Received: 25 October 2021 | Revised: 9 December 2021 | Accepted: 20 December 2021

DOI: 10.1111/exsy.12933

ORIGINAL ARTICLE

Expert Systems WILEY

Diagnosis of depression level using multimodal approaches using deep learning techniques with multiple selective features

Pratiksha Meshram[1] | Radha Krishna Rambola[2]

[1]Department of IT, SVKM's NMIMS, Mukesh Patel School of Technology Management & Engineering, Mumbai, India

[2]Department of CS, SVKM's NMIMS, Mukesh Patel School of Technology Management & Engineering, Mumbai, India

Correspondence
Pratiksha Meshram, Department of IT, SVKM's NMIMS, Mukesh Patel School of Technology Management & Engineering, Mumbai, India.
Email: pratiksha_mesh@yahoo.com

Abstract

Depression is a serious mental health condition that may lead to poor mental and emotional functioning at work, at school and in the family causing the mental imbalance. In worst scenarios, depression may lead to severe anxiety or suicide. Hence, it is necessary to diagnose depression at early stages. This paper elaborates the development of a novel approach for a convolutional neural network model that can examine facial images from the recorded interview sessions to discover facial patterns that could indicate depression level. The user-generated data helps to distinguish between different depressive groups with depression symptoms that can manifest people with various mental illnesses in different ways. In particular, we want to automatically predict the depression scale and differentiate depression from other mental disorders using the patient's psychiatric illness history and dynamic textual descriptions extracted from the user inputs. We apply the k-nearest neighbour algorithm on the dynamic textual descriptors to make a linguistic analysis for classifying mental illness into different classes. We apply dimensionality reduction and regression using the Random Forest algorithm to predict the depression scale. The proposed framework is an extension to pre-existing frameworks, replacing the handcrafted feature extraction technique with the deep feature extraction. The model performs 2.7% better than existing frameworks in facial detection and feature extraction.

KEYWORDS
depression, deep learning, emotion recognition, facial expression, regression

1 | INTRODUCTION

In India, depression is common in adults between the ages of 16–25, considered to be a leading cause of disability. For years, researchers have to identify and map the relationship between brain function and structure using neuroimaging data. The researchers at the University of Texas have identified a unique technique to categorize people susceptible to developing depression and anxiety using deep learning with a supercomputer. Now they are using the Stampede supercomputer at the Texas advanced computing centre to train deep learning algorithms that can identify similarities among hundreds of patients using magnetic resonance imaging genomics data and other factors to predict patients at risk of depression and psychological disorders (Valstar et al., 2013). In the past, the researchers have worked on the development of a model that takes raw text and audio segments as input and analyzes the wave-forms that predict depression. The work provides a method for deep learning-based segmentation to detect depression, as well as irregular segmentation and masks used for Gabor wavelength detection (Long et al, 2014). A Gabor filter, named after Dennis Gabor, is a linear local texture filter in image processing. It analyses if the model includes any specific frequency content in specific directions in a localized region around the point or region of analysis. The researchers trained the deep learning algorithm using extracted audio

and textual features from clinical patients with suicide attempts and higher rate of mood swings (Ma et al, 2016). Extreme Gradient boosting technique is used for identifying and categorizing the important parameters of depression and predicting depression cases by re-sampling methods using different balanced samples but the researchers were not able to precisely predict the depression scale. Also differentiating the mental disorders is a tedious task due to similarities in symptoms (Marcus et al., 2012).

Deep learning algorithms can help in identifying mental illness cases, biomarkers relevant to distinguish between them and identify cases of self-reported depression that are not reported in their section. Moreover, these biomarkers are reported on a case-by-case basis, so that we can predict mental illness and healthy cases using machine learning techniques. Deep learning is a machine learning and artificial intelligence (AI) method that replicates how humans acquire certain types of knowledge (Han et al, 2015; He et al, 2015). Deep learning algorithms are stacked in a hierarchy of increasing complexity and abstraction, contrasting typical machine learning algorithms, that are linear. To address this, most existing studies have reported that researchers have used tools that pre-process MRI data, such as MRI scanning tomography, to extract useful functions from the DL models fed with the data. However, all these tools extract traits based on prior knowledge, and therefore, some information relevant to predicting mental illness or healthy cases of depression may be omitted (Kessler & Bromet, 2013). For increasing the accuracy values for analysis of multi-model framework, cooperation was included in the analysis. The researchers believe that the use of text and images, which are user-generated data, helps to distinguish between different depressive groups and that depression symptoms can manifest in people with different mental illnesses in different ways. In particular, they investigate cases where people may express symptoms of depression outside the established criteria. Tweets can be commented on and used to detect feelings, which can help prevent the use of digitally rendered mathematical constructs as a diagnostic tool for depression. Words that struck us in those tweets were "depression," "dissociation" and "disagreement" (i.e. depression and anxiety). Such definitions have been adopted by the Institute for Health Metrics and Evaluation to analyse mental illness (Jenkins & Goldner, 2012). The researchers fed data from laboratory tests into the extreme gradient model to predict the outcomes of depression in healthy cases. After collecting tweets using a combination of social media analysis and a multi-modal model, we used the model to make predictions. There are several challenges in using the DL algorithm to analyse EEG data for the study of mental illness. They converted the h5py formatted data to TensorFlow, which they used for the classification experiments. As we deal with multiple mental illnesses, they have to face different challenges, such as the use of different types of data, the number of participants and the quality of the data. If researchers consider mental illness such as depression as a large category, they are unable to explain the many variations of depression and the combination of factors, which means that the structure of a model for monitoring depression is only vaguely defined. Social media data are typically identified and given the lack of data on social media use in the study of mental illness, there is no easy way to confirm the presence of a particular mental condition in many participants (Luxton, 2015).

To address the above-mentioned challenges, the researchers have integrated the electronic health record information along with facial features extracted from the video segments. The data set contains rich facial features with the fusion of multiple other features that differentiate the biological differences turned into biomarkers to help in protecting the depression scale. Deep learning and machine learning algorithms can make great strides in this area. Machine learning is a technique for constructing algorithms that take in input data and apply statistical analysis to predict the result based on the type of data provided. Computer Vision is one of the applications of Deep Learning algorithms. Image processing methods, Text Analysis & Understanding are the cornerstones of computer vision. Speech recognition, pattern recognition and autonomous cars are all aspects of speech recognition (Nasir et al, 2016; Williamson et al, 2013). This is just one example of a potentially exciting advance in predicting mental disorders by studying the relationships between brain structures by looking at neuroimaging data on brain function. In this research, we have developed a novel convolutional neural network model to examine facial images from the recorded interview sessions to discover facial patterns that could indicate depression level (Simonyan et al, 2015). The user-generated data helps to distinguish between different depressive groups with depression symptoms that can manifest people with various mental illnesses in different ways. Our model automatically detects the depression scale and differentiate depression from other mental disorders using the patient's psychiatric illness history and dynamic textual descriptions extracted from the user inputs. We have used the k-nearest neighbour algorithm on the dynamic textual descriptors to make a linguistic analysis for classifying mental illness into different classes. We apply dimensionality reduction and regression using the Random Forest algorithm to predict the depression scale. The proposed framework is an extension to pre-existing frameworks, replacing the handcrafted feature extraction technique with the deep feature extraction. The model performs 2.7% better than existing frameworks in facial detection and feature extraction. This is important for people who suffer from depression or who are unaware that they are experiencing depression or those who need treatment for depression. Early detection of depression can be crucial to tackle mental illness and providing support to people with terrible mental illness. The research paper has been overall summarized into five sections. The Section 2 of this paper discusses previous work, various approaches, research and pre-developed models by the researchers. Section 3 describes the dataset for video samples, the facial detection and feature extraction technique and regression technique used for prediction and classification, Section 4 discusses the results and general outcomes and in Section 5 we draw our conclusion describing the model.

