Computers, Materials & Continua
DOI: 10.32604/cmc.2022.028229
Article

Tech Science Press

An Automated and Real-time Approach of Depression Detection from Facial Micro-expressions

Ghulam Gilanie[1], Mahmood ul Hassan[2], Mutyyba Asghar[1], Ali Mustafa Qamar[3,*], Hafeez Ullah[4], Rehan Ullah Khan[5], Nida Aslam[6] and Irfan Ullah Khan[6]

[1]Department of Artificial Intelligence, Faculty of Computing, The Islamia University of Bahawalpur, Pakistan
[2]Department of Computer Skills, Deanship of Preparatory Year, Najran University, Najran, Saudi Arabia
[3]Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
[4]Biophotonics Imaging Techniques Laboratory, Institute of Physics, The Islamia University of Bahawalpur, Pakistan
[5]Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
[6]Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia
*Corresponding Author: Ali Mustafa Qamar. Email: al.khan@qu.edu.sa
Received: 05 February 2022; Accepted: 10 March 2022

Abstract: Depression is a mental psychological disorder that may cause a physical disorder or lead to death. It is highly impactful on the social-economical life of a person; therefore, its effective and timely detection is needful. Despite speech and gait, facial expressions have valuable clues to depression. This study proposes a depression detection system based on facial expression analysis. Facial features have been used for depression detection using Support Vector Machine (SVM) and Convolutional Neural Network (CNN). We extracted micro-expressions using Facial Action Coding System (FACS) as Action Units (AUs) correlated with the sad, disgust, and contempt features for depression detection. A CNN-based model is also proposed in this study to auto classify depressed subjects from images or videos in real-time. Experiments have been performed on the dataset obtained from Bahawal Victoria Hospital, Bahawalpur, Pakistan, as per the patient health questionnaire depression scale (PHQ-8); for inferring the mental condition of a patient. The experiments revealed 99.9% validation accuracy on the proposed CNN model, while extracted features obtained 100% accuracy on SVM. Moreover, the results proved the superiority of the reported approach over state-of-the-art methods.

Keywords: Depression detection; facial micro-expressions; facial landmarked images

1 Introduction

Mental disorders are extreme psychotic issues that cause patients to think, precept, and behave abnormally. When the disease has progressed to a severe condition, persons suffering from serious

mental disorders struggle to maintain a connection to the real world and are generally unable to perform well in everyday life. Depression, schizophrenia, bipolar disorder, different dementia, psychoses, and formative disorders such as chemical imbalance are examples of mental disorders. Depression is a common mental disorder and one of the leading causes of disability worldwide. Globally, about 350 million people are suffering from depression [1]. Depression fundamentally influences an individual's family, connections, and other general well-being perspectives. In the worst case, depression can lead to suicide for many of them. In underdeveloped countries, people suffering from depression cannot get adequate medical facilities due to social stigma, cost, and lack of treatment. Human beings must set the dynamic state of a well-integrated element of mind, body, and spirit held in emotional balance, but due to depression, this balance is no more carried out. Depressed persons may lose self-esteem, grief, sadness, and the feeling of worthlessness overwhelmed. Early diagnosis of depression is necessary because of its bio-psychosocial factor, as many other diseases such as hypertension, anemia, diabetes mellitus, and heart disease may emerge [2]. These occur due to changes in the hormones, immune factors, metabolism, and neurotransmitter aggravated by depression, which all are associated with socioeconomic stressors. Moreover, these physical disorders are more focused than the mental disorders that are the base of the former one. Depression may influence the biological, social, and financial factors financial factor, such as insomnia, sleeping routine, dietary habits, child-rearing practices [3], personality development, and occupational hazards [4].

Both clinical and non-clinical assessment strategies are used to diagnose depression. The clinical method, which is based on the seriousness of symptoms, comprises PHQ-8 and depends only on patient reports or clinical decisions taken under the light of the Diagnostic and Statistical Manual of Mental disorder (DSM-5). Recently, electroencephalogram and eye tracking are also used for depression detection. Symptomatic methods of depression have vibrant afflictions related to patient refusal, poor sensitivity, emotional predispositions [5], over-diagnosis, and misdiagnosis [6]. People may not want to be screened or diagnosed with psychiatric conditions due to the associated stigma. The related stigma refers to the myth that a person with depression is posed as a mental in society or the mental disorder is taken as an abuse by the society. Imaging-based assessment strategies include brain Magnetic Resonance Imaging (MRI), gait, and facial expression analysis. MRI methods also play a vital role in diagnosing mental illness [7]. They are non-invasive and have a relatively good spatial and temporal resolution to find connectivity/interaction between brain and psychological variables. However, these are expensive, time-consuming, and may be harmful, as patients suffering from mental health problems may create serious dilemmas during MRI scanning.

All these hindrances make depression detection a toil exhaustive work. Recently, artificial intelligence, computer vision, and machine learning approaches have been applied to facial appearances for depression detection. Computer vision can recognize facial micro-expressions that can accurately detect the severity level and type of depression. Using computer-aided diagnostic systems, developed by integrating these approaches, every person can be screened in hesitation and hindrance-free manners. These systems could save the cost and time of both patient and the psychiatrist. Moreover, privacy can be ensured, which is most necessary due to associated stigma. Hence, such systems could prove themselves as effective, reliable, secure, non-invasive, real-time, robust, and facilitate early-stage detection.

