He, Lang; Guo, Chenguang; Tiwari, Prayag; Su, Rui; Pandey, Hari Mohan; Dang, Wei

DepNet

Published in:
International Journal of Intelligent Systems

DOI:
10.1002/int.22704

Published: 01/07/2022

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
He, L., Guo, C., Tiwari, P., Su, R., Pandey, H. M., & Dang, W. (2022). DepNet: An automated industrial intelligent system using deep learning for video-based depression analysis. *International Journal of Intelligent Systems*, *37*(7), 3815-3835. https://doi.org/10.1002/int.22704

ORIGINAL ARTICLE

Journal Section

DepNet: An Automated Intelligent System using Deep Learning for Video-based Depression Analysis

Lang He[1*] | Chenguang Guo[2*] | Prayag Tiwari [3*†] | Rui Su [4*†] | Hari Mohan Pandey [5†] | Wei Dang[6]

[1]School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China

[1]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an 710121, Shaanxi, China

[1]Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an 710121, Shaanxi, China

[2]School of Electronics and Information, Northwestern Polytechnical University, Xi'an Shaanxi, China

[3]Department of Computer Science, Aalto University, Espoo, Finland

[4]School of Foreign Languages, Northwest University, Xi'an Shaanxi, China

[5]Department of Computer Science, Edge Hill University, Ormskirk, United Kingdom

[6]Xi'an Mental Health Center, Xi'an Shaanxi, China

Correspondence
Hari Mohan Pandey, Prayag Tiwari, Rui Su
Email: pandeyh@edgehill.ac.uk, prayag.tiwari@aalto.fi, sabrina@nwu.edu.cn

[*]Equally contributing authors.

[†]Corresponding author

Funding Information

As a common mental disorder, depression has attracted many researchers from affective computing field to estimate the severity of depression. However, existing approaches based on Deep Learning (DL) are mainly focused on single facial image without the consideration of the sequence information for predicting the depression scale. In this paper, an integrated framework, termed as DepNet, for automatic diagnosis of depression adopting facial images sequence from videos, is proposed. Specifically, several pretrained models are adopted to represent the Low Level features (LLF), while Feature Aggregation Module (FAM) is proposed to capture the high level characteristic informa-

tion for depression analysis. More importantly, the distinct characteristic of depression on faces can be mined to assist the clinicians to diagnose the severity of the depressed subjects. Multi-scale experiments, carried out on AVEC2013 and AVEC2014 databases have shown the excellent performance of the intelligent approach. The root mean square error (RMSE) between the predicted values and the BDI-II scores is respectively 9.17 and 9.01 on the two databases, which are lower that those of the state of the art video-based depression recognition methods.

KEYWORDS
Depression, Industrial intelligent system (IIS), Deep Learning (DL), Pattern recognition, Feature aggregation module (FAM)

1 | INTRODUCTION

Major depression disorder (MDD) (also simply known as depression) is highly prevalent all over the world nowadays. As a mental disorder, depression can affect people's career, life, study and so on in a bad way. According to the report of World Health Organization (WHO) in 2017, there are about 350 million depressive patients worldwide and depression will become the second leading cause of death by 2030 [1]. The symptoms of depression is complicated and diversified, but the pity is that the diagnostic methodologies are mostly depending on patient's self-report or the judgement by clinicians, both of which are subjective in nature, therefore, unavoidably bearing limitations to some extent. In some general cases, the clinicians adopt some common diagnosis approaches, such as the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-V) [2], the Beck Depression Inventory (BDI) [3], and the PHQ–8 [4], etc. To further support and help clinicians and psychologists to diagnose depression symptoms in a comprehensive way, various methods are proposed by researchers from the affective computing community and deep learning field.

In recent years, non-verbal behaviours, facial expressions activities [5], head pose and movement [6], eye and gaze activity [7], have also been highly indicated for predicting the depression scale. Previous works has illustrated that some subtle patterns are implied around the facial region [8]. In [9], the authors considered that dynamic activation of facial nonverbal behaviour is signifi-

cant for predicting the severity of depression. Therefore, the paper focuses on only the nonverbal behaviours around facial regions for depression analysis. From the perspective of machine learning, depression recognition problem can be taken as classification or regression issue. The goal of the study is to predict Beck Depression Inventory–II (BDI–II, Table 1) score of each video clip on AVEC2013 [10] and AVEC2014 [11] databases.

To estimate the scale of depression from facial images by videos and conventional methods includes the following steps: 1) feature extraction, 2) feature aggregation, and 3) regression (or classification). Feature extraction plays a significant role for depression recognition in video. Mining a robust and compacting feature descriptor is crucial and meaningful. For the time being, the feature extraction can be classified into two: hand-crafted [12], [13] and deep-learned [14], [15], [16], [17], [18], [19], [20]. For hand-crafted features, Local Binary Patterns from three orthogonal planes (LBP-TOP) feature descriptor has been considered effective for predicting the scale of depression [21]. Wen et al. [12] extracted the dynamic appearance using the Local Phase Quantisation from Three Orthogonal Planes (LPQ-TOP) features estimated from facial region sub-volumes to capture the temporal dynamics. Meanwhile, Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) is adopted as visual features for predicting the scale of depression [13]. However, there exists the problem that the above mentioned steps are independent for depression recognition.

To mine the characteristic representation of depression from frame-level dynamic feature descriptors, feature aggregation method is commonly adopted to estimate the depression scale, such as BoW, VLAD, FV [13]. After the feature aggregation process, some regression (classification) (e.g., Support Vector Regression (SVR)) technology are used to estimate the scale of depression severity. In effect, as a matter of fact, the above three steps are developed respectively, and integrated an ensemble framework for the overall estimation of depression. Although hand-crafted features have produced satisfactory results, there still exist some limitations, like labor consuming and subjectivity. For deep-learned features, Zhu et al. [14] introduce an architecture using 2D convolutional neural networks (2D-CNN), for automatic depression recognition. Also, Al Jazaery et al. [20] design a framework named Recurrent Neural Network-3D convolutional neural network (RNN-C3D) to model the local and global spatiotemporal information from consecutive face expressions for depression recognition. However, the deep learned method requires a great number of facial images to fine-tune the 2D-CNN or C3D model to predict the depression severity. In comparison with hand-crafted features, deep-learned features do not need domain knowledge, and obtain better performance for depression recognition. More importantly, 2D-CNN can learn and mine high-level characteristic patterns from facial regions, and deep learned features can be more discriminative and effective for predicting depression degree than hand-crafted features. Furthermore, the stand three individual procedures of feature extraction, feature aggregation, regression (classification) for predicting the scale of depression are integrated into a deep framework. Through the proposed framework, the performance of depression recognition could be improved readily in an integrated way, and the effort of manually designing can be obviously reduced. As aforementioned, the current deep learning methods need a great amount images to train the depression recognition models. Therefore, it is considered that there is an urgent requirement for designing an integrated framework for automated depression diagnosis system to address the above-mentioned issues, which is

TABLE 1 BDI–II Score Ranges and Depression Severity.

Ranges	Depression Severity
0 – 13	None or minimal
14 – 19	Mild
20 – 28	Moderate
29 – 63	Severe

the motivation of this paper.