The paper is structured as follows: the first section describes about the literature survey, second section describes about the design methodology, third section describes about the university ideological, fourth section is result and discussion, fifth section is conclusion and future work.

2 | LITERATURE SURVEY

In recent years, researchers have witnessed an evolutionary growth in the estimation of verbal intelligence and intellectual fitness estimation from facial expressions and audio recognition from their classification. They have developed a unique deep-learning-based computational technique to identify facial features from video data to analyse the sentiment variations among patients with psychiatric illness and classifying them according to their severity of illness (M. Kaletsch et al., 2014). In depression evaluation, their findings endorse that developing an automated depression prediction model is achievable using facial and audio feature extraction from video data using deep learning (Luxton, 2015). Researchers have explored versions inside the vocal speech of patients and located slight predictability of the depression ratings (Cun et al, 1990). They analysed each guide and automated facial expressions for a time interval from video data of a depressed patient. After analysis, they concluded that patients with high developing symptoms for depression show specific extra feelings related to depression, and smile much less. Researchers have tested the outcomes as a result of depression to more youthful patients of each gender. They fused several characteristic features such as eye moment, lip moment, eye-brow moment, head moment and spectral shape from video segments for calculating high amounts of precision and accuracy rate (Han et al., 2014). We fed data from laboratory tests into the extreme gradient model to predict the outcomes of depression in healthy cases. After collecting tweets using a combination of social media analysis and a multi-modal model, we used the model to make predictions. The researchers have analysed the correlation between the head and eye moment, eyebrow moment, lip moment, facial expression and audio segments and conclude that these features are considered to be the most important features for classifying the depression level among patients using AVEC 2016 and AVEC 2017 datasets resulted in better results (Cohn et al., 2009; Davies et al., 2016).

The researchers have reviewed facial recognition techniques that are used to extract facial features. They have used PCA and LBP as the handcrafted feature extraction technique. Even when every person only has one training image, the PCA + CNN and SOM + CNN methods are superior to the eigenfaces technique. When contrasted to PCA + CNN, the SOM + CNN method consistently beats it. The PCA is considered to be an orthogonal transformation procedure where a combination of observations are converted into the principal components. The initial not many head parts have the biggest fluctuation subsequently addressed pictures with modest number of highlights. LBP is a nearby paired example which encodes neighbourhood pictures into a twofold example. LBP endures against changes in dim scale varieties. By permitting profound figuring out how to naturally find the picture portrayals from crude information in this way DeepFace is a scholarly element. Now and again, information might be not able to characterize explicit highlights particularly for face portrayal. DeepFace is an elective strategy where highlights are produced through preparing/learning measures without depending on explicit calculations. Learned highlights essentially beat the handcraft one where the test set is concealed. PCA, LBP and DeepFace will be thought about as far as exactness and computational time.

According to the findings, the researchers looked over the correlation between time scales from overall frequency waveband and waves of Delta-Mel spectrum. They also have experimented on nerves generating EEG signals affecting the vocal capabilities and facial expressions from which they derived a relation between shape and time function from their vocal data (Jan et al, 2017). Based on multivariate regression analysis they built a model on audio signal processing using Gaussian plane combination and staircase regression, which resulted in accurate prediction of depression scale from voice samples. Researchers have used motion history histogram for studying the visual and vocal features of a patient (Velvizhi et al, 2021; Yang et al, 2013). The motion history histogram is fused with processed audio signal features and is used to seize motion dynamics for studying ever-changing features. Linear and partial regression techniques are used for calculating the depression scale. The researchers have identified several features that have a great impact on calculating the depression scale prediction values. They fused collectively numerous capabilities that include head movement, finger movement, lip movement, eye-brow movement and Audio features producing lexicon from the linguistic modality (Srisuk, et al., 2018).

The researchers fed data from laboratory tests into the extreme gradient model to predict the outcomes of depression in healthy cases. After collecting tweets using a combination of social media analysis and a multi-modal model, we used the model to make predictions. They have used LGBP-TOP with Local Phase Quantization (LPQ) on inner facial regions like the forehead, lips, eyes and eyebrows. Regression techniques such as Correlated factors and Moore-Penrose are pseudo-inverse techniques in the multi-modal framework. The Moore–Penrose inverse of a matrix is a concept in use in mathematics, specifically linear algebra. The pseudoinverse is frequently used to find a "best fit" (least squares) solution to a system of linear equations which has no solution. They proposed a methodology to use fisher vectors to decode LGBP-TOP and dense and visible trajectories with low stage audio descriptor capabilities (Jan & Meng, 2015).

After carefully studying the video dataset, the researchers confirmed that the videos with slower movements have better scale prediction as compared to other videos. They used multiple techniques for searching of movement and speed facts that take place at the facial region (Jan et al, 2014). In addition to 12 attributes received from the audio information which includes a wide variety of silence period and overall smile period. However, using the visible function extraction strategy they included the texture, surface and facet facts. The deep learning methodologies have made an immense development on reputation, behavioural changes classification and the usage of neural networks to categorize human emotions and visions using processing techniques. These neural networks have a better application in describing the visible and facial features (Meng et al., 2014; Weber et al, 2016).

The researchers have proposed the usage of multitasking getting to know primarily based totally on audio and visible information. They used lengthy short-time period reminiscence models with pre-trained convolutional neural networks to train the architecture with the FER2013

dataset. This dataset has over 35,890 48 × 48 grayscale images. The overall performance of the dataset was better than AVEC 2013–14 dataset, however, it is nevertheless some distance far from neural models. Some drawbacks from the model are that the visual length they follow is small which reduces the dimensionality of the AVEC dataset images which results in the reduction of a huge amount of feature characteristics. This will have a very bad impact on the model as the features are slow, subtle and small. Hence, the model will not have expected accuracy and precision Ekman and Friesen (1978).

This paper elaborates the development of a novel approach for a neural network model that can examine the visual image from interview sessions to discover facial patterns that could indicate depression level. The user-generated data helps to distinguish between different depressive groups and depression symptoms can manifest in people with various mental illnesses in different ways. In particular, we want to automatically predict the depression scale and differentiate depression from other mental disorders using the patient's psychiatric illness history and dynamic descriptions extracted from the user inputs. The proposed framework is an extension to the pre-existing frameworks, replacing the handcrafted feature extraction with the deep feature extraction technique.