Micro-expressions are transient and involuntary facial expressions, frequently happening in high-stakes circumstances when individuals attempt to cover or hide their actual emotions. These actions drive by the amygdala. They are excessively short (1/25 to 1/2 s) and unobtrusive for natural eyes to see [8]. The outflow of emotion can be limited to a part of the face, or maybe a quick-expression spread over the whole face such as changed gazing pattern, furrow on the head, eyebrow moment, and the

wrinkles and jaw movement. These expressions are used for lie detection and emotional intelligence [9]. In high stake situations, like when suspects are being interviewed, a micro-expression flees across the face of the suspect un-intentionally even if (s)he tries to conceal it. This subtle movement tells another story than what the suspect is telling.

According to Darwin [10], an expression is visible by the right combination of facial muscles. Paul [11] recommended facial actions and proposed a facial action coding system, which codes all the expressions of the face experienced by a human. These expressions are universal, irrespective of gender, culture, and ethnicity. In the facial action coding system, there are 44 Action Units (AUs), which enfold all the expressions (facial movements), including the seven universal emotions [12]. These AUs are distinguished based on contraction and relaxation of facial muscles, such as tightening of lips in anger, furrowing the eyebrows in anger or change in the movement of eyelids, and movement of cheeks in sadness. Out of 44 AUs, 30 are related to the contraction of muscles, 12 are for the upper face, and 18 are for the lower face. They can appear as an action unit or a combination of these on the face to embody an emotion. Despite all this, these AUs can be additive or non-additive. If AUs do not change an individual's appearance when integrated with others are called additive; otherwise, non-additive [13]. Paul also discovered universal micro-expressions that appear on the face involuntarily within the quarter or half of a second [14]. It is challenging to detect these micro-expressions using a naked eye, but the high-resolution camera can capture them. Therefore, in the proposed research activity, we do facial expressions analysis through video frames of depressed patients and healthy controls for depression detection.

The paper is organized as follows: Section 2 contains a review of related studies, while Section 3 embodies details about the data acquisition, pre-processing, and the proposed CNN model used for depression detection. In Section 4, experimental work is performed; this section also represents experimental results, their discussions, and comparison with the state-of-the-art methods. Conclusions and future directions are part of Section 5.

2 Literature Review

In the research article [15], interpersonal context and clinical interviews, comprised of head movements, facial dynamics, and vocal parody, were focused on which depression severity has been detected. A total of 57 depressed patients were requested to participate in the study, and their interviews were done using the Hamilton depression rating scale. In this research, 3D registration from the 2D video is done using Z-Face technology, then stacked Denoising Auto-encoders (SDAE) were used for encoding the facial and head movement by carrying out the mapping between features and improved fisher vector coding. The accuracy achieved by this research is 78.67%. In the research [16], the authors proposed an automatic depression detection system by approximating the Beck Depression Inventory score (BDI). BDI score is based on the analysis of facial expression features. Median Robust Local Binary Patterns from three orthogonal Planes (MRLBP-OP), which can extract both macrostructure and microstructure of facial appearance and dynamics, are used to extract facial dynamic features in this model. This work also proposes using Dirichlet Process Fisher Encoding (DPFP) to aggregate the MRLBP-Top over an image sequence. DPFV uses Dirichlet Process Gaussian Mixture Models (DPGMM) to obtain the number of GMMs and model parameters automatically. The depression databases AVEC-2013 and AVEC-2014 were used for the experiments and analyses. The results were evaluated as per Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), which are 9.20 and 7.55, respectively.

In the research [17], the author suggested a model estimate beck depression inventory score from video data for depression level analysis. The framework automatically learns spatiotemporal features of the face region using recurrent neural networks and 3D convolution neural network RNN-C3D. Experiments were conducted on AVEC-2013 and AVEC-2014 depression datasets. The results obtained have been evaluated as per RMSE (9.28) and MAE (7.37). The authors of this study [6] proposed a method that uses a deep and shallow model to classify and estimate depression from text and video descriptors. A deep CNN model for detecting depression is presented based on audio and visual characteristics. The model also used a random forest for depression classification and a paragraph vector and support vector machine (SVM)-based model to assess the patients' mental and physical health. The AVEC-2016 dataset was used in the experiments, and the F1-measure was 0.746.

The authors of this paper [18] proposed a method for obtaining a multimodal representation of depression indicators for individual depression level prediction using a spatiotemporal attention network (STA) and a multimodal attention feature fusion (MAFF). The author visualizes audio data by splitting the voice amplitude into fixed-length segments and putting these segments into an STA network that uses an action mechanism to integrate spatial and temporal information. To evaluate how each dimension of the audio-visual segment feature has changed, the EIGEN evolution pooling approach was used in this model. The MAFF is also processed using support vector regression. Experiments were conducted on the AVEC-2013 and AVEC-2014 depression databases, with RMSE of 8.16 and MAE of 6.14. In another study [19], the authors suggested a model for estimating depression levels from video data using a two-stream deep spatiotemporal network. To extract spatial information, the model used the Inception-ResNet-V2 network.

Furthermore, the system uses dynamic feature descriptors based on volume local directional number (VLDN) to capture facial movements. To gain more discriminative features, the feature map obtained from VLDN is fed into a CNN. The model also gathers temporal information by incorporating the temporal median pooling (TMP) technique through a multilayer bidirectional long short-term memory (BI-LSTM). The TMP method was used on spatial and temporal feature temporal fragments. The experiment is conducted on the AVEC-2013 and AVEC-2014 datasets, with an RMSE of 8.93 and an MAE of 7.04.