The aim of the present paper is to propose a unified architecture for designing automatic protosystem capable of efficiently assessing depression severity (i.e., BDI–II Score) given an sampled facial image sequence from video. Based on the novel advances in deep learning field and facial expression analysis [22, 23, 24], it is proposed to use many deep learning approaches (DCNN, LSTM, etc) to learn high-level semantics features from the facial image sequences. More importantly, to overcome the issues with the shortage of small samples, we use a fine-tuning approach that adopts pre-trained networks SE-ResNet-50 (SENet for short) [25],[26]. Also, a novel Feature Aggregation Module (FAM) layer is proposed to capture long-term characteristic representation of depression. More importantly, the discriminative representations of depression on faces can be revealed to help clinicians to diagnose the severity of the depressed subjects.

1.1 | Contribution

The main contributions of this study can be concluded as follows:

1. An integrated and integrated and intelligent framework termed DepNet that effectively models facial dynamics information as a non-verbal behaviour measure for assessing the severity of depression.
2. To model the temporal dynamic discriminative representations from facial sequences, a novel Feature Aggregation Module (FAM) based on DCNN, is employed, which can aggregate low level features from video data and capture the temporal characteristic of the facial appearance and dynamics.
3. The proposed DepNet architecture is evaluated on two depression databases. Experimental results show that the proposed intelligent system significantly advance the recognition accuracy, when compared with the majority of the visual cue-based methods.

1.2 | Organization

The rest of the present paper is organized as follows. The background of depression is discussed in Section 2. The related works on visual-based depression assessment are discussed in Section 3. The proposed approach (DepNet) is described in Section 4. The used databases and the experiments

are shown in Section 5. In Section 6, discussions and conclusions are presented.

2 | BACKGROUND OF DEPRESSION

Depression is a mental disorder, which has been studied by many researchers in different fields. As for the definition of depression, it has not reached an agreement from different areas. The common definition of depression is coming from clinicians in clinical trials. Depression might be a disorder of the brain, but its harms aren't confined to the cranium. Prolonged depression has been linked with a slew of health problems, from impaired immune function to gastrointestinal dysfunction. It's also been linked with cardiovascular disease (CVD), even increasing the risk for heart attack and a disrupted heart rate.

For the diagnosis of depression, clinicians often use medication to assist depressed subjects. Also, some bio-makers (i.e., speech, gamma-amino butyric acid (GABA), etc) are adopted for assessing the severity of depression. Though the current diagnostic methods have achieved a great success, other methods are still called for to help the clinicians with assessing the severity of depression. Therefore, various studies have been proposed to assess the severity of depression (see Section 3).

3 | RELATED WORK

Automated methods for behavior analysis have been proposed for the assessment and understanding of depression. Eye gaze, head and body movements, facial expressions, posture, and gesture have been commonly used in depression recognition task. In particular, automatic face analysis for depression recognition has attracted attentions of many researchers from affective computing field.

To describe the dynamic facial appearance, various approaches using visual information are proposed. AVEC2013 [10] used Local Phase Quantisation (LPQ) [27] feature descriptor as visual features. For each video, the authors first pre-process the video images using face detection, fitting and alignment method to extract robust features. Then LPQ features are represented using many blocks around the facial regions. After the above procedures, the generated LPQ histograms are concatenated to obtain the facial features from frame sequences, and the Support Vector Regression (SVR) was used for estimating and prediction.

Local Binary Pattern (LBP) [28] based methods have been also widely used [29], [30]. For a gray scale image, the most fundamental LBP operator performed in a 3×3 pixel block, if the non-center pixels greater than center pixel, the threshold value is set to 1, otherwise the value is set to 0. After that, eight binary numbers are generated and computed to get a LBP value. The LBP descriptor which include different possible binary patterns correspond to each bin to generate a histogram to obtain a 256-dimensional texture descriptor.

In [30], the authors adopted LBP and Edge Orientation Histogram (EOH) as frame-based features and Motion History Histogram (MHH) to represent dynamic and discriminative pattern from videos. Among the LBP extensions, Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) [31],

and its variants, have been also adopted. LBP-TOP illustrates a video by calculating local binary pattern from three orthogonal planes (XY, XT and YT). The dynamic appearance descriptors provided an important performance improvement for facial expression analysis, as well as depression analysis in [32] and [21]. Similarly, LPQ has been extended to LPQ-TOP in [33, 34] and successfully adopted for depression recognition and analysis in [12], where the researchers extracted frame-based LPQ-TOP features, and used sparse coding to represent the video, along with a decision fusion scheme for predicting the depression scale. The Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [35], another variant of LBP-TOP, has been employed in [36, 32, 37, 38], and [39], to estimate the severity of depression from facial images. Local Gabor Binary Patterns (LGBP) are extracted by generating a list of Gabor magnitude response images, and then the LBP is adopted to each of them. To a certain degree, all above-mentioned approaches based on hand-crafted feature descriptors bear some limitations for diagnosis of depression. To design the hand-crafted features, a wide range knowledge related to the domain needs to be deeply interpreted. Meanwhile, the hand-crafted feature descriptors are difficult to learn and capture the discriminative representations of depression. In this scenario, Zhu et al. [14] proposed DCNN to represent both the facial appearance patterns and the dynamics to analysis the severity of depression. They carried out experiments on both AVEC2013 and AVEC2014 datasets, reported comparable results than other visual-based methods.

In [15], the authors proposed an artificial intelligent system based on audio and video cues. The feature dynamic history histogram (FDHH), low-level descriptors (LLD), are extracted, and fused to estimate BDI–II scores.

Based on the recent solutions [14], [15], [40], [41], it has been noticed that 2D-CNN played a vital part for depression analysis. However, these methods is explicitly take the many number of facial images for depression analysis.

In [42], the authors propose a new feature descriptor from each frame and introduce spectral heatmaps and spectral vectors to learn the discriminative representations based on action units (AU). The spectral representations are input into the convolution neural networks (CNN) and artificial neural networks (ANN) for predicting the depression scale. Extensive experiments are carried out on the two depression databases (i.e., AVEC2013 and AVEC2014), and obtained comparable performance in the depression recognition task.

In [43], the authors design a two-stream framework to model the spatiotemporal representations for depression recognition. The temporal median pooling (TMP) method is adopted to model some temporal patterns of the generated features via CNN. Lastly, experimental results of the two depression databases (i.e., AVEC2013 and AVEC2014) shown the proposed method's efficiency.

Meanwhile, these DCNN-based depression estimation methods are found difficult for clinicians to know the underlying information of feature learning, and the clinicians find it hard to define which patch of the facial region is discriminative for estimating the depression scale. As is mentioned above, the paper focuses on developing an automated framework, which can effectively and efficiently reveal the severity of depression.

