71.6	88.0	97.5
Happy surprise	82.4	72.0	88.6	98.0
awful	83.0	72.7	87.0	97.4
Surprise angry	83.6	72.3	87.5	97.2
sad	83.4	73.1	86.0	97.6
Surprise angry sad	82.0	73.7	86.4	98.1

Tab. 7. Analysis of recall for DIAEMO Dataset.

Categories	1D CNN	RNN	GBAN	DC-AlexNet
happy	87.0	84.3	76.8	98.0
surprise	87.4	84.6	78.3	98.6
Happy surprise	87.3	85.0	78.0	98.5
awful	86.4	85.9	72.4	98.3
Surprise angry	86.9	85.7	77.0	98.0
sad	86.4	82.4	76.3	97.3
Surprise angry sad	86.1	84.2	75.2	97.2

Tab. 8. Analysis of F1-score for DIAEMO Dataset.

Categories	1D CNN	RNN	GBAN	DC-AlexNet
happy	78.3	84.3	76.8	98.5
surprise	76.4	84.6	78.3	97.0
Happy surprise	78.4	85.0	78.0	97.3
awful	77.8	85.9	72.4	97.2
Surprise angry	77.3	85.7	77.0	95.4
sad	77.0	82.4	76.3	96.5
Surprise angry sad	76.4	84.2	75.2	96.3

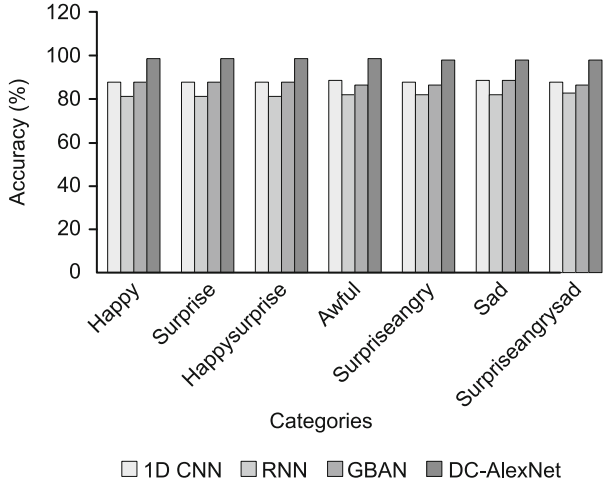

Fig. 7. Comparison of Accuracy for DIAEMO Dataset.

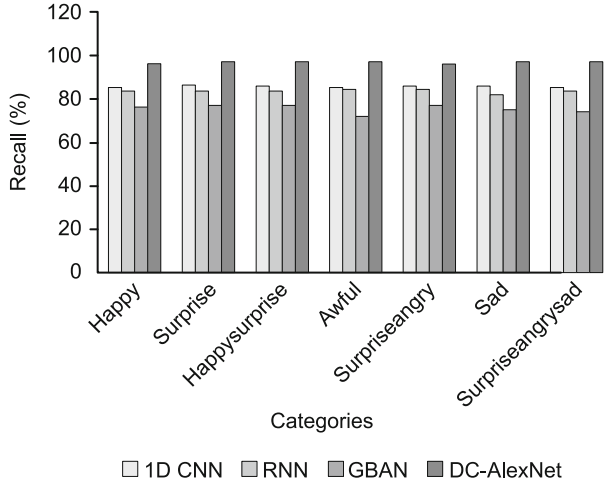

Fig. 9. Comparison of recall for DIAEMO Dataset.

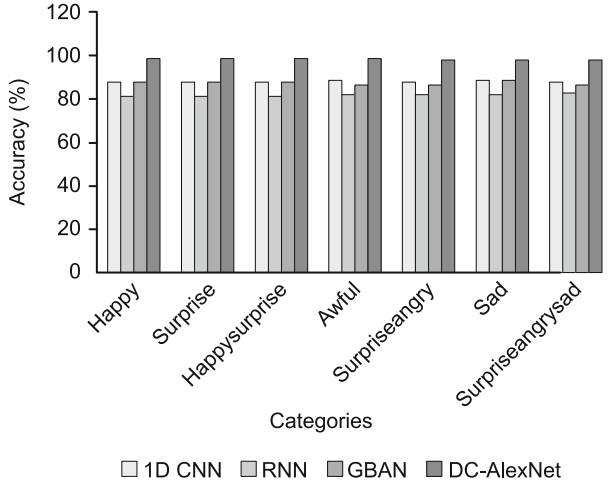

Fig. 8. Comparison of precision for DIAEMO Dataset.