3 The Proposed Methodology

The proposed methodology consists of several steps to classify depression from facial images.

3.1 Dataset

The dataset consisting of 3–5 min' videos of each depressed patient was taken from the Department of Psychiatry, Bahawal Victoria Hospital, Bahawalpur. All the patients have undergone a consent form by providing their informed consent to take videos when conducting interviews based on PHQ-8. There were 180 depressed patients, as per a clinical trial conducted by a team of psychiatrists, and 200 healthy controls (normal). During the acquisition of the videos, interviews based on the questionnaire prepared according to the PHQ-9 and DSM-V were conducted. The details of the patients and healthy controls are shown in Tab. 1.

Table 1: Details of patients and healthy controls

Controls	No.	Gender	Age group	Marital status	Job occupation
Patient	10	Male	22–39	Unmarried	Job holder
Patient	20	Male	32–43	Married	Job holder
Patient	20	Male	40–55	Married	Job holder
Patient	10	Male	50–64	Married	Job holder
Patient	40	Female	20–35	Married	Housewife
Patient	30	Female	30–45	Married	Job holder
Patient	20	Female	40–55	Married	Housewife
Patient	20	Female	50–65	Married	Housewife
Patient	10	Female	50–70	Married	Housewife
Normal	70	Male	20–40	Unmarried	Job holder
Normal	10	Female	20–55	Unmarried	Job holder
Normal	20	Male	30–45	Married	Job holder
Normal	30	Female	40–60	Married	Housewife
Normal	40	Male	45–65	Married	Job holder
Normal	30	Female	50–65	Married	Housewife

3.2 Preprocessing

The recorded videos were converted to color frames of dimensions 256 × 256 each. The average time of each video is 3 to 5 min, with 24 frames per second. Approximately 1,036,800 depressed and 1,152,000 healthy controls images were obtained from these videos. This was a massive dataset concerning the computations involved during its analysis. So, another light-weighted version of this dataset was also obtained, where one frame out of two consecutive frames was part of this version of the dataset. Since micro expressions are transient and subtle movements, we may lose these expressions if we drop more frames. This justifies that we cannot drop more frames to reduce the volume.

3.3 Features Extraction

To have noise-free information, the OpenFace tool [20] has been used for aligned faces, 2D landmarking, and AUs detection and their estimation. According to [11], AU's are based on micro-expressions during sadness, fear, disgust, contempt, and anger. Hence, their analysis revealed 13 filters. The obtained aligned faces, (samples shown in Fig. 1) are saved in the JPEG of dimensions 256 × 256 each. To have landmarks, each of the frames is cropped and aligned using the mouth, ears, and eyes facial landmarks. Multiple features, including AUs, gaze, pose, and landmarks, have been obtained and saved in feature files (CSV format) for further experiments to learn their contributing patterns.

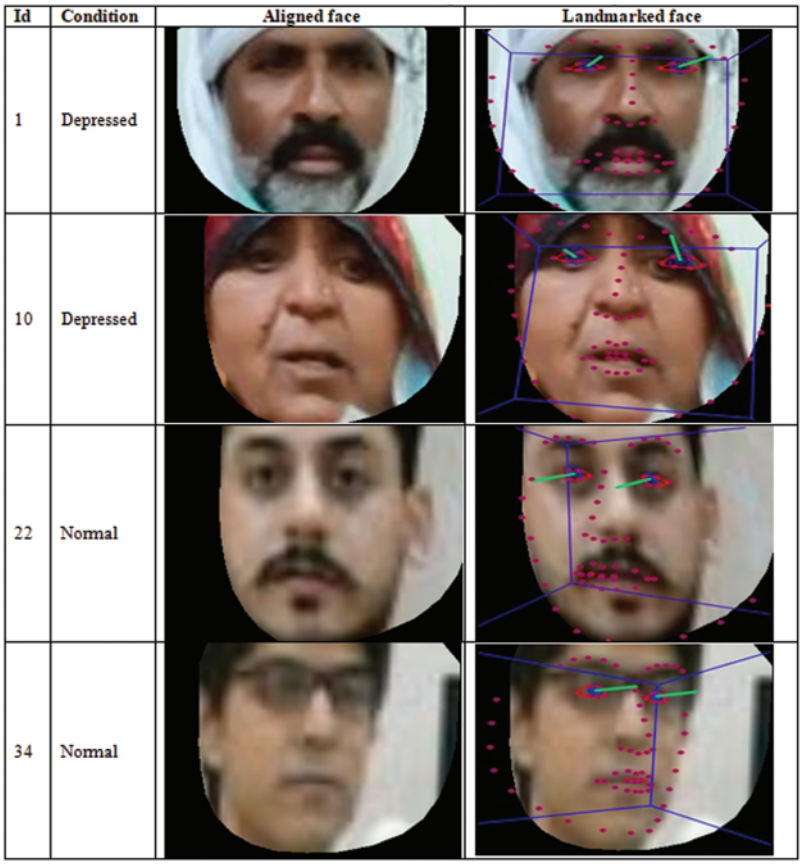

Figure 1: Aligned and landmarked faces of sample patients

3.4 The Proposed CNN Architecture

CNN is a deep neural network that works on images and is applicable in every machine vision-based problem. It analyzes the image and maps it into further processable form for better prediction without losing essential features. CNN comprises two significant steps, i.e., convolution and pooling. The convolution process works as a feature extraction layer, which extracts features to preserve spatial information. The pooling layer minimizes the dimensions of the substantial features according to the application. Pooling functions are categorized into max-pooling and average-pooling. The reduced features are then fed into a fully connected layer that consists of an activation function.