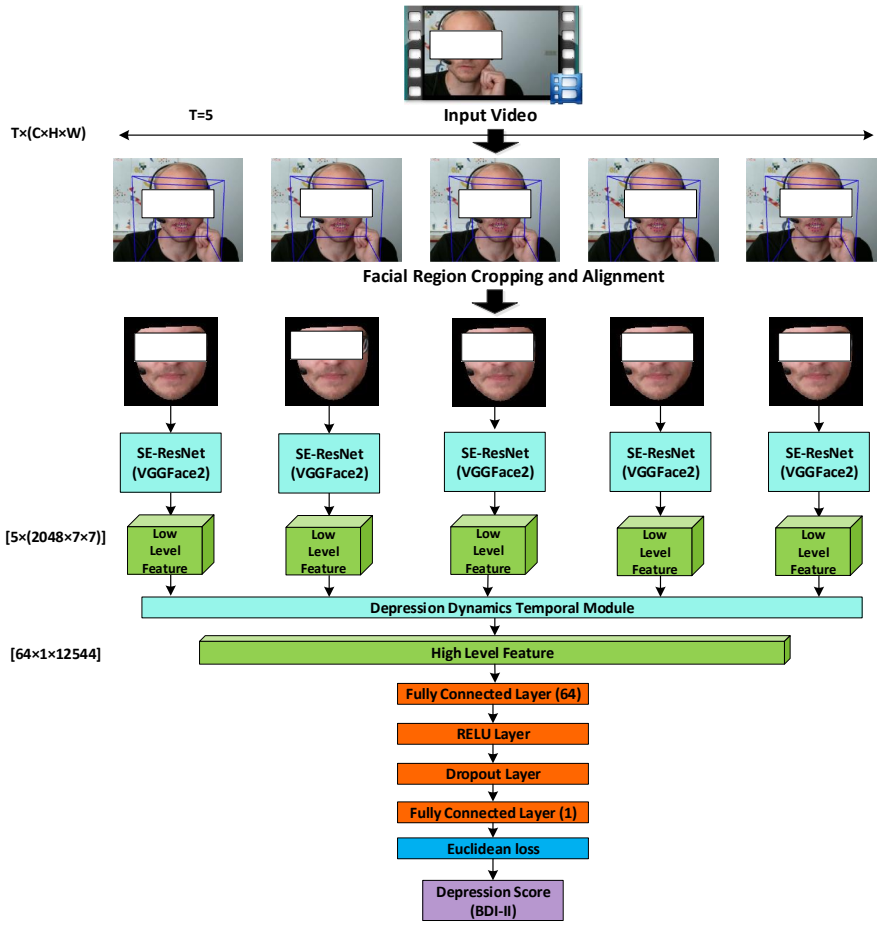

FIGURE 1 The pipeline of the proposed scheme for estimating the severity of depression. The dimensions of each layer are shown in square brackets on the left bracket.

4 | OUR APPROACH

The proposed depression recognition framework (i.e., DepNet based 2D-CNN) is illustrated in Figure. 1. We first adopt the OpenFace toolkit [44] to pre-process the facial images for feature extraction step. In order to generate a robust and discriminative representations, we first extract the Low Level features (LLF) from sampled facial images. Then we use FAM to learn and capture a robust and discriminative representation of High Level features (HLF) for each video clip. Lastly, the BDI–II score is predicted for every video clip.

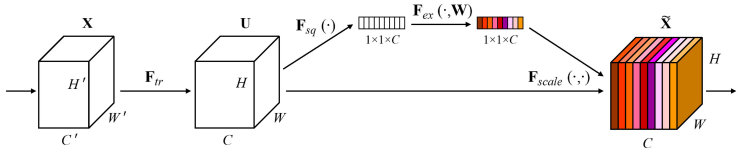

FIGURE 2 A Squeeze-and-Excitation block [25].

4.1 | Deep Depression Feature Representation by Transfer Learning

4.1.1 | SENet

The key component of SENet architecture is the SENet block. To make a clear description, in the following part, we first introduce the SENet block, and describe the proposed method. The architecture of the SENet block is showed in Fig. 2. Let $\mathbf{F}_{tr}$ be the transformation, which can map the image $\mathbf{X} \in \mathbb{R}^{H' \times W' \times C'}$ to feature map $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$. A *squeeze* operation is first performed on the features $\mathbf{U}$, and a channel feature descriptor is generated via aggregating feature maps at their spatial dimensions. The action of the feature descriptor is to generate an embedding of the global distribution of channel-wise feature responses, allowing information from the global receptive field of the network to be adopted by all its layers. Then an excitation operation is performed, which can use a self-gating mechanism to implement embedding as input and generate a collection of per-channel modulation weights. These weights are adopted on the feature map $\mathbf{U}$ to produce the output of the Squeeze-and-Excitation (SE) block which can be input into another subsequent layers of the network. Let $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_C]$ be the filter kernels, where v_C is the parameters of the c-th filter. Hence, the output can be written as $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_C]$, where

$$\mathbf{u}_c = \mathbf{v}_c * \mathbf{X} = \sum_{s=1}^{C'} \mathbf{v}_c^s * \mathbf{x}^s \tag{1}$$

Here $*$ represents convolution, $\mathbf{v}_c = [\mathbf{v}_c^1, \mathbf{v}_c^2, ..., \mathbf{v}_c^{C'}]$, $\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^c]$ and $\mathbf{u}_c \in \mathbb{R}^{H \times W}$. $\mathbf{v}_c^s$ represents a 2D spatial kernel that performs on a single channel of $\mathbf{v}_c$ of the corresponding channel of image $\mathbf{X}$. In this case, to better understand the notation, bias terms are ignored. After that, a summation operation is performed on all channels to generate the output. The relationships among channels are represented by convolution operations. Two steps, i.e., squeeze and excitation, are explained in the following.

(1) Squeeze: Global Information Embedding

To address the problem of investigating channel dependencies, the signal to every channel from the output features are considered. Each of the learned filters performs with a local receptive field, and consequently each unit of the transformation output $\mathbf{U}$ is unable to investigate contextual information outside of this region.

To further overcome this problem, *squeeze* operation is performed to aggregate global spatial information into a channel descriptor. Therefore, global average pooling is used to produce channel-

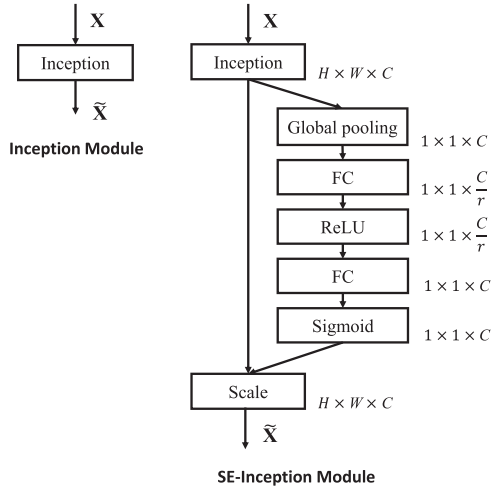

FIGURE 3 The comparison with the Inception module and the SE-Inception module [25].

wise features. In maths, a statistic $\mathbf{z} \in \mathbb{R}^C$ is produced via shrinking $\mathbf{U}$ at the spatial direction $H \times W$. Hence, the c-th operater of $\mathbf{z}$ can be written as:

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{2}$$

(2) Excitation: Adaptive Recalibration

To further use the patterns aggregated based on the *squeeze* operation, channel-wise dependencies are captured via a second operation *excitation*. To achieve this goal, the operation should meet the following standard: 1) it should be flexible; 2) it must study a non-mutually-exclusive relationship. After that, a gating mechanism with a sigmoid activation is adopted to met the criteria, which can be written as:

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 z)) \tag{3}$$

where σ is the ReLU function, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$,$\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. The final output of the block can be written as:

$$\widetilde{\mathbf{x}}_c = \mathbf{F}_{scale}(\mathbf{u}_c, s_c) = s_c \mathbf{u}_c \tag{4}$$

where $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, ..., \widetilde{\mathbf{x}}_C]$, and $\mathbf{F}_{scale}(\mathbf{u}_c, s_c)$ represents the channel-wise multiplication between the scalar s_c and the feature map $\mathbf{u}_c \in \mathbb{R}^{H \times W}$.