3 | DESIGN METHODOLOGY

3.1 | Depression dataset

For the development of the depression scale prediction model, classification of emotions and extracting facial features we have used the AVEC 2016–17 dataset to train the model. The AVEC 2016–17 dataset contains visual features computed using the scikit-learn toolbox4. In particular, we fit a linear support vector machine with stochastic gradient descent, that is, the loss is computed one sample at a time and the model is sequentially updated. We validated the model on the development set and conducted a grid search for optimal hyper-parameters on the development set of video data. Features of both modalities are taken from the provided challenge baseline features. Classification and training was performed on a frame-wise basis. We also performed temporal fusion through simple majority voting of all the frames within an entire screening interview. We have developed a dataset containing more than 300 video clips with each video clip containing a recorded interactive session by patient-computer interaction with low, mild and severe level mental disorders with a duration ranging from 1 to 18 min. The dataset consists of patients with an age group of 16–64 with a mean age calculated to be 34 years having an inflation of 15.7 years. The anxiety-based BDI-II scale ranges from 0 to 63 where each range has its significance. The range of 0–10 is considered to be normal with no anxiety, the range of 11–16 is considered to be mild mood disturbance or stress, the range of 17–20 is considered as borderline clinical anxiety, range of 21–30 is considered as moderate anxiety, range of 31–40 is considered for acute anxiety and >40 is serious anxiety. The BDI-II is a 21-item self-report instrument that is commonly used to assess the severity of depression in adolescents and adults. In 1996, the BDI-II was updated to align with the DSMIV depression criteria. The highest predicted scale from the dataset for anxiety is 47 which indicates that the dataset includes patients coming under each category. The dataset undergoes an audio signal processing technique using deep learning algorithms to calculate the anxiety BDI-II score of a patient. The dataset contains overall 300 video clips from which we have divided the data set into three equal parts that are for training, testing and validation in the ratio of 80:10:10. Overall, the dataset consists of 2,457,362 frames. The common period of these visuals is 2 min.

3.2 | Architectural overview

The automatic depression and visual assessment system take a visible input at the side of alternative sorts of visuals. As the dataset used was not as per the system requirement, the pre-processing of video data may be a required step. The video data was dampened into visual frames extracted from AVEC 2016–17 dataset. Deep Visual extraction and dimension reduction algorithms are eventually enforced to all visual images reducing their dimensionality. Dimensionality reduction helps to reduce the number of dimensions in numerical input data while preserving the data's key relationships. There really are numerous dimensionality reduction algorithms available, and there is no single ideal solution for all datasets. The model carries fusion of different features extracted from the dataset with completely different modality to ensure the choice. Dynamic variations in patterns are extracted throughout the model by reducing spatiality and victimization regression techniques for depression assessment (Figure 1).

At this stage where we carry out the feature extraction of the video clips. We break down the features into visual images from the discovered patterns using deep neural networks by mapping and reducing the dimensionality. Later, these frames are processed under convolutional neural networks like ResNet architecture for extracting excessive stage characteristics. As soon as the feature capabilities are taken out of visual frames from the video dataset, the miles are ranked and normalized among zero, and one earlier than the FDHH set of rules is implemented throughout every set of capabilities in keeping with video. The output is converted right into an unmatched row vector, on the way to constitute the temporal characteristic of a video. The uni-modal framework approach focuses on facial expressions associated mostly with the eye, lip, chin and eye-brow movement. A single distinct peak or most frequent value indicates a unimodal distribution. At first, the values rise until they achieve a single peak,

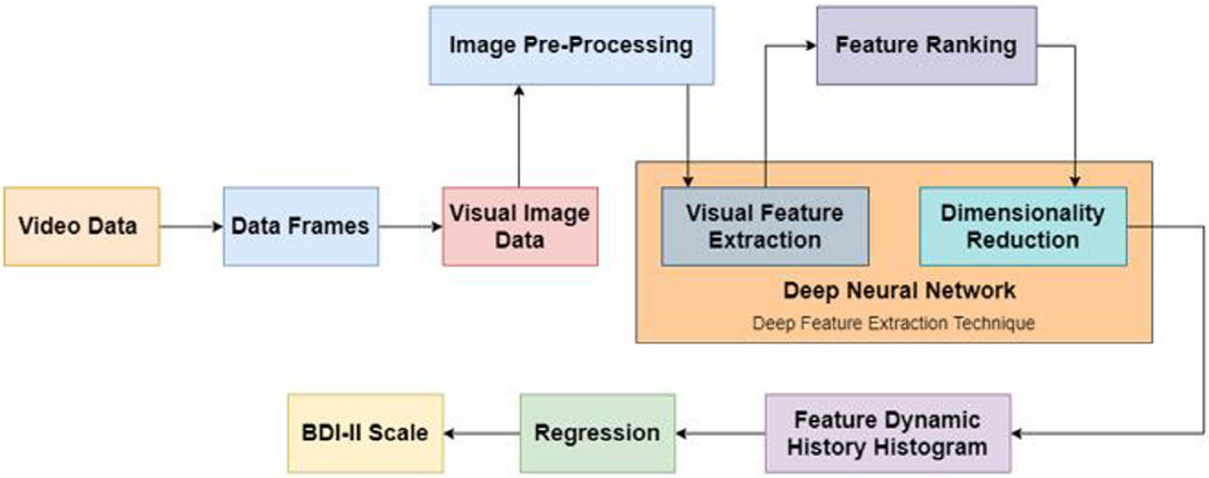

FIGURE 1 Block architectural diagram for visual processing and feature extraction

then after they fall. The normal curve has only one local maximum, that is an example of a unimodal distribution (peak). The model classifies the facial expressions into depressed and non-depressed classes according to the percentage scale focusing on each frame from the video. The efforts are blended at a characteristic stage with the aid of concatenating the capabilities produced with the aid of using every framework simply before principal component factor analysis regression (PCA) is enforced (Figure 2).

Our methodology takes a cycle model and a going with text based portrayal as input sources. It first subjects the two contributions to a linguistic analysis. PCA is a type of Singular Value Decomposition that is used to decompose data into a lower component (SVD). Factor Analysis is used to figure out what the underlying 'cause' is for these factors (latent or constituent) to absorb so much data from a number of variables in a dataset. The objective of this examination is to handle sentences and activities so that their similarities can be precisely evaluated. Second, we process precision scores that measure the semantic similarities between singular exercises and sentences. Third, we develop an action sentence arrangement. We do as such by supplementing the similarity scores with a thought of the requesting relations that exist between the different cycle steps caught in the model and text. In the fourth and last advance, we distinguish irregularities between the model-based and text based cycle portrayal. To accomplish this, we utilize the K-closest neighbour calculation that assesses the nature of the acquired action sentence arrangement. The end-product of the methodology is a bunch of anticipated irregularities, both at a cycle level, just as at an all the more fine granular action level (Figure 3).