3.4.1 Convolutional Layer

The component that carries out "convolution" in the convolutional layer is the kernel/filter. This component performs multiplication between the input data and the two-dimensional array of weights in the case of a two-dimensional image. The convolutional layer is responsible for extracting low-level features, including edges, color, and gradient orientation. More layers are added to get the high-level features comprising micro-expressions that can easily identify the severity and type of depression.

Consequently, a feature map is created for further processing. The convolutional process is shown in Eq. (1).

$$conv(I,K)_{x,y} = \sum_{i=1}^{n_h} \sum_{j=1}^{n_w} \sum_{k=1}^{n_c} K_{i,j,k} I_{x+i-1,\, y+j-1,k} \tag{1}$$

where I = image, K = filter/kernel, x = x-coordinate, y = y-coordinate, n_h = height of image, n_w = width of image, and n_c = the number of channels.

3.4.2 Pooling Layer

In this layer, the spatial size of the convolved feature is reduced for managing the computing power to process the massive data through minimalized dimensions. Moreover, it is necessary for extracting the dominant features that are position or rotation invariant. This helps in the effective processing of the model. There are two types of pooling, i.e., max-pooling and average-pooling. In max-pooling, the maximum value is returned by the potion of an image covered by the kernel, while in average-pooling, the average of all numbers from the portion of an image is given, covered by the kernel. The performance of max-pooling is better than the average-pooling; therefore, max-pooling has been used in the proposed architecture.

3.4.3 RELU Layer

The activation function converts the data from linear to nonlinear form. It is essential to convert data into non-linear forms to solve complex problems. Common activation functions include tanh, sigmoid, softmax, Rectified Linear Unit (RELU), and Exponential Linear Unit (ELU). During feature extraction by the proposed CNN model, rectified linear function is used to increase the non-linearity to learn the complex relationship in data. RELU can minimize the interaction effects as it returns 0 for negative values. This function act as a linear and non-linear function for two halves of the input data. It is primarily implemented in the hidden layer due to the formation of dead neurons. RELU function is computationally efficient because it has zero derivatives for negative numbers and 1 for positive numbers, and due to this, some of the neurons get activated. Its function is shown in Eq. (2).

$$f(x) = max(0,\ x) \tag{2}$$

3.4.4 Fully Connected Layer

The pooled feature map is flattened before setting foot into the fully connected layer. The image is flattened into a column vector suitable for multi-level perceptron and fed to the fully connected layer. Then back-propagation applies to each iteration of training and the fully connected layer is responsible for learning a non-linear combination of high-level features. A fully connected layer has all the connections with the activation unit of previous layers. Over a series of epochs, the model can classify dominating and recessive features using the Softmax Classification technique. It allocates the values to the feature map vectors for each category to normalize them. Normalization is done to obtain a mean close to zero that accelerates the process of learning and is a prime factor for faster convergence. Convergence is the state when the neural network has reached a constant learning rate and does not improve further.

3.4.5 Sigmoid Function

It is a feed-forward and non-linear activation function defined for real input values. It is used for predicting probability-based outputs. Its value lies between 0 and 1. It is a smoothing function that facilitates derivation and is suitable for classification. The derivation is needed in the neural network for calculating gradients. To optimize the neural network's learning with the constant derivative, it is impossible to calculate the optimal parameters. It has a vanishing gradient problem and is not zero-centered. Therefore, the learning is minimal and time-consuming. Mainly, it is used in the output layer and is better for binary classification. The sigmoid function is shown in Eq. (3). In the proposed CNN architecture, this function has been used for classification.

$$S(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

3.5 Hyperparameters of the CNN Architecture

3.5.1 Dropout

This hyperparameter of CNN architecture is used to prevent over-fitting. This makes the model efficient and more generalized. Its value lies between 0.0 and 0.9. Nodes are emitted temporarily from the neural network with all its incoming and outgoing connections according to the provided fixed probability value for thinning of the network weights. In the proposed CNN architecture, the dropout value has been fixed empirically as 0.5.

3.5.2 Batch Size

In gradient descent, the batch size is a hyperparameter that refers to the number of samples in the forward/backward pass before appraising the internal model limitations. The batch size may be 32, 64, 128, 256, or 512. In the proposed CNN architecture, the batch size value has been fixed as 128.

3.5.3 Momentum

Momentum is used to overcome the noisy gradient or bounce the gradient by accelerating the movement of search in a direction to build inertia (constant movement). For the optimal performance of the gradient, descent momentum is used. It generally reduces the error and speeds up the performance of the learning algorithm. It prevents hedging of the optimization process. It helps the neural nets get out of the local minima point to find the global minimum. Mostly the momentum values are chosen close to 1 such as 0.9 or 0.99. In the proposed CNN architecture, the value of momentum has been fixed empirically as 0.9.

3.5.4 Learning Rate

The learning rate is augmented or abated concerning the error gradient by adjusting the neural network's weights. It is responsible for a stable and smooth training and is the most important hyperparameter of gradient descent. Learning rate effects, the accuracy, and time required to train the model. The learning rate lies between 0.0 and 1.0. In the proposed CNN architecture, the learning rate value has been used as 0.001.