(3) Instantiations

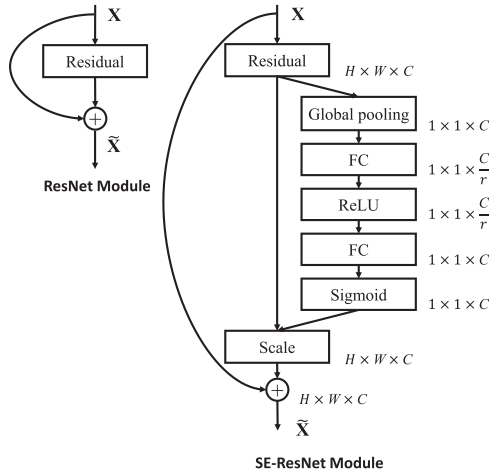

FIGURE 4 The comparison with the Residual module and the SE-ResNet module [25].

Output size	ResNet-50	SE-ResNet-50	SE-ResNeXt-50 (32 × 4d)
112×112	conv, 7×7, 64, stride 2		
56×56	max pool, 3×3, stride 2		
	$\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \\ fc, [16, 256] \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \quad C = 32 \\ \text{conv}, 1 \times 1, 256 \\ fc, [16, 256] \end{bmatrix} \times 3$
28×28	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \\ fc, [32, 512] \end{bmatrix} \times 4$	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \quad C = 32 \\ \text{conv}, 1 \times 1, 512 \\ fc, [32, 512] \end{bmatrix} \times 4$
14×14	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 1024 \\ fc, [64, 1024] \end{bmatrix} \times 6$	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \quad C = 32 \\ \text{conv}, 1 \times 1, 1024 \\ fc, [64, 1024] \end{bmatrix} \times 6$
7×7	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \\ fc, [128, 2048] \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 1024 \\ \text{conv}, 3 \times 3, 1024 \quad C = 32 \\ \text{conv}, 1 \times 1, 2048 \\ fc, [128, 2048] \end{bmatrix} \times 3$
1×1	global average pool, 1000-d fc, softmax		

FIGURE 5 (Left) ResNet-50. (Middle) SE-ResNet-50. (Right) SE-ResNeXt-50 with a $32 \times 4d$ template [25].

As is mentioned above, the SE block can be inserted into the common used architecture, e.g., VGGNet [45]. To illustrate the meaning, two examples are shown in Fig. 3 and Fig. 4. The transformation $\mathbf{F}_t r$ are adopted to be an entire Inception module in the architecture, as illustrated in Fig. 3. Also, SE blocks can also be adopted with residual networks, as shown in Fig. 4. Moreover, various variants are also integrated with SE-blocks, such as ResNeXt [46], Inception-ResNet [47], MobileNet [48] and ShuffleNet [49]. To further understand the SENet architectures, SEResNet-50 and SE-ResNeXt-50 are illustrated in Fig. 5.

To make an integrated, automated, and intelligent depression diagnosis framework from video, we adopt 2D-CNN, which has been used comprehensively in a great number of computer vision

scenes (e.g., face recognition [14], image classification [50], etc). Most importantly, as a popular used technology, 2D-CNN can perform a robust feature extraction, learn a comprehensive feature representation, and achieve a promising prediction performance. As a commonly adopted neural network technique, 2D-CNN requires numerous and various type of samples for training. However, the database the paper used is very small and not diverse (see Section 5.1). This is because the privacy issues and database collection need huge investment in time, human resource, and cost as well. Very few training samples make the depression model learning by 2D-CNN to be overfitting. Moreover, existing works of 2D-CNN only used various facial images to single model (i.e., VGG [15], [14], etc.) to predict the depression scale.

To tackle the problems, a novel pre-trained 2D-CNN model, SENet [25], is pre-trained on a large-scale face dataset named VGGFace2 [26], as a base model for feature learning of depression. The VGGFace2 dataset contains total 3.31 million images of 9131 subjects, with a large extent from VGGFace. Specifically, the last convolutional layers of the SENet is adopted to modify the full connected layers in our work. Different from the traditional SENet technology that uses one facial image as input for facial recognition, the input of the SENet is proposed to receive video frame sequences as input data. In our work, T images are sampled randomly from a video clip. And then the images are cropped and aligned. Finally, each of facial image is fed to the SENet and perform feature extraction at the last convolutional layer. To capture and aggregate facial dynamic temporal features based on LLFs, a novel dynamics feature extraction layer FAM is proposed, which consists of convolutional and pooling layers. A detailed description of these databases is referred to Section 4.2. After above process, a new discriminative feature representation, namely, HLF is generated, followed by two fully connected layers, a dropout layer, a rectified linear units (ReLU) layer. As a deep network, the loss function plays a significant role for the final classification or regression. In our task, depression analysis can be regarded as a regression issue. Therefore, Euclidean loss is used as the loss function, which is considered as suitable for our work. Formally, the Euclidean loss function L calculates the sum of squared differences between predicted and ground truth values, which can be expressed as:

$$L = \frac{1}{2M} \sum_{i=1}^{M} \| \hat{p}_i - p_i \|^2 \tag{5}$$

where M represents the number of samples, $\hat{p}_i$ is the predicted value of the network, and p_i denotes as the label (BDI–II score).

Finally, we fine-tune the proposed DepNet for automated depression diagnosis. The fine-tuning technology is commonly known as Transfer Learning [51], and is an effective solution for depression analysis [14]. The overall framework is illustrated in Figure. 1.

4.2 | Feature Aggregation Module (FAM)

In order to avoid overfitting of the network, and to capture the long term patterns "encoded" in the facial images, FAM module is proposed to be used for aggregating the discriminative feature representation from LLF to HLF. To prevent overfitting, the number of parameters is first decreased via reducing the dimensionality of the input channels of the convolutional layers. Here, some of the dimension reduction layers are compared, and the max-pooling layer of the channel direction (C. Max Pool) obtains the best performance for the final prediction of depression. Second, filters of various temporal scales are used to learn various temporal dynamics features. To extract the discriminative temporal features, different sized of convolutional filters (i.e., 3×1, 5×1, and 9×1) and pooling operation layers (max, average) are used. Based on [52], three temporal convolutional layers are designed for capturing global motion patterns in video sample. Meanwhile, inspired from [53], two temporal convolutional layers are adopted to extract statistical characteristic related to depression.

In Figure. 6, a visually illustration for FAM is provided. The module includes three parts: (i) dimension reduction (C. Max Pool: max pooling in the channel direction), (ii) dynamics temporal feature extraction (three convolutional layers followed by one max-pooling layers, two pooling layers), and (iii) dimension reduction operation (one 1×1 convolutional layer).