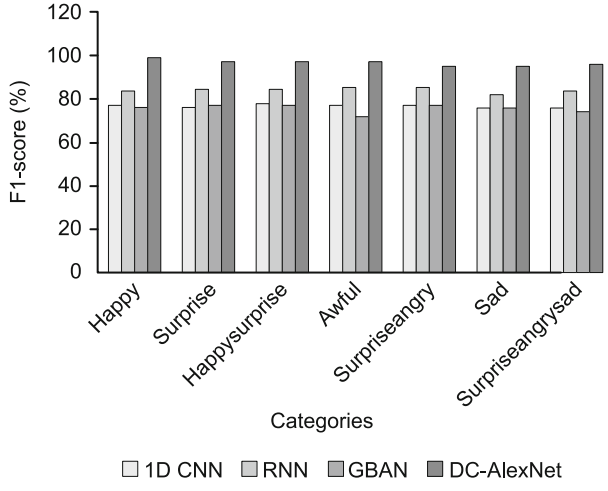

Fig. 10. Comparison of F1-score for DIAEMO Dataset.

Quantitative performance on DIAEMO Dataset

Tables 5 and 6 show the evaluation of accuracy and precision for DIAEMO dataset for the proposed DC-AlexNet and the existing 1D CNN, RNN, and GBAN accordingly.

As depicted in Figure 7, the existing 1D CNN, RNN, and GBAN methods each attain accuracy levels of 82.3 %, 86.3 %, and 87 %, respectively, while the proposed DC-AlexNet technique obtains accuracy level of 95.4 %, which is 13.1 % higher than the 1D CNN, 9.1 % higher than the RNN method, and 8.4 % higher than the GBAN method. As seen in Figure 8, the suggested DC-AlexNet technique obtains 97.2 % of precision, which is 14.2 % better than 1D CNN, 24.2 % better than RNN, and 10.3 % better than GBAN method. In comparison, the existing 1D CNN, RNN, and GBAN methods reach 83.4 %, 73.4 %, and 87.5 % of precision, respectively.

Tables 7 and 8 show the evaluation of accuracy and precision for DIAEMO dataset for the proposed DC-AlexNet and the existing 1D CNN, RNN, and GBAN accordingly.

As seen in Figure 9, the suggested DC-AlexNet technique obtains 98.5 % of recall, which is 12.1 % better than 1D CNN, 12.1 % better than RNN, and 20 % better than GBAN method. In comparison, the existing 1D CNN, RNN, and GBAN methods obtain 86.4 %, 85.3 %, and 78.5 % of recall, correspondingly. As seen in Figure 10, the proposed DC-AlexNet technique obtains 97.5 % of the F1-score, which is 20 % better than the existing 1D CNN, RNN, and GBAN methods and 11 % better than the existing RNN and GBAN methods, correspondingly.

Conclusion and futurework

This research work used facial features to recognize emotions in autistic children using deep learning. One large trunk network with two little network branches make up the multimodal Deep Coupled AlexNet emotion identification model that has been proposed. The trunk network has only been trained once to learn the distinguishing characteristics that are shared by face photos of various resolutions. The outcomes of the model categorization demonstrated to us the possibility for employing computer vision and deep learning techniques as processing algorithms for both professionals and families to much more quickly and accurately identify autism. The successful completion of complicated behavioral and psychological investigations for the diagnosis of autism, which takes more time and effort, is facilitated by computer tools.

References

1. Fridenson Hayo S, Berggren S, Lassalle A, Tal S, Pigat D. Basic and complex autism spectrum disorderin children with autism: Cross-cultural findings. Mol Autism 2016; 7 (52). https://doi.org/10.1186/s13229-016-0113-9.