The proposed light-weighted CNN architecture is shown in Fig. 2, while an overview of the proposed approach is demonstrated in Fig. 3.

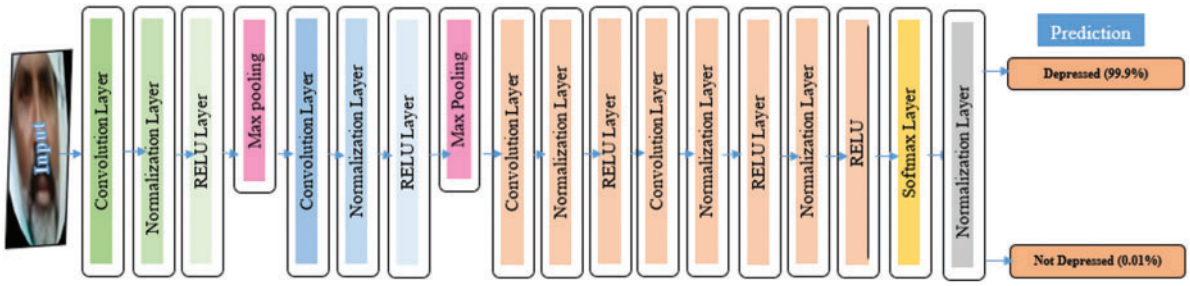

Figure 2: The proposed CNN architecture for depression detection

4 Results and Discussion

To have subsamples from the video given to the system, we extracted frames to discard one frame of two consecutive frames. So, a total of 1,09,440 data frames (51,840 depressed and 57,600 normal) were obtained from 38 subjects' data. Preliminary, the experiments for depression detection from the whole frames were performed using pre-trained state-of-the-art CNN-based models, i.e., VGG16, VGG19, Alex-net, and Google-net and the proposed CNN model, but the model got over-turned due to the noise present in the frames. To remove noise, these frames are then incorporated into the OpenFace tool for aligned faces, 2D landmarks, AUs and gaze, pose, and landmarked features detection and extraction.

Aligned faces have been used for depression detection using the proposed and state-of-the-art CNN modes, i.e., VGG16, VGG19, Alex-net, and Google-net. AUs, gaze, pose, and landmarked based features have also been used to train a model using SVM. Moreover, to analyze, all these features were combined to verify their applicability to depression detection.

4.1 Results

70% of the data is used for training while the rest of 30% is divided equally into testing and validation. The results of all experiments are shown in Tab. 2. It is evident from Tab. 2 that only six pose features have obtained 95.4% validation accuracy when extracted from aligned images. Similarly, gaze features, 288 in number, gained lower classification rates. When only AUs features were extracted, we obtained 90% validation accuracy, which is reasonable. It is worth stating that features have an excellent classification rate, i.e., 98.5%. To verify whether all these features collectively improve accuracy, these were combined, which achieved a remarkable classification rate of 100%, maximum as it can be using SVM. All about SVM seems good; however, feature engineering is required. Therefore, aligned images are also classified using CNN-based models, which do not need features to extract explicitly.

When aligned facial images were classified through VGG16, VGG19, Alex-net, and Google-net obtained 94.67%, 96%, 96.23%, and 93.57% validation accuracy. This is also reasonable; however, all these pre-trained models have a large volume of learnable, i.e., parameters, which significantly increases the training time of a respective model. However, the proposed model consumes only 16 layers and 6.5 million parameters. Therefore, it could be considered as light-weighted and requires lesser training time, achieving 99% validation accuracy. The accuracy obtained through the proposed CNN model is significantly high compared to the pre-trained model. Therefore, it is more favorable for real-time depression detection from facial images. The accuracy and loss plots and confusion matrix of the results is shown in Fig. 4.