In the following, a detailed introduction is made for FAM. First, every LLF [2048×7×7] is performed by C. Max Pool, the parameters is kernel_size with 4, stride with 4, and pad with 0. After the dimension reduction process, the dimension of each LLF is [256×7×7]. In the reshape and concatenation layer, the LLF is reshaped from 4D [1×256×7×7] feature to a 1D feature [1×12544], then the five (T=5) reshaped features are concatenated at the temporal direction to a 2D feature [5×12544]. For feature extraction in the dynamic temporal layer, the 2D feature is performed by three convolutional layers, and followed by two pooling layers. Note that all of the convolutional layers and pooling layers use a kernel at the the temporal direction, and with kernel size one at the spatial direction. The kernel size of the convolutional layer is 3×1, 5×1, and 9×1, respectively. For our task, the size of every kernel is effected via the time window size (T=5). The size of kernel for pooling layers is 5×1 to extract valuable feature representation implied in the overall time window. The three convolutional layers followed by the max pooling layers are devised to alter the output to 1D feature at the temporal direction. The size of stride is 1, and the size of pad is 0 in all of the layers. After above transformation process, all outputs are concatenated into a 3D feature. To create the feature maps with a small channel, a 1×1 convolutional layer is used. At last, a HLF [64×1×12544] is generated via the FAM.

5 | EXPERIMENTS

This section introduces the experimental performance of the proposed approach for depression recognition. The datasets used in the experiments is described in Section 5.1. In Section 5.2, we schematically detail the experimental setup and evaluation measures. Lastly, the evaluations of the introduced scheme are discussed in Section 5.3.

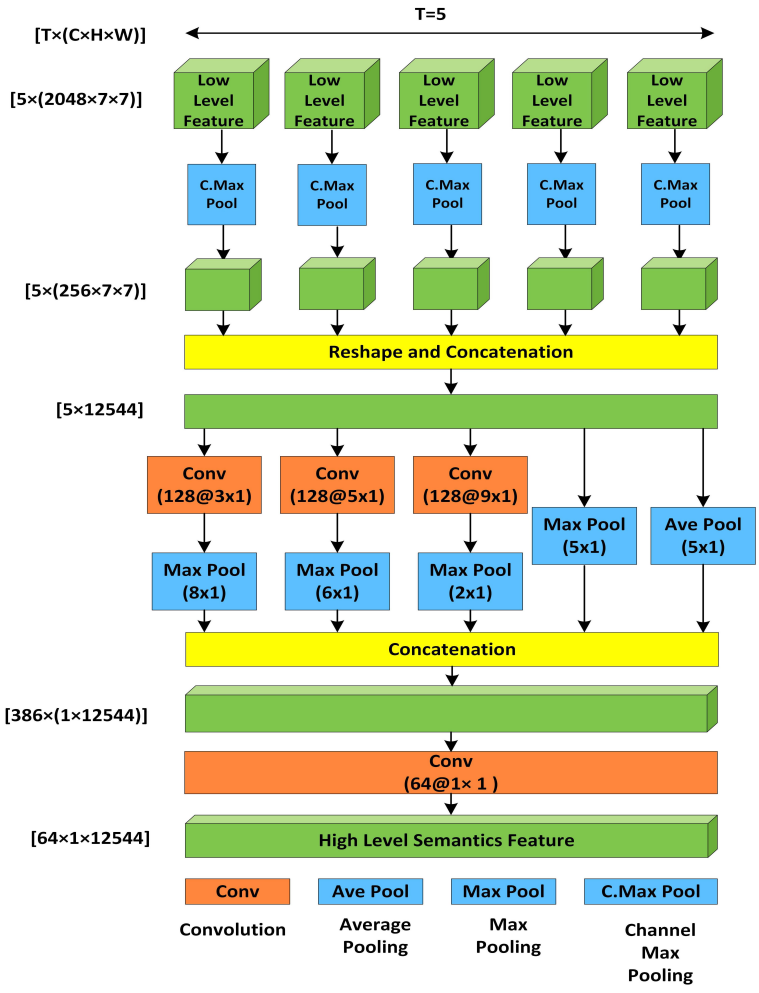

FIGURE 6 Schematic illustration of the FAM. Boxes and green cubes is the input/output features. Colored boxes is the intermediate layers. The numbers of the square brackets on the left denote the dimensions of the features of each layer.

5.1 | Databases

In our task, the evaluations of all of the experiments are carried out on the two publicly depression databases, i.e., AVEC2013 and AVEC2014. The average age of participant is 31.5 years with the range between 18 and 63 years. A webcam and a microphone are used for recording the audio and appearance signals. BDI-II is used as labeling for each sample.

In AVEC2013 depression corpus, there exists 150 video clips from 82 patients totally. The audiovisual recordings have been divided into three partitions by the publisher, i.e., training, devel-

opment, and test sets. For every partition, it has 50 recordings.

For the AVEC2014 depression corpus, there are two tasks included, i.e., Freeform and Northwind. For the two tasks, there are 150 video clips from 84 subjects. Similarly, AVEC2013 also has three partitions, i.e., training, development, and test sets. Therefore, a total of 100 samples are included in the partitions.

5.2 | Experimental Setup and Evaluation Measures

5.2.1 | Experimental Setup

On the AVEC2013 and AVEC2014 datasets, face detection and landmarks localization are first performed by OpenFace toolkit [44]. Then we crop the aligned facial images to the size of 224×224 with RGB color channels.

After the above process, the features of facial image sequences are represented for every video clip of the database. After that, we random sample T (T=5) images from each video clip, feed each facial image into the SENet, and extract the LLF at the convolutional layer. Totally, the AVEC2013 and AVEC2014 included 750 and 1,500 images, respectively.

The architecture of the DepNet is illustrated in Figure 1, which differs from the original SENet in that our DepNet adopt the facial image sequence as input for depression analysis. Only one neuron is set for giving the BDI–II score because depression recognition is a regression problem from the machine learning perspective. Specially, DepNet is fine-tuned on the VGGFace2 dataset [26] that contains a total of 3.31 million images of 9131 subjects, and fine-tuned on the AVEC2013 and AVEC2014 database. The Networks are trained by caffe deep learning toolbox [54] with Stochastic Gradient Descent (SGD). The learning rate is used to 0.00001, and reduced by the inverse decay rule with gamma of 0.0001, and power of 0.75. The momentum is set to 0.9, and weight decay is set to 0.0005. We conduct the experiments with two Titan-X GPU (each with 12G memory). In the fine-tuning procedure, we freeze the SENet, and train the rest of the DepNet (from the FAM to Euclidean loss in Figure. 1). In the training stage, an early-stop strategy is addopted to overcome the issue of overfitting. If the loss no longer decreased in two hundred consecutive iterations, the training process was stopped.

5.2.2 | Evaluation Measures

To make a fair comparison, the root mean square error (RMSE) and mean absolute error (MAE) are adopted to evaluate the capability of depression recognition methods, as shown in Equ.6 and Equ. 7, respectively, where M denotes the number of subjects, p_i is the label, and $\tilde{p}_i$ is the assessed value of the j-th subjects.

$$MAE = \frac{1}{M}\sum_{j=1}^{M}\left|p_j - \tilde{p}_j\right| \tag{6}$$

TABLE 2 Performance of deep architectures for visual-based depression diagnosis on the test set of AVEC2013.