3.3 | Dataset pre-processing

We have reduced the frame rate for a video clip from 35 frames per second to seven frames per second; consistent in comparison yields more robust and improved video data analytics as the system extracts more visual features from some frames. Short videos of duration 56 s to 2 min frames were extracted at a rate of six frames per second for collecting more sample frames. Consider a frame sample extracted from a video from a dataset of duration 1 min has extracted 16 frames. Every fourth frame of the sample starting at sample F1, are much different from pair frames in similar video samples (Figure 4).

We have split video within 16 frames keeping a four and eight frames lapse between two video clips from the respective dataset. Using this methodology, we can extract up to 80–30,000 frames from the dataset. The model extracts features from the front and side face landmarks using three-dimensional convolutional neural network-based architecture applied using ResNet architectures. A residual neural network (ResNet) is a type of neural network (ANN) that is based on pyramidal cell constructions in the frontal lobe. Skip connections, or shortcuts, are being used by residual neural nets to jump over some layers. After applying the neural network, the model crops and focuses on the facial regions such as lip, eye, eyebrow, forehead and other facial landmarks applied on each video from the training and testing dataset. In facial region, it includes the different techniques namely: Neural network, eigenfaces, automatic facing processing. Furthermore, facial recognition technology has an enhancement to prove accuracy and reliability.

On the other hand, we have trained the three-dimensional convolutional neural network (3D-CNN) based VGG face model using the video frames extracted from the dataset. We have used the pre-trained weights extracted from the UCF-101 dataset to train our model for facial and emotion recognition and initialized the weights on AVEC 2016–17 dataset. We have changed to a 40 convolutional neural network from 4086 layers to 512 layers to avoid over-fitting and loss of data. We have altered the loss operations into the Euclidean distance for performing

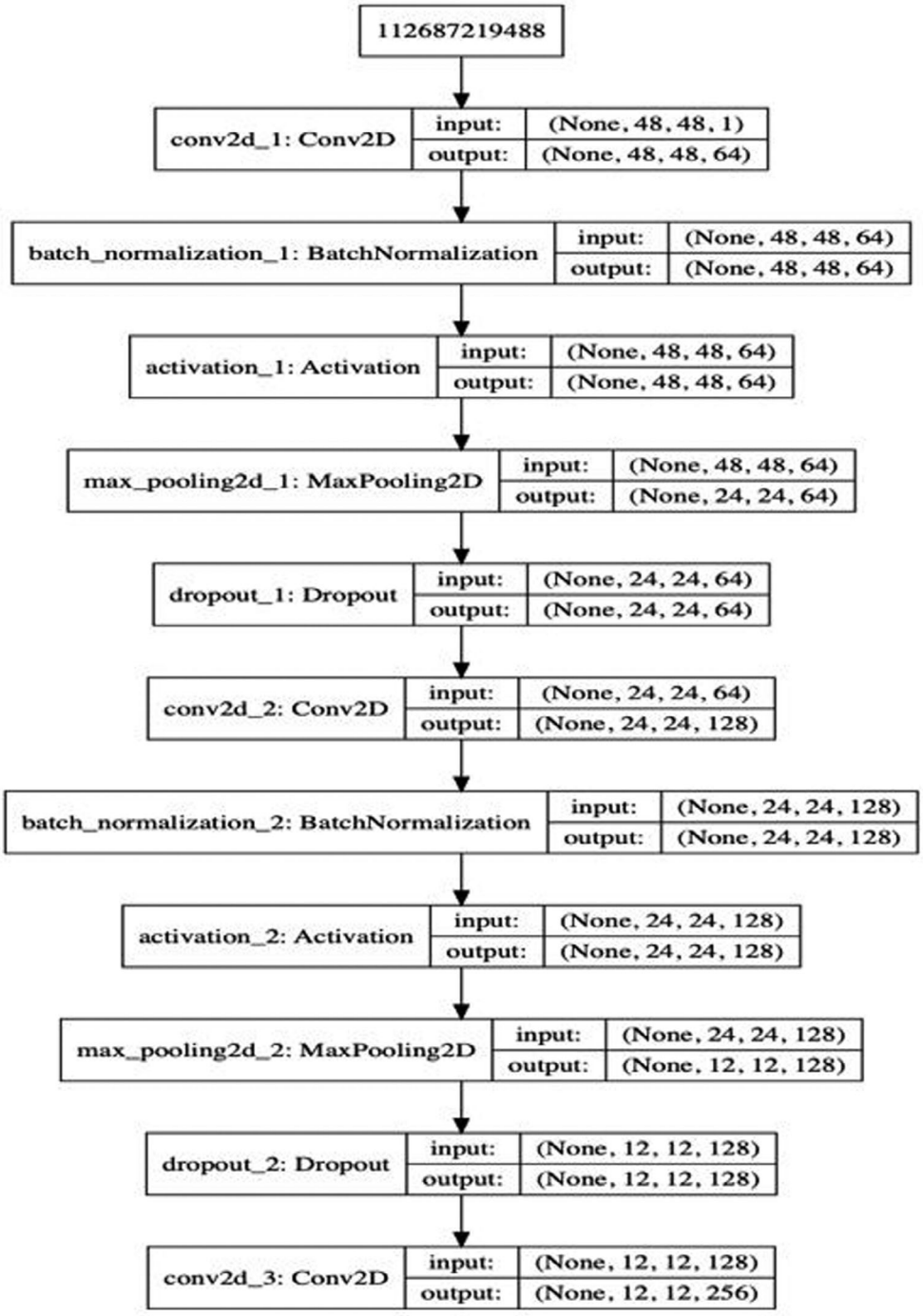

FIGURE 2 Convolutional layers applied for feature extraction

regression. We have implemented the changes using the Caffe tool test model as it supports three-dimensional neural networks conveniently. Convolutions are filters (matrix/vectors) with trainable parameters that are used to extract close to the bottom features from input data in neural networks and deep learning. The geographic or temporal links between input data points are maintained by them.