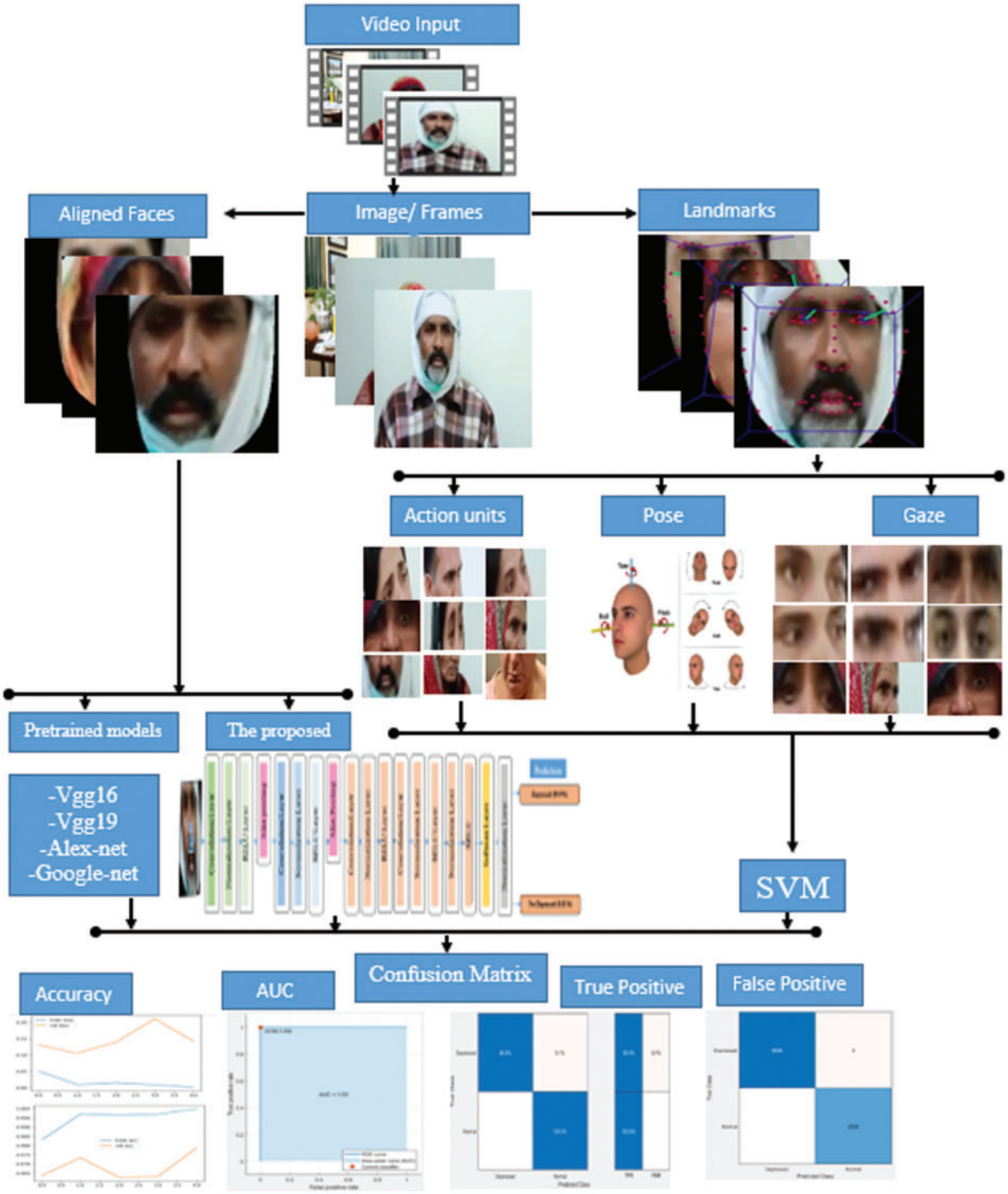

Figure 3: Overview of the proposed methodology

Table 2: Results of depression detection

Classification through	Features	Model trained through	Parameters (millions)	Testing accuracy	Validation accuracy	Remarks
Features	Pose (6 features)	SVM with fine gaussian	-	96.3	95.4	Validated
	Gaze (288 features)	SVM with fine gaussian	-	83.2	82.1	Validated
	AUs (35 features)	SVM with cubic kernel	-	91.2	90.0	Validated
	Landmark features (136 features)	SVM with fine gaussian	-	99.9	98.5	Validated
	Combined (Pose + Gaze + AUs + Landmark)	SVM with cubic kernel and fine gaussian	-	100	100	Validated
Frames	"Noisy" frames of the dataset	VGG16	138	100	100	Over-tuned
		VGG19	144	100	100	Over-tuned
		Alex-net	61	100	100	Over-tuned
		Google-net	7	100	100	Over-tuned
		CNN based model	6.5	100	100	Over-tuned
	Frames having aligned faces	VGG16	138	97.56	94.67	Validated
		VGG19	144	99.0	96.0	Validated
		Alex-net	61	98.10	96.23	Validated
		Google-net	7	95.46	93.57	Validated
		The proposed CNN model	6.5	**99.9**	**99.0**	Validated

4.2 Comparison of the Results with the State-of-the-Art

These results obtained through the proposed model are also compared with the recent studies of depression detection from videos or imagery data, as shown in Tab. 3.