Deep Architecture	RMSE	MAE
SENet	9.43	7.58
DepNet (SENet-based)	***9.17***	***7.36***
ResNet-50 (single image)	9.83	7.90
ResNet-50 (image sequence)	9.54	7.66

TABLE 3 Performance of deep architectures for visual-based depression diagnosis on the test set of AVEC2014.

Deep Architecture	RMSE	MAE
SENet	9.30	7.47
DepNet (SENet-based)	***9.03***	***7.26***
ResNet-50 (single image)	9.57	7.68
ResNet-50 (image sequence)	9.34	7.48

$$RMSE = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(p_j - \tilde{p}_j)^2} \tag{7}$$

5.3 | Experimental Results

To show the capability of DepNet, experiments are carried out on the AVEC2013 and AVEC2014 databases. Moreover, we conduct the experiments using SENet and DepNet to compare the performance of single facial images and facial image sequence, respectively.

5.3.1 | Overall Performance for Depression Recognition

The performances of depression analysis on the two databases (i.e., AVEC2013 and AVEC2014) are demonstrated in Table 2 and 3, respectively. Firstly, we fine-tune the SENet only using the facial images on the both databases. We retrain the proposed DepNet to use the facial image sequence in videos. Secondly, we compare the performance among the four deep architectures.

For AVEC2013, as shown in Table 2, one can see that, the performance of DepNet is the best among the four deep architectures (MAE 7.36 and RMSE 9.17) on the test set. The reason of comparison is that SENet is extended by ResNet. As for AVEC2014, as demonstrated in Table 3, similar

TABLE 4 Depression recognition results comparision with all the previous works on the test set of AVEC2013. Note that the listed methods use video data only.

Methods/Features	RMSE	MAE
Baseline [10]/LPQ-TOP	13.61	10.88
Cummins et al. [55]/STIP and PHOG	10.45	N/A
Meng et al. [30]/EOH and LBP	11.19	9.14
Wen et al. [12]/LPQ-TOP	10.27	8.22
Zhu et al. [14]/Optical flow	9.82	7.58
Mohamad et al. [20]/C3D, RNN	9.28	7.37
Zhou et al. [18]/2D-CNN	8.19	6.30
Song et al. [42]/CNN	8.10	6.16
Md et al. [43]/LSTM	8.93	7.04
Ours	9.17	7.36

to AVEC2013, DepNet outperforms the others (i.e., ResNet-50 (single image), ResNet-50 (image sequence), SENet, DepNet (SENet-based)), obtains in the MAE of 7.26, and RMSE of 9.03, for estimating the depression scale. In comparison with the results of AVEC2013, AVEC2014 gets comparable performances. This is because that AVEC2014 includes more data samples than AVEC2013 database for training the deep models. This important observation indicates that the temporal patterns is significant for predicting the scale of depression, and DepNet can encode the facial dynamics well. All in all, the performances on both datasets (AVEC2013 and AVEC2014) demonstrate that the availability of the proposed DepNet for depression recognition from facial image sequence.

5.3.2 | Comparison with Previous Works

In this part, we compare our capability of the proposed framework, using the proposed DepNet, with the previous algorithms using other visual features. Table 4 and Table 5 present the predicted value for the AVEC2013 and AVEC2014 databases.

Based on the two databases (i.e., AVEC2013 and AVEC2014), as shown in Table 4 and 5, our proposed framework obtains the comparable performances when compared with the-state-of-the-art approaches. The efficiency of our framework is that it can capture discriminative visual depression patterns from facial appearances. Some explanations are provided below. First, as an integrated framework, DepNet can better learn the behavior pattern than conventional approaches. It is proved that having a reasoning capability is important for automatically learning the characteristic "encoded" in facial expressions. Second, FAM can capture the dynamic temporal movements around facial regions.

In Figure. 7 and Figure. 8, we introduce our results on the two depression databases (i.e.,

TABLE 5 Depression recognition results comparision with all the previous works on the test set of AVEC2014. Note that these methods use video cues only.

Methods/Features	RMSE	MAE
Baseline [11]/LGBP-TOP	10.86	8.86
Sidorov et al. [56]/LGBP-TOP	10.83	8.32
Jan et al. [57]/EOH, LBP and LPQ	10.50	8.44
kaya et al. [58]/LGBP-TOP and LPQ	10.27	8.20
Zhu et al. [14]/Optical flow	9.55	7.47
Mohamad et al. [20]/C3D, RNN	9.20	7.22
Zhou et al. [18]/2D-CNN	8.39	6.21
Song et al. [42]/CNN	7.15	5.95
Md et al. [43]/LSTM	8.78	6.86
Ours	9.03	7.26

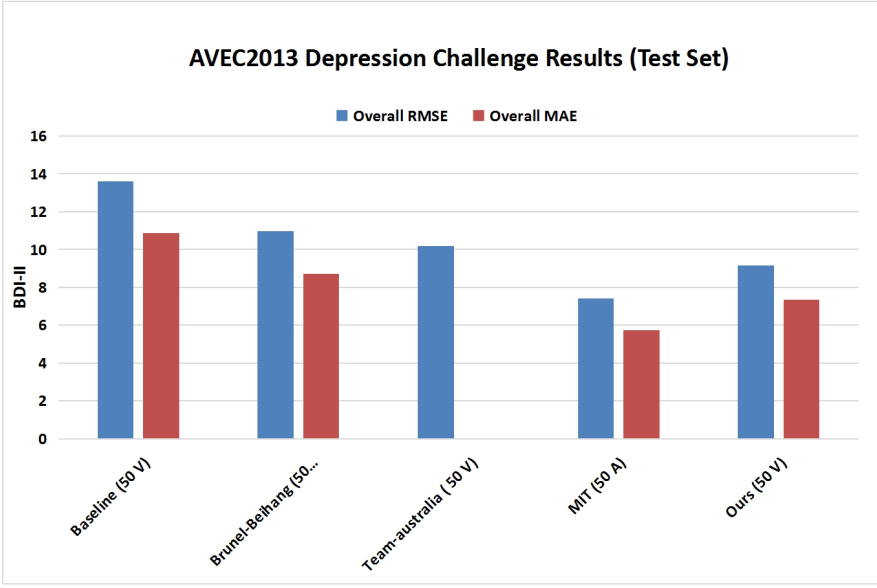

FIGURE 7 Comparison with techniques of depression recognition results in AVEC2013 challenge. Note that several of the listed methods use the audio data while our method only uses the visual data. (V) and (A) denote the video and audio data, respectively. (50 V) represents the 50 data samples used in the studies.

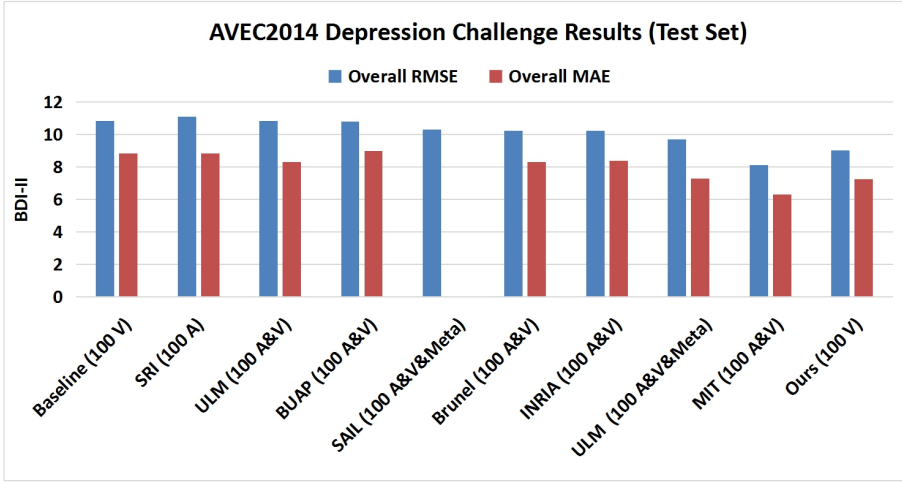

FIGURE 8 Comparison with techniques of depression recognition results in AVEC2014 challenge. (100 V) and (100 A) represent the 100 data samples used in the studies.