3.4 | Visual feature extraction

In this section, we have used different pre-trained convolutional neural network models particularly the architectures and their selected applications. We have used VGG Net and Alex Net for comparative analysis between them. Visual Geometry Group has a deep network of pre-trained models and neural networks such as Visual Geometry Group-Slow, Fast, Medium (VGG-S, F, M) networks. Also, we have VGG-D, E convolutional

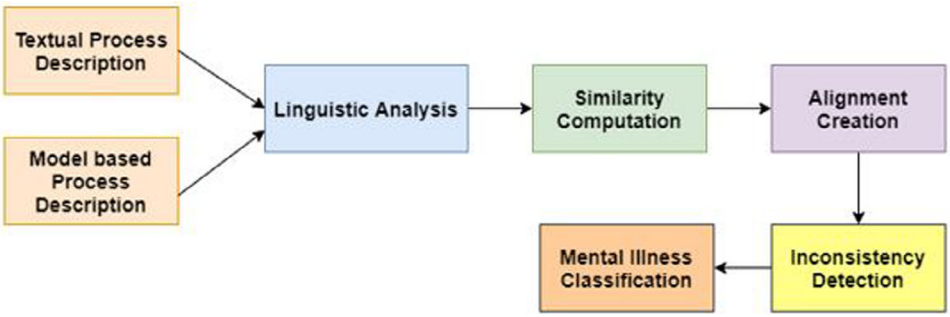

FIGURE 3 Block architectural diagram for extracting dynamic textual features from user inputs

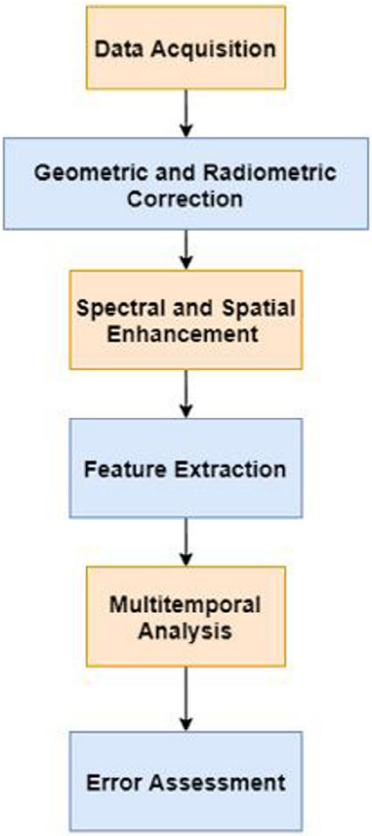

FIGURE 4 Flowchart for dataset pre-processing

neural networks. Visual Geometry Group is an abbreviation for Visual Geometry Group (a group of researchers at Oxford who developed this architecture). The VGG architecture can be divided into blocks, each of which is made up of 2D Convolution and Max Pooling layers. Using the ImageNet dataset, the VGG neural network architectures are tr8ained on 2.6 million facial images and tested for classification and face recognition modules as critical object detection modules that can best be used for detecting and analysing depression (Krizhevsky et al, 2012). The Visual Geometry Group-Face architecture contains a complete set of 36 convolutional neural network layers in which 16 are convolutional and 3 are connected having a variable filter depth from 64 to 512 layers with 3×3 kernel size.

The rest of the 20 convolution layers comprises of inter-linked activating maximum pooling layers. For predicting the depressive cases, the top-most layer is a softmax layer with a high and low level throughout the whole convolutional network. Before the output layer, Softmax is implemented using a neural network layer. The activation function and the Softmax layer must both have the same number of nodes. The edges and binary large objects are activated over the network to remake the visuals. The network tries to match the extracted features of the images with the features already present in the softmax layer with overall 2622 trained images containing special and highlighting features from the face from network-1 to network-5. The data for features extracted might be too huge to directly represent as a vector for representing faces. (Kumar et al., 2021).

In the development of this model, we have used 3D convolution neural networks with connected layers as FC1, FC2 and FC3 which are heritable (Shanmugam et al., 2020). In the model, we are also using a widespread network called AlexNet architecture consisting of 21 deep neural

network layers. The AlexNet architecture is used for the classification of features extracted from the dataset. The difference between VGG and AlexNet lies in the depth of their architectural design and filter sizes. AlexNet was the first to join the network to employ graphics processing (GPU) to improve performance. AlexNet has five convolutional layers, three max-pooling layers, two connected layer, two fully connected layers and the first SoftMax layer in its design. The main drawbacks of this AlexNet was the first convolutional network to utilize the graphics processing (GPU) to improve performance. 1. AlexNet has five convolutional layers, three max-pooling layers, two batch normalization, two fully connected layers and one SoftMax layer in its design.

3.5 | Feature history dynamic histogram

MHH are the descriptive temporary features of movement for recognizing visible movements. It was initially projected and implemented for recognizing facial movements. It statistics the grayscale cost modifications for every element inside the video. In contrast with different facial features which include recorded visual movement of lips, eyes, eyebrows, face, etc., Consisting of Greater dynamic records for pixels and calculates the overall performance in recognizing the human movements and facial expressions (Peng et al., 2014). Medizinische Hochschule Hannover now no longer most effectively affords wealthy movement statistics, however, additionally stays computationally inexpensive. MHH typically includes shooting movement facts of every pixel of 2-D images.

We have proposed ways for the statistic representations for a visible sequence to seize the dynamic versions. The prediction of values mostly depends on factors such as training and testing datasets, quality and amount of streams and features used and their relation, input data, and data pre-processing, number of epochs carried out for training and testing the model, selection for the design of architecture, the total number of neural networks, hidden layers and batch size. We have differentiated long-range features with single labels used as training samples (Figure 5).

The architectural neural network and Visual group geometry architectural model consist of 36 layers in which 16 layers are convolutional and 20 are geometrical. The length of the convolutional filter is set to 3×3 having a kernel length of 3×3. The intensity of the filter increases as the layers increase from 64 to 512. The layers connected to the model are of kernel size 1×1 and associated considering the dimensionality of 4096 dimensions connected to the closing layer with 2622 layers. The Last 20 layers consist of a mixture of co-linear activation layers, pooling layers and softmax layer for scale prediction. The entire structure verifies the excessive and occasional stage functions that seem to be throughout the community. We are resizing and remaking the images which may be visible in responses that are activated to produce edges and blobs.

4 | RESULT AND DISCUSSION

We have experimented according to the setup on the BDI-II scale we need to classify the facial features extracted from the dataset denoting negative and positive emotions in a particular video frame. The results for each video in the dataset are compared and analysed by calculating their