2. Omer A, Aydin A, Hasan D, Tahir Cetin A. Analysis of Gait Dynamics of ALS Disease and Classification of Artificial Neural Networks. Tehnički vjesnik 2018; 25: 183–187. https://doi.org/10.17559/TV-20160914144554

3. Goran D. Interface Providers in Alternative Communication for Children with Autism Spectrum Disorders. Acta Clin Croat 2013; 52 (1).

4. Daniels MA, Mandell DS. Explaining differences in age at autism spectrum disorder diagnosis: a critical review. Autism 2014; 18 (5): 583–597. DOI: 10.1177/1362361313480277.

5. Noor H, Shahbodin F, Pee N. Serious game for autism children: review of Literature. World Academy of Science, Engineering and Technology 2012; 6 (4): 554–559. https://doi.org/10.5281/zenodo.1333272.

6. Giannopoulos P, Perikos I, Hatzilygeroudis I. Deep learning approaches for facial emotion recognition: a case study on FER2013. In: Hatzilygeroudis I, Palade V (Eds). Advances in hybridization of intelligent methods, Springer, Berlin, 2018, 1–16. https://doi.org/10.1007/978-3-319-66790-4_1.

7. Littlewort G, Whitehill J, Fasel I, Wu T, Frank M, Movellan J et al. The computer expression recognition toolbox (CERT) In: Face and gesture. IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, Santa Barbara, California, USA, 2011, 298–305. http://dx.doi.org/10.1109 %2FFG.2011.5771414.

8. Zafar B, Ashraf R, Ali N, Iqbal M, Sajid M et al. A novel discriminating and relative global spatial image representation with applications in CBIR. Appl Sci 2018; 8 (11): 1–23. https://doi.org/10.3390/app8112242.

9. Hu X, Chen J, Wang F, Zhang D. Ten challenges for eeg-based affective computing. Brain Sci Adv 2019; 5 (1): 1–20. https://doi.org/10.26599/BSA.2019.9050005.

10. Alnuaim A, Zakariah A, Alhadlaq M, Shashidhar A, Hatamleh C et al. Human-computer interaction with detection of speaker emotions using convolution neural networks. Comput Intelligence Neurosci 2022; 1–16. http://dx.doi.org/10.1155/2022/7463091.

11. Savchenko AV, Makarov IA. Neural network model for video-based analysis of student's emotions in E-learning. Optical Memory Neural Networks 2022; 31 (3): 237–244. https://doi.org/10.3103/S1060992X22030055.

12. Liu P, Li K, Meng H. Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition. arXiv preprint arXiv:2201.06309, 2022; 1–5. https://doi.org/10.1155/2023/9645611.

13. Jia N, Zheng C, Sun W. A multimodal autism spectrum disordermodel integrating speech, video and mocap. Multimedia Tools Applications 2022; 81: 1–22. https://doi.org/10.1007/s11042-022-13091-9.

14. Wang P, Han W. Construction of a new financial e-commerce model for small and medium-sized enterprise financing based on multiple linear logistic regression. Journal of Organizational and End User Computing 2021; 33 (6): 1–18. https://doi.org/10.4018/JOEUC.20211101.oa4.

15. Rocha M, Valentim P, Barreto F, Mitjans A, Cruz-Sandoval D et al. Towards Enhancing the Multimodal Interaction of a Social Robot to Assist Children with Autism in Emotion Regulation. In International Conference on Pervasive Computing Technologies for Healthcare, Tel Aviv, Israel, Springer, Cham, 2022; 398-415. http://dx.doi.org/10.1007/978-3-030-99194-4_25.

16. Silva V, Soares F, Esteves JS, Santos CP, Pereira AP et al. Fostering autism spectrum disorderin children with autism spectrum disorder. Multimodal Technologies Interaction 2021; 5 (10): 1–18. https://doi.org/10.3390/mti5100057.

Received April 23, 2023.
Accepted May 14, 2023.