TABLE 6 The results of statistical significance test from the BDI-II prediction on the AVEC2013 database.

Statistical Significance Test Method	P-value
Shapiro-Wilk Normality Test	0.005
Dickey-Fuller Unit Root test	0.003
Analysis of Variance Test (ANOVA)	1.000
Chi-Squared Test	1.000
Mann-Whitney U Test	0.499

AVEC2013 and AVEC2014) and compare with the previous works using audiovisual modalities. It is noted that from the two figures that adopting only video cues, our framework obtains comparable performances to multi-modal methods for predicting the severity of depression. The observation further demonstrates the superiority of the proposed approach for predicting the scale of depression.

5.3.3 | Statistical Significance Test

Statistical tests have been performed to further illustrate the performance of the proposed method on the AVEC2013 and AVEC2014 databases. We conducted different tests which includes normality (Shapiro-Wilk Normality test), correlation (Chi-Squared test), stationary (Dickey-Fuller Unit

TABLE 7 The results of statistical significance test from the BDI-II prediction on the AVEC2014 database.

Statistical Significance Test Method	P-value
Shapiro-Wilk Normality Test	0.002
Dickey-Fuller Unit Root test	0.001
Analysis of Variance Test (ANOVA)	1.000
Chi-Squared Test	1.000
Mann-Whitney U Test	0.498

Root test), parametric (Analysis of Variance Test (ANOVA)) and non-parametric (Mann-Whitney U Test) tests. Table 6 and 7 respectively present the result statistics for AVEC2013 and AVEC2014 databases in the terms of p-value. This result statistics were achieved on 30 samples were collected from each of the algorithm. On the AVEC2014 database, from the P-value of 0.002 of the Shapiro-Wilk Normality test, we can conclude that the prediction of BDI-II is not Gaussian. From the Chi-Squared test, the P-value of 1.000, we can see that the prediction of BDI-II is dependent. For P-value of 0.001 of Dickey-Fuller Unit Root test, we can conclude that the prediction of BDI-II is stationary. Also, from the performance of ANOVA, the P-value of 1.000, then we can obtain the conclusion is that the prediction of BDI-II are from different distributions. Also, on the AVEC2013 database, we can obtain the same observations.The statistical tests result indicates that the proposed method has demonstrated superior results.

6 | CONCLUSIONS AND FUTURE WORKS

In this paper, an architecture named DepNet to capture temporal facial expressions dynamics, for depression analysis is proposed. We argue that such a reasoning ability is significant for capturing the characteristic patterns of depression "encoded" in facial expressions. It has been demonstrated that the proposed architecture outperforms the most of the approaches on the two databases, i.e., AVEC2013 and AVEC2014. Experimental results have supported such observation. From the experimental performances, the following major observations have been made :

1. The DepNet can model a discriminative depression pattern with visual explanation that benefits clinical diagnosis of depression severity in video sequence with fewer facial images.
2. To mine the important characteristic information of depression, FAM is proposed to aggregate LLF into HLF. Unlike the conventional convolutional and filter methods, FAM can capture high-level semantic information from facial image sequence.

However, the architecture has the limitation that the deep models contains amount of parameters to apply the industrial usage. In our future studies, various features using deep learning and

feature aggregation approaches based on deep learning are explored. Additionally, we will explore more explicable representation patterns, and more robust regression models to further promote the performance of depression analysis. More importantly, the proposed technology is proved to function well in assisting clinicians to assess the depressed subjects in a more effective way. We will collaborate with hospitals to use the proposed system to collect the depressed case and train the deep models for clinic usage. Our aim is that the clinicians can adopt the proposed system to help their diagnosis procedure. Furthermore, an attempt is made to adopt the introduced IIS in industrial domains. In addition, we will focus on adopting audio and video cues for multimodal depression recognition.

Acknowledgements

This work is supported by the Shaanxi Provincial Social Science Foundation (grant 2021K015), the Special Construction Fund for Key Disciplines of Shaanxi Provincial Higher Education, and the Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No. 20JG030). This work was supported by the Academy of Finland (grants 336033, 315896), Business Finland (grant 884/31/2018), and EU H2020 (grant 101016775).

Conflict of Interest

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

All authors contributed to the preparation of this manuscript. Lang He, Chenguang Guo, Prayag Tiwari, and Rui Su share equal contributions as first co-authors. Lang He, Chenguang Guo, and Prayag Tiwari contributed to the methodology, experiments, and writing the manuscript. Rui Su, Hari Mohan Pandey assisted in methodology, writing manuscript, statistical analysis and proof reading, and Wei Dang assisted in methodology, writing the manuscript, and proofreading.

ORCID

Lang He (https://orcid.org/0000-0003-2515-8579)
Chenguang Guo (https://orcid.org/0000-0002-5711-7977)
Prayag Tiwari (https://orcid.org/0000-0002-2851-4260)
Rui Su (https://orcid.org/0000-0002-4557-8215)
Hari Mohan Pandey (https://orcid.org/0000-0002-9128-068X)
Wei Dang (https://orcid.org/0000-0003-2477-8485)

references

[1] Mathers C, Fat DM, Boerma JT. The global burden of disease: 2004 update. *World Health Organization*; 2008.

[2] Bogduk N. Diagnostic and Statistical Manual of Mental Disorders. *American Psychiatric Association*,; 2013.

[3] Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients. *Journal of Personality Assessment* 1996;67(3):588–97.

[4] Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* 2009;114(1):163–173.

[5] Stratou G, Scherer S, Gratch J, Morency LP. Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences. In: *Humaine Association Conference on Affective Computing & Intelligent Interaction*; 2013. p. 147–152.

[6] Alghowinem S, Goecke R, Wagner M, Parkerx G, Breakspear M. Head pose and movement analysis as an indicator of depression. In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on IEEE*; 2013. p. 283–288.

[7] Alghowinem S, Goecke R, Wagner M, Parker G, Breakspear M. Eye movement analysis for depression detection. In: *Image Processing (ICIP), 2013 20th IEEE International Conference on IEEE*; 2013. p. 4220–4224.

[8] Girard JM, Cohn JF, Mahoor MH. Nonverbal social withdrawal in depression: evidence from manual and automatic analyses. *Image and Vision Computing* 2014;32(10):641–647.

[9] Ellgring H. Non-verbal communication in depression. Cambridge University Press; 2007.

[10] Valstar M, Schuller B, Smith K, Eyben F, Jiang B, Bilakhia S, et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: *Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge ACM*; 2013. p. 3–10.