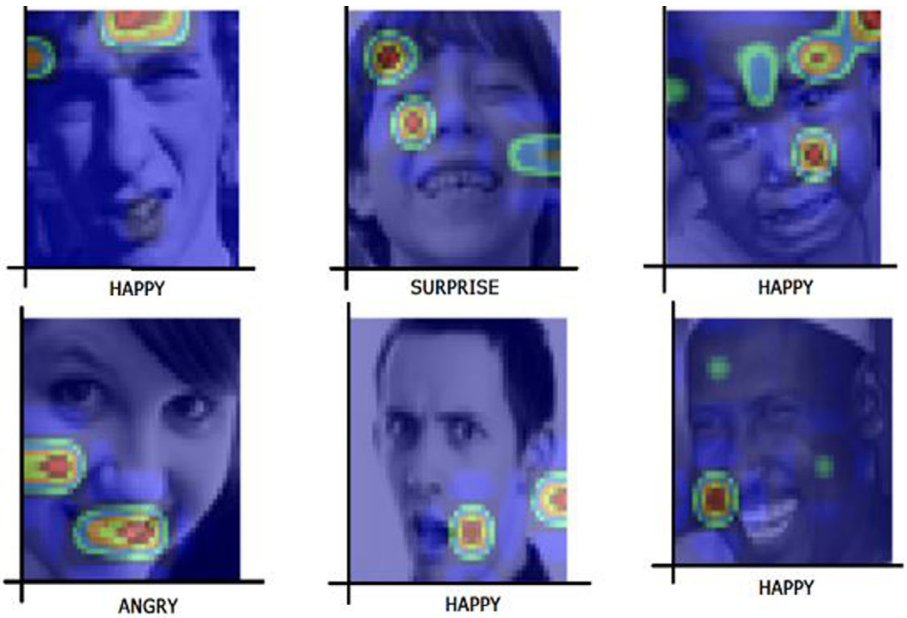

FIGURE 5 Feature extraction from face using chin, eye, eye-brow and lip moment

root square mean error and mean absolute error opposite for the ground truths. We have partitioned the dataset as train data, test data and validation data in the ratio of 80:10:10. To improve the efficiency of the model we have implemented several data pre-processing methodologies and split the large videos into small frames for improving the efficiency of the memory. We have implemented window normalization and minimum-maximum normalization techniques for evaluating dynamic changes in the features. Each partition in the dataset contains 100 video clips dividing the dataset into northwind and freeform databases respectively. We have used MatConvNet for managing the deep neural network to access the features over any convolutional neural network layer offering both 3D-CNN and VGG face.

Not every feature present in the dataset may have the same predictive power that dominating features have. Hence, using dimensionality reduction, we remove those features which do not affect the prediction of the depression scale and performance of our model from the dataset. We choose subset features that are correlated to each other and target variable but it has more probability of generating highly collinear feature space which affects the learning algorithm and reduces the accuracy of the model. Hence, we rank the features by using feature importance ranking methods according to their dominance and remove the irrelevant variables. We calculate the root mean square error at every node for the data once the features are rearranged. We averaged the difference between the mean errors at each node and normalized the differences using standard deviations. We have used five videos for testing our model architecture. We have extracted in total of 160 frames from the video dataset considered for testing. We have divided the video into three equal segments. The facial expressions we recognized for each part of the video and positive and negative parts were discovered from the video (Tables 1 and 2).

We have performed normalization to the dataset to bring all the features into a common range. We have trained the model using the input data with a size of 200 video frames for predicting depression scale. To obtain the real predicted values the model performs a denormalization process. The training is carried out up to 2000 epochs to improve model accuracy. The size of input data was fixed by calculating the mean absolute error percentage for each window size with a range of 50–250 having a window size from 10 to 40 with a minimum error value of 200 compared with other error values representing dynamic features. Initially we observed that the meta-data analysis performed during the experiment permitted a rough comparative analysis of methodologies through Cohen's chance robust metric approach. Several approaches tested on the dataset proved that the AVEC visual features achieved higher performance in comparison with AVEC 2014 dataset as shown in Table 3 (Table 4).

We have pre-processed the dataset to make it according to the system compatibility and extract minimum optimum features from the video dataset using pre-trained networks. Each video is divided into small data frames for extracting their deep features. We have resized and reshaped the video frames into $227 \times 227 \times 3$ represents height, width and colour channels. AlexNet architecture usually takes a frame input of $227 \times 227 \times 3$ into its network. Convolutional layers, pooling layers, fully connected layers and standardization layers are frequent convolution layer in CNNs. It simply means that convolution and pooling functions are implemented as artificial neurons instead of the regular activation functions mentioned above. We have resized and reshaped the photographs according to the input requirements of the MatConvNet toolbox and subtracted the mean image with each network. In the next steps, we are filtering each video frame through created network and procedure options.

TABLE 1 Use of different methodologies for calculating root mean square error and mean absolute error values for AVEC 2016 dataset

Methodology	Root mean square error (RMSE)	Mean absolute error (MAE)
3D convolutional tight-face	11.02	8.08
Recurrent neural network-3D convolutional tight-face	11.04	8.09
3D convolutional loose-face	10.04	9.15
Recurrent neural network-3D convolutional loose-face	11.09	8.75
Merge of 3-D convolutional tight-loose face weights	10.06	8.40
Recurrent neural network + 3D convolutional merge of tight-loose face weights	10.28	8.37

TABLE 2 Use of different methodologies for calculating root mean square error and mean absolute error values for AVEC 2017 dataset

Methodology	Root mean square error	Mean absolute error
3D convolutional tight-face	11.68	8.46
Recurrent neural network-3D convolutional tight-face	11.94	8.40
3D convolutional loose-face	10.86	8.76
Recurrent neural network-3D convolutional loose-face	11.67	8.69
Merge of 3-D convolutional tight-loose face weights	10.45	8.34
Recurrent neural network + 3D convolutional merge of tight-loose face weights	11.21	8.23

TABLE 3 Sensitivity, specificity, precision and accuracy values for depression scale prediction model

	Performance evaluation metrics			
Folds	**Specificity**	**Sensitivity**	**Precision**	**Accuracy**
Fold-I	96.74	97.23	97.89	96.47
Fold-II	95.87	94.83	92.04	94.23
Fold-III	92.70	93.75	91.17	91.54
Fold-IV	94.58	95.64	95.32	89.12
Fold-V	95.56	94.32	94.62	92.5
Overlapped	NIL	NIL	NIL	NIL
Average	93.93	94.38	93.35	92.56

TABLE 4 Metric evaluation performance of model using different methodologies

Partition	Methodology	Mean absolute error (MAE)	Root mean square error (RMSE)
Train	Local phase quantization	8.05	10.08
Train	Edge oriented histogram: Linear regression	10.01	13.06
Train	Edge oriented histogram: Partial least square	8.02	11.02
Train	Local binary pattern: Partial least square	10.09	12.05
Train	Edge oriented histogram: Partial least square + linear regression	10.05	11.09
Train	Local binary patterns partial least square + linear regression	10.02	12.14
Train	Local binary patterns: Linear regression	10.08	13.09

We have employed the spatial domain for each video clip from the space at every approach. In the case of AlexNet, we have preserved the connected layers from the sixteenth and eighteenth layer from 4096-dimensional connected layer feature vectors. We have selected the 16th convolutional layer from 4096 layer architecture to justify whether the layer gives higher features compared to another convolutional neural feature. According to our consideration the 35th, 34th and 32nd layer of the 4096 convolutional layers from the VGG face net are extracted. The 34th layer acts as an output for directly connected layer, the 35th layer acts as an output for correlated RELU activation platform and the 32nd layer is output for the primary fully connected layer. The VGG-Network as a convolutional neural network model in his study "Very Deep Convolutional Networks for Large-Scale Image Recognition" [1]. In ImageNet, which contains over 14 million images belonging to 1000 classes, this architecture achieved top-5 test accuracy of 92.7 percentile.

After carefully studying the video dataset, we confirmed that the videos with slower moments have better scale prediction as compared to other videos. They used multiple techniques for searching for movement and speed facts that take place at the facial region, in addition to 12 attributes received from the audio information which includes a wide variety of silence period and overall smile period. However, using the visible function extraction strategy they included the texture, surface and facet facts. The deep learning methodologies have made an immense development on reputation and behavioural changes classification and the usage of neural networks to categorize human emotions and visions using processing techniques. These neural networks have a better application in describing visible and facial features (Figure 6).