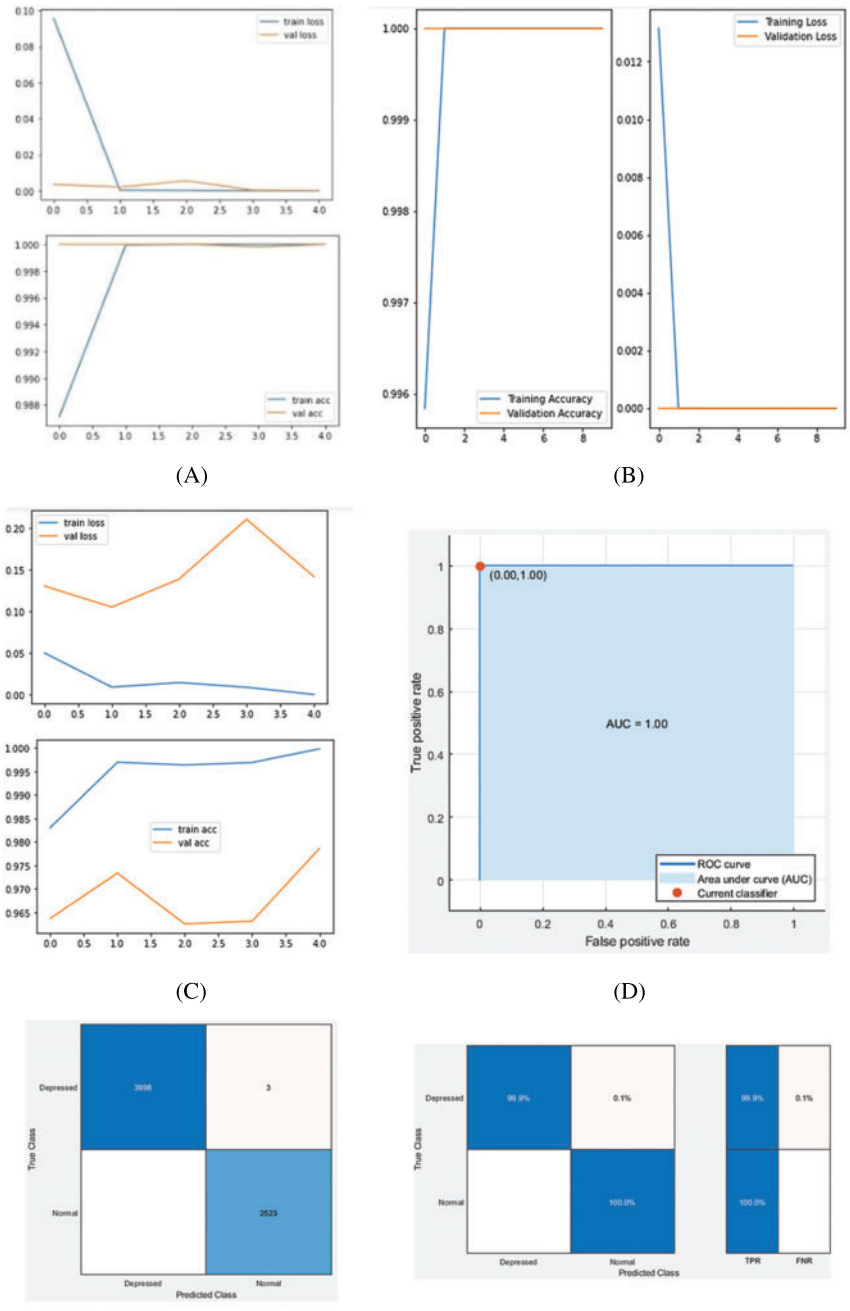

Figure 4: Accuracy and loss plots and confusion matrix of the results, (A) Accuracy and loss plot of VGG19 model when trained on noisy data, (B) Accuracy and loss plot of the proposed CNN based model when trained on noisy data, (C) Accuracy and loss plot of VGG19 model when trained on aligned faces, (D) ROC obtained on combined (Pose + Gaze + AUs + Landmark) features classified using SVM, (E) Confusion matrix of the results obtained through Landmark feature classified using SVM, (F) True positive rate of the results obtained through Landmark feature classified using SVM

Table 3: Comparison of the results with recent studies of depression detection

Related reference	Dataset used	Clinical methods	Methodology	Evaluation measures
[15]	57 depressed patients	Interviews were conducted using the Hamilton Rating Scale for Depression	ZFace + Stacked Denoising Autoencoders + Fisher Vector coding	Accuracy (78.67%)
[17]	AVEC-2013, AVEC-2014, Audio-Visual depressive language corpus	Depression Videos, BDI-II	Convolutional 3D network + RNN	MAE (7.22) RMSE (9.20)
[6]	AVEC-2016	Audio video based clinical trials	DCNN + PV-SVM + Histogram of Displacement Range	F1 (89%) Precision (91%) Recall (86%)
[16]	AVEC-2013 and AVEC-2014	Clinical interviews/Questionnaires	Median Robust Local Binary Patterns from Three Orthogonal Planes + Dirichlet Process Fisher Vector	MAE (7.55) RMSE (9.20)
[18]	AVEC-2013 and AVEC-2014	Depression Videos	Spatio-Temporal Attention network + Multimodal Attention Feature Fusion + Eigen Evolution Pooling + 2D CNN + LSTM	MAE (6.14) RMSE (8.16)
[19]	AVEC-2013 and AVEC-2014	Depression Videos	Inception-ResNet-v2 network + CNN + Bi-LSTM + Temporal median pooling + Deep spatiotemporal network	MAE (7.04) RMSE (8.93)

(Continued)

Table 3: Continued

Related reference	Dataset used	Clinical methods	Methodology	Evaluation measures
[21]	Locally developed	Depression Videos	Clinical methodology (self-rating depression scale) + 3DCNN	Accuracy (92.50%)
The proposed model	Locally developed	Clinical interviews/Questionnaires	SVM + CNN	Accuracy SVM (100%) CNN = **99.9%**

In the research [15], the SDAE was used for mapping between features and improved fisher vector coding for encoding the facial and head movements, overall, with an accuracy of 78.67%. This study is with low accuracy. In another research [17], 3D CNN and recurrent neural networks are used for depression detection from videos. It has been evaluated as per mean absolute error (7.22) and root mean square error (9.20). This study is with a high measure of errors. In another paper [6], a deep CNN, a paragraph vector, and SVM-based models are used on video features to predict patient mental health conditions. Overall, evaluation measures, i.e., F1, precision, and recall obtained are 89.2%, 91.7%, and 86.8%, respectively. Although precision is good, however, other achieved measures are low.

In another paper [16], the Dirichlet process fisher encoding scheme to aggregate the MRLB-TOP over an image sequence has been proposed. It has been evaluated as per mean absolute error (7.55) and root mean square error (9.20). This study is also with a high measure of errors. A study conducted by [18] proposed a framework that used STA and a MAFF to obtain the multimodal representation of depression cues to predict the individual depression level. It has also been evaluated as per mean absolute error and root mean square error, which got values of 6.14 and 8.16, respectively. This study is with a reasonable measure of errors. A framework for estimating the depression level from video data using a two-stream deep spatiotemporal network has been developed in another paper (Uddin et al., [19]). This model obtains temporal information by integrating the TMP approach through a multilayer BI-LSTM.