[11] Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, et al. AVEC 2014: 3D dimensional affect and depression recognition challenge. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM*; 2014. p. 3–10.

[12] Wen L, Li X, Guo G, Zhu Y. Automated depression diagnosis based on facial dynamic analysis and sparse coding. *IEEE Transactions on Information Forensics and Security* 2015;10(7):1432–1441.

[13] He L, Jiang D, Sahli H. Multimodal depression recognition with dynamic visual and audio cues. In: *International Conference on Affective Computing and Intelligent Interaction*; 2015. p. 260–266.

[14] Zhu Y, Shang Y, Shao Z, Guo G. Automated Depression Diagnosis based on Deep Networks to Encode Facial Appearance and Dynamics. *IEEE Transactions on Affective Computing* 2017;.

[15] Jan A, Meng H, Gaus YFA, Zhang F. Artificial Intelligent System for Automatic Depression Level Analysis through Visual and Vocal Expressions. *IEEE Transactions on Cognitive and Developmental Systems* 2017;PP(99):1–1.

[16] Yang Y, Tan Z, Tiwari P, Pandey HM, Wan J, Lei Z, et al. Cascaded Split-and-Aggregate Learning with Feature Recombination for Pedestrian Attribute Recognition. *International Journal of Computer Vision* 2021;p. 1–14.

[17] Wang G, Yang Y, Zhang T, Cheng J, Hou Z, Tiwari P, et al. Cross-modality paired-images generation and augmentation for RGB-infrared person re-identification. *Neural Networks: the Official Journal of the International Neural Network Society* 2020;128:294–304.

[18] Zhou X, Jin K, Shang Y, Guo G. Visually Interpretable Representation Learning for Depression Recognition from Facial Images. *IEEE Transactions on Affective Computing* 2020;11(3):542–552.

[19] Wu J, Yang Y, Lei Z, Wang J, Li SZ, Tiwari P, et al. An end-to-end exemplar association for unsupervised person Re-identification. *Neural Networks* 2020;129:43–54.

[20] Al Jazaery M, Guo G. Video-Based Depression Level Analysis by Encoding Deep Spatiotemporal Features. *IEEE Transactions on Affective Computing* 2018;p. 1–1.

[21] Dhall A, Goecke R. A temporally piece-wise fisher vector approach for depression analysis. In: *Affective Computing & Intelligent Interfaces*; 2015. p. 255–259.

[22] He L, Niu M, Tiwari P, Marttinen P, Su R, Jiang J, et al. Deep Learning for Depression Recognition with Audiovisual Cues: A Review. arXiv preprint arXiv:210600610 2021;.

[23] Su H, Qi W, Hu Y, Karimi HR, Ferrigno G, De Momi E. An incremental learning framework for human-like redundancy optimization of anthropomorphic manipulators. *IEEE Transactions on Industrial Informatics* 2020;.

[24] He L, Guo C, Tiwari P, Pandey HM, Dang W. Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence. *International Journal of Intelligent Systems* 2021;.

[25] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. arXiv preprint arXiv:170901507 2017;7.

[26] Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: *IEEE International Conference on Automatic Face & Gesture Recognition*; 2018. p. 67–74.

[27] Ojansivu V, Rahtu E, Heikkila J. Rotation invariant local phase quantization for blur insensitive texture analysis. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on *IEEE*; 2008. p. 1–4.

[28] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence* 2002;24(7):971–987.

[29] Joshi J. An automated framework for depression analysis. In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on IEEE*; 2013. p. 630–635.

[30] Meng H, Huang D, Wang H, Yang H, AI-Shuraifi M, Wang Y. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In: *Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge ACM*; 2013. p. 21–30.

[31] Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2007;29(6):915–928.

[32] Jain V, Crowley JL, Dey AK, Lux A. Depression estimation using audiovisual features and fisher vector encoding. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM*; 2014. p. 87–91.

[33] Jiang B, Valstar MF, Pantic M. Action unit detection using sparse appearance descriptors in space-time video volumes. In: *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on IEEE*; 2011. p. 314–321.

[34] Jiang B, Valstar M, Martinez B, Pantic M. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Transactions on Cybernetics* 2014;44(2):161–174.

[35] Almaev TR, Valstar MF. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on IEEE*; 2013. p. 356–361.

[36] Gupta R, Malandrakis N, Xiao B, Guha T, Van Segbroeck M, Black M, et al. Multimodal prediction of affective dimensions and depression in human-computer interactions. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM*; 2014. p. 33–40.

[37] Perez H, Escalante HJ, Villasenor-Pineda L, et al. Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM*; 2014. p. 49–55.

[38] Senoussaoui M, Sarria-Paja M, Santos JF, Falk TH. Model Fusion for Multimodal Depression Classification and Level Detection. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM*; 2014. p. 57–63.

[39] Williamson JR, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD. Vocal and facial biomarkers of depression based on motor incoordination and timing. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM*; 2014. p. 65–72.

[40] Ma X, Yang H, Chen Q, Huang D, Wang Y. Depaudionet: An efficient deep model for audio based depression classification. In: *Proceedings of the 6th international workshop on audio/visual emotion challenge*; 2016. p. 35–42.

[41] He L, Chan JCW, Wang Z. Automatic depression recognition using CNN with attention mechanism from videos. *Neurocomputing* 2021;422:165–175.

[42] Song S, Jaiswal S, Shen L, Valstar M. Spectral Representation of Behaviour Primitives for Depression Analysis. *IEEE Transactions on Affective Computing* 2020;p. 1–1.

[43] Uddin MA, Joolee JB, Lee YK. Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. IEEE Transactions on Affective Computing 2020;p. 1–1.

[44] Baltrušaitis T, Robinson P, Morency LP. Openface: an open source facial behavior analysis toolkit. In: *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on IEEE*; 2016. p. 1–10.

[45] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* 2015;abs/1409.1556.

[46] Xie S, Girshick RB, Dollár P, Tu Z, He K. Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2017;p. 5987–5995.

[47] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* AAAI'17, AAAI Press; 2017. p. 4278–4284.

[48] Howard A, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv 2017;abs/1704.04861.

[49] Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2018;p. 6848–6856.

[50] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2016. p. 770–778.

[51] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: *Advances in neural information processing systems*; 2014. p. 3320–3328.

[52] Husain F, Dellen B, Torras C. Action recognition based on efficient deep feature learning in the spatio-temporal domain. *IEEE Robotics and Automation Letters* 2016;1(2):984–991.

[53] Ryoo MS, Rothrock B, Matthies L. Pooled motion features for first-person videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2015. p. 896–904.

[54] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2015. p. 1–9.

[55] Cummins N, Joshi J, Dhall A, Sethu V, Goecke R, Epps J. Diagnosis of depression by behavioural signals: a multimodal approach. In: *Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge ACM*; 2013. p. 11–20.

[56] Sidorov M, Minker W. Emotion Recognition and Depression Diagnosis by Acoustic and Visual Features: A Multimodal Approach. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM*; 2014. p. 81–86.

[57] Jan A, Meng H, Gaus YFA, Zhang F, Turabzadeh S. Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM*; 2014. p. 73–80.

[58] Kaya H, Çilli F, Salah AA. Ensemble CCA for continuous emotion prediction. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM*; 2014. p. 19–26.