5 | CONCLUSION AND FUTURE WORK

Here, we have developed an AI-based automated system for predicting the depression scale. The model is purely based on recognizing facial expressions from recorded video sessions of depressed patients. We have used deep learning algorithms for extracting dynamic facial features and recognizing expressions producing the feature dynamic history histogram applied to feature dynamic vector sequences applied to the video sample. After calculating the feature dynamic history histogram, we apply regression techniques to correlate between feature and depression scales. The experimental results showed that the model had performed well on the test video dataset. The deep feature extraction technique significantly performs much better than the handcrafted feature extraction technique. The feature attributes are directly extracted from the responses provided by the convolutional layers perform best rather the performance of the neural networks are closely connected giving benchmarking results.

During the development of the model, we are highlighting three major contributions. The first contribution is we have created a model that is efficient in predicting the depression scale using facial expressions. The model performs 2.7% better than existing frameworks in facial detection

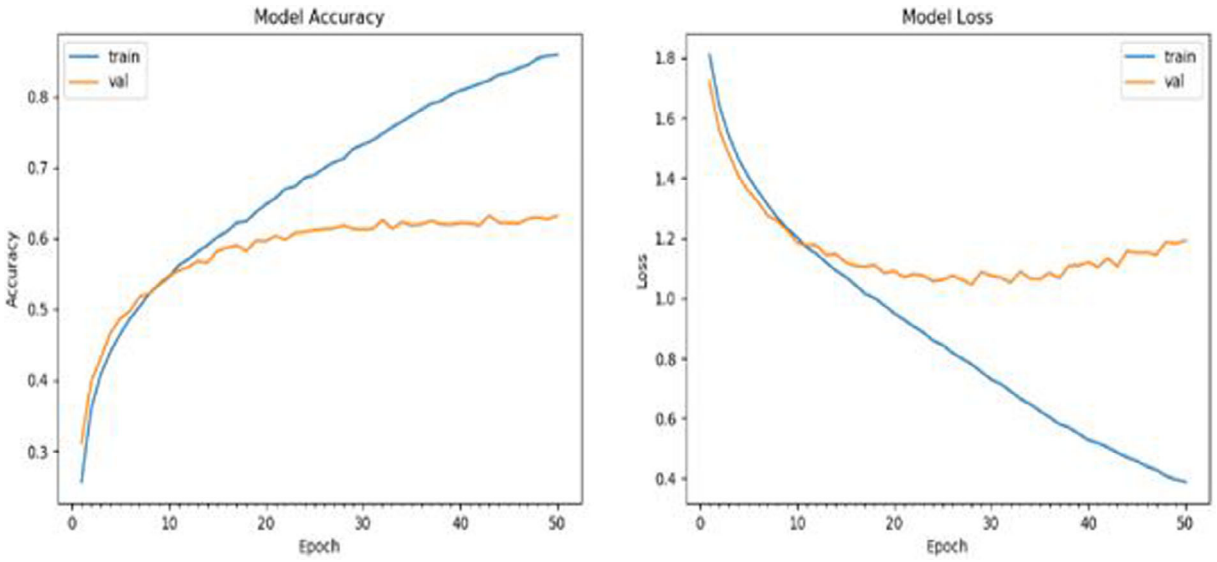

FIGURE 6 Model accuracy and model loss for VGG network

and feature extraction. The second contribution is that we are using the deep feature extraction method eliminating the previous traditional hand-crafted feature extraction method. We draw feature dynamic history using dynamic features using the main concept of MHH on videos. The third and major contribution is to differentiate depression from other mental disorders using the patient's psychiatric illness history and dynamic textual descriptions extracted from the user inputs. We apply the k-nearest neighbour algorithm on the dynamic textual descriptors to make a linguistic analysis for classifying mental illness into different classes. Our novel neural network model has shown remarkable performance by setting up important benchmarks focusing on slow-changing subtle facial expressions in small patterns. There are some limitations in the model that affect the model accuracy and system performance. The measurement of the BDI-II scale depends on the questions asked to the patients and how the patient responds to the questions. Sometimes the questions asked might not show the true depression level. The recorded sessions from the dataset have portrayed the depression level only from German ethnicity. The system might respond differently while testing on other ethnicities. The highest recorded depression scale in the dataset is 45 which is also a restriction on development and partition testing. All the things can be taken into consideration for further improvement in the system. We can improve the system performance by additionally training the VGG face neural network architecture with more videos. For a regression model, we can re-train the network with raw data. Adding more layers of convolutional neural networks may increase the extraction of facial features. Feature fusion techniques may also be added to the model for more accurate results.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Pratiksha Meshram https://orcid.org/0000-0002-3493-849X

REFERENCES

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F. & La Torre, F. D. (2009). Detecting depression from facial actions and vocal prosody. In Proc. 3rd int. conf. affective comput. intell. interact. workshops (ACII), Amsterdam, The Netherlands, pp. 1–7.

Cun, Y. L., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In Proc. Adv. Neural Inf. Process. Syst., Denver, CO, USA, 1990, pp. 396–404.

Davies, H., Wolz, I., Leppanen, J., Fernandez-Aranda, F., Schmidt, U., & Tchanturia, K. (2016). Facial expression to emotional stimuli in non-psychotic disorders: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *64*, 252–271.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system*. Consulting Psychol. Press.

Han, J., Li, K., Shao, L., Hu, X., He, S., Guo, L., Han, J., & Liu, T. (2014). Video abstraction based on FMRI-driven visual attention model. *Information Science*, *281*, 781–796.

Han, J., Chen, C, Shao, L, Hu, X, Han, J, & Liu, T. (2015). Learning computational models of video memorability from fMRI brain imaging. *IEEE Transactions on Cybernetics*, *45*(8), 1692–1703.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. In Proc. IEEE conf. comput. vis. pattern recognit. (CVPR), Las Vegas, NV, USA, pp. 770–778.

Jan, A. & Meng, H. (2015). Automatic 3D facial expression recognition using geometric and textured feature fusion. In Proc. 11th IEEE int. conf. workshops autom. face gesture recognit. (FG), vol. 5. May, pp. 1–6.

Jan, A., Meng, H., Gaus, Y. F. A., Zhang, F. & Turabzadeh, S. (2014). Automatic depression scale prediction using facial expression dynamics and regression. In Proc. 4th int. workshop audio/vis. emotion challenge (AVEC), Orlando, FL, USA, pp. 73–80.

Jan, A., Meng, H., Gaus, Y. F. B. A., & Zhang, F. (2017). Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, *10*, 679.