Moreover, the framework captures the facial motion using VLDN based dynamic feature descriptors. The feature map obtained from VLDN is nurtured into a CNN to obtain more discriminative features. It has been evaluated as per mean absolute error (7.04) and root mean square error (8.93). This study is with a reasonable measure of errors. This research [21] uses a software-developed camera (SDC) for depression detection. It generates the video based on the self-rating depression scale (SDS) and processes it using 3D CNN for redundancy-aware self-attention and local and temporal pattern exploration. The accuracy obtained by this model is about 92.5%. This accuracy can be further improved.

The proposed study refers to the employment of own designed CNN model, previously developed state-of-the-art CNN-based models, and SVM. Experiments have been performed in a variety of ways. The whole dataset containing noise is used to classify through pre-trained models and the reported one. Here, in both ways, the model gets overturned. Therefore, the dataset is incorporated into the OpenFace tool for the aligned faces and feature extraction. The aligned faces are then used for training

the proposed model that achieved 99% validation accuracy. Moreover, the extracted features have also been fed into the SVM model for training and testing with 100% accuracy.

5 Conclusion and Future Work

Depression is becoming a significant health disorder nowadays. Facial AUs have major clues for the detection of depressive disorders. Many studies have been conducted to detect depression through these visual clues, but their models do not perform well as per evaluation measures. Therefore, we proposed a multimodal depression detection system based on CNN and SVM, trained on a locally developed dataset. The experiments revealed 99.9% validation accuracy on the proposed CNN model, while extracted features obtained 100% accuracy on SVM. Moreover, the results proved the superiority of the reported approach over state-of-the-art methods. The proposed light-weighted CNN-based model with fewer layers and parameters is perfectly tuned, providing robust performance. Therefore, this model is more suitable for real-time application of depression detection from videos captured through digital cameras even having low resolution.

In the future, we aim to further improve to detect rest of the psychological disorders in real time environment from gait, spoken language, and body movement Moreover, using electroencephalogram (EEG) sensor, it is also aimed to diagnose these disorders in their early stage.

Acknowledgement: The researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project.

Funding Statement: The researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Marcus, M. T. Yasamy, M. v. van Ommeren, D. Chisholm and S. Saxena, "Depression: A global public health concern," 2012.

[2] B. E. Cohen, D. Edmondson and I. M. Kronish, "State of the art review: Depression, stress, anxiety, and cardiovascular disease," *American Journal of Hypertension*, vol. 28, no. 11, pp. 1295–1302, 2015.

[3] T. Field, "Postpartum depression effects on early interactions, parenting, and safety practices: A review," *Infant Behavior and Development*, vol. 33, no. 1, pp. 1–6, 2010.

[4] R. D. Caplan and K. W. Jones, "Effects of work load, role ambiguity, and type A personality on anxiety, depression, and heart rate," *Journal of Applied Psychology*, vol. 60, no. 6, pp. 713–719, 1975.

[5] J. Zhu, Z. Wang, T. Gong, S. Zeng, X. Li *et al.,* "An improved classification model for depression detection using EEG and eye tracking data," *IEEE Transactions on Nanobioscience*, vol. 19, no. 3, pp. 527–537, 2020.

[6] L. Yang, D. Jiang and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—Hybrid architectures," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 239–253, 2018.

[7] X. Zhan and R. Yu, "A window into the brain: Advances in psychiatric fMRI," *BioMed Research International*, vol. 2015, pp. 542467, 2015.

[8] G. Zhao and X. Li, "Automatic micro-expression analysis: Open challenges," *Frontiers in Psychology*, vol. 10, pp. 1833, 2019.

[9] J. Wojciechowski, M. Stolarski and G. Matthews, "Emotional intelligence and mismatching expressive and verbal messages: A contribution to detection of deception," *PLoS One*, vol. 9, no. 3, pp. e92570, 2014.

[10] C. Darwin, *The Expression of the Emotions in Man and Animals*. Chicago, IL, USA: The University of Chicago Press, 2015.
[11] E. Paul, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. 2nd ed., NY, USA: OWL Books, 2007.
[12] K. M. Goh, C. H. Ng, L. L. Lim and U. U. Sheikh, "Micro-expression recognition: An updated review of current trends, challenges and solutions," *The Visual Computer*, vol. 36, pp. 445–468, 2020.
[13] Y. -I. Tian, T. Kanade and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.
[14] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister *et al.,* "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 563–577, 2017.
[15] H. Dibeklioğlu, Z. Hammal and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, 2017.
[16] L. He, D. Jiang and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1476–1486, 2018.
[17] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 262–268, 2021.
[18] M. Niu, J. Tao, B. Liu, J. Huang and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Transactions on Affective Computing*, 2020.
[19] M. A. Uddin, J. B. Joolee and Y. -K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer Bi-LTSM," *IEEE Transactions on Affective Computing*, 2020.
[20] T. Baltrušaitis, P. Robinson and L. -P. Morency, "Openface: An open source facial behavior analysis toolkit," in *Proc. of the 2016 IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, pp. 1–10, 2016.
[21] W. Xie, L. Liang, Y. Lu, C. Wang, J. Shen *et al.,* "Interpreting depression from question-wise long-term video recording of SDS evaluation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 865–875, 2022.