Jenkins, E., & Goldner, E. M. (2012). Approaches to understanding and addressing treatment-resistant depression: A scoping review. *Depression Research and Treatment*, *2012*, 1–7. https://doi.org/10.1155/2012/469680

Kaletsch, M., Pilgramm, S., Bischoff, M., Kindermann, S., Sauerbier, I., Stark, R., Lis, S., Gallhofer, B., Sammer, G., Zentgraf, K., Munzert, J., & Lorey, B. (2014). Major depressive disorder alters perception of emotional body movements. *Frontiers in Psychiatry*, *5*, 4.

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual Review of Public Health*, *34*, 119–138. https://doi.org/10.1146/annurev-publhealth-031912-114409

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Proc. adv. neural inf. process. syst., pp. 1097–1105. Audio/Vis. Emotion Challenge (AVEC), Amsterdam, The Netherlands.

Kumar, P., Chauhan, R., Stephan, T., Shankar, A. & Thakur, S. (2021). A Machine Learning Implementation for Mental Health Care. Application: Smart Watch for Depression Detection. In 2021 11th international conference on cloud computing, data science & engineering.

Long, J., Shelhamer, E. & Darrell, T. (2014). Fully convolutional networks for semantic segmentation. In Proc. IEEE conf. comput. vis. pattern recognit. (CVPR), Boston, MA, USA, 2014, pp. 3431–3440.

Luxton, D. D. (2015). *Artificial intelligence in behavioral and mental health care*. Academic Press.

Ma, X., Yang, H., Chen, Q, Huang, D., & Wang, Y. (2016). Depaudionet: An efficient deep model for audio based depression classification. In *Proc. 6th int. workshop audio/vis. emotion challenge (AVEC)*, Amsterdam, The Netherlands, pp. 35–42.

Marcus, M., Yasamy, M. T., Ommeren, M. V., Chisholm, D., & Saxena, S. (2012). *Depression: A global public health concern* (Vol. 1, pp. 6–8). WHO Dept. Mental Health Substance Abuse.

Meng, H., Huang, D., Wang, H., Yang, H., Al-Shuraifi, M. & Wang, Y. (2014). Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In Proc. 3rd ACM int. workshop audio/vis. emotion challenge (AVEC), Barcelona, Spain, vol. 2013, pp. 21–30.

Nasir, M., Jati, A., Shivakumar, P. G., Chakravarthula, S. N. & Georgiou, P. (2016). Multimodal and multiresolution depression detection from speech and facial landmark features. In Proc. 6th int. workshop.

Peng, J., El-Latif, A., Li, Q., Ahmed, A., & Niu, X. (2014). Multimodal biometric authentication based on score level fusion of finger biometrics. *Optik*, *125*(23), 6891–6897.

Sivakami, A., Balamurugan, K. S., Shanmugam, B., & Pitchaimuthu, S. (2020). Deep learning techniques for biomedical image analysis in healthcare. In *Deep neural networks for multimodal imaging and biomedical applications advances in bioinformatics and biomedical engineering* (pp. 31–46). IGI-Global.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Proceedings of ICLR*, 1–14.

Srisuk, S., Boonkong, A., Arunyagool, D. & Ongkittikul, S. (2018). Handcraft and learned feature extraction techniques for robust face recognition: A review. In 2018 international electrical engineering congress (iEECON), Krabi, Thailand, pp. 1–4. https://doi.org/10.1109/IEECON.2018.8712272.

Valstar, M. F., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R. & Pantic, M. (2013). AVEC (2013): The continuous audio/visual emotion and depression recognition challenge. In Proc. Int. Conf. ACM Multimedia Audio/Vis. Emotion Challenge Workshop, pp. 3–10.

Velvizhi, V., Billewar, S. R., Londhe, G., Kshirsagar, P., & Kumar, N. (2021). Big data for time series and trend analysis of poly waste management in India. *Materials Today: Proceedings*, *37*(2), 2607–2611. https://doi.org/10.1016/j.matpr.2020.08.507

Weber, R., Barrielle, V., Soladié, C., & Séguier, R. (2016). High-level geometry based features of video modality for emotion prediction. In *Proc. 6th int. workshop audio/vis. emotion challenge (AVEC)*, Amsterdam, The Netherlands, pp. 51–58. *Rev*., vol. 64, pp. 252–271, May.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G. & Mehta, D. D. (2013). Vocal and facial biomarkers of depression based on motor incoordination and timing. In Proc. 4th ACM int. workshop Audio/Vis. Emotion challenge (AVEC), Orlando, FL, USA, pp. 41–47.

Yang, Y., Fairbairn, C., & Cohn, J. F. (2013). Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, *4*(2), 142–150. https://doi.org/10.1109/T-AFFC.2012.38

AUTHOR BIOGRAPHIES

Pratiksha Meshram is working as Assistant professor in Department of IT at SVKM's NMIMS MPSTME Shirpur campus. She has passion towards development in the field of AI, Machine learning & deep learning for the applications in health care. She has organized & conducted several workshops & guest lectures for students & faculties. She has guided several projects of undergraduate students including inter disciplinary projects. She has filed one patent & published more than 12 research articles in emerging technologies.

Dr Radha Krishna Rambola, Associate Professor, CSE Dept, NMIMS University Mumbai, Shirpur. He had also worked as Associate Professor at Galgotias University, Noida (2014 – 2017) and Assistant Professor at Asia Pacific Institute of Information Technology, Panipat (2009 – 2014). Earned his B.E. (Computer Sci. and Engg) from University of Madras; M.Tech. (Computer Sci. and Engg) from Allahabad Agricultural Institute, Allahabad and PhD from T. M. B. University, Bhagalpur.He had worked as Software Engineer at Aspyre Systems, Chennai (1998 – 2000); Corporate Teaching Manager at Software Solution Integrated Ltd, Chennai, Hyderabad, Kanpur and New Delhi (2000 – 2003);

Sr. Lecturer, Fr. Agnel Institute of Tech, New Delhi (2003 – 2004 and 2006 – 2007); Project Manager at Eastern Software System Ltd, Delhi, Mumbai; also worked in Congo in Central Africa (2007 – 2009). Has over 21 years of experience in Teaching, Industry and Research with Engineering technology uses Languages, Software Engineering, System Analysis and Design, Software Development Project, Project Management, Database Management System, Software testing Methodology and Engineering related subjects with deep knowledge of ERP for realistic approach development in terms of Information Technology. Published more than 45 research papers in National and International journals & conferences. Delivered many Invited Talk viz. Invited as a Speaker in the Scientific Information Resource Division, BARC (Bhabha Atomic Research Centre) Seminar on 'Knowledge Networking for Excellence' on 10th August 2019 at BARC, Anushaktinagar, Mumbai; National Conference by University of C.V. Raman University, Bilaspur, Chattisgarh; IEEE Conference ICACCCN held at Galgotia College of Engineering & Technology, APJ Abdul Kalam Technical University, U.P. in 2018 and MTMI Conference organised by University of Maryland Eastern Shore, USA at Dubai in 2017.

How to cite this article: Meshram, P., & Rambola, R. K. (2022). Diagnosis of depression level using multimodal approaches using deep learning techniques with multiple selective features. *Expert Systems*, e12933. https://doi.org/10.1111/exsy.12933