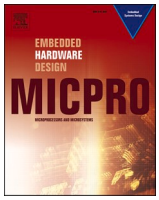
Microprocessors and Microsystems 95 (2022) 104681

Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro



Detection of stress, anxiety and depression (SAD) in video surveillance using ResNet-101

Astha Singh *, Divya Kumar

Department of Computer Science, Motilal Nehru National Institute of Technology Allahabad, India

ARTICLE INFO

Keywords:
Stress
Anxiety
Depression
Kanade-Lucas-Tomasi
ResNet 101
Facial feature extraction
Kalman filter
Video based analysis

ABSTRACT

Emotional disruptions are associated with the psychological state of a person that comes out in the form of non-verbal signals. The usage of medical resources for the identification of emotional activities is a complex and expensive task. Computer vision techniques equipped with artificial intelligence are capable of bringing automatic and fast identification of emotional variations of the human mind. Emotional variations may contain overlapping stages in which multiple non-separable emotional symptoms are more difficult to classify. The objective is to draw up an investigation of such a non-verbal body signal and correlate it with the psychological state of the person. Artificial intelligence techniques explore the identification of psychological states using pixel intensity information from datasets of facial expressions. The proposed study explores the classification of emotional symptoms into stress, anxiety and depression from facial expressions in a real-time video surveillance dataset. The second objective of the proposed study is to maintain classification accuracy for variation of real-time noise that may distort feature information. The study exhibits the use of the Kalman filter for the localization of intensity-based features and the use of the bilateral filter, contrast enhancement and adaptive filter algorithms for the removal of noise. Finally, ResNet 101 architecture has been used to classify symptoms of stress, anxiety and depression. The robustness of the proposed classification algorithm has been compared with other algorithms, such as PCA, Gradient boosting algorithm, KNN, Decision tree, Naïve Bayes, and SVM. It has been observed that ResNet 101 outperformed other models with a notable 98.4% accuracy.

1. Introduction

Recognition of human non-verbal behavior has been studied widely in various scenarios. The challenging study includes investigating negative non-verbal patterns and justifying their reasonable cause. Stress, anxiety, and depression are severe psychological issues obtained due to some event in a person's life. These mental issues are very complex to investigate and explore the treatment. Medical science has evaluated various psycho tests such as beck depression inventory [1], Hamilton rating scale [2], Raskin depression rating [3], Barthel index score [4], etc. All these tests require the patient's cooperation in conducting their non-verbal signals. In such a case, getting the patient's involvement is complex. Generally, stressed or depressed people show un-willingness to participate in any interrogatory psycho test. Therefore, the disease is mostly ignored and converted into high-order depression. The number of suicidal cases is increasing due to delays in recognizing psychological disorders. The psychological disorder is complex, and its investigation takes much time. In psychological disorders, anxiety is the early stage that is generated during day-to-day adverse events. This creates negative non-verbal signals [5] from a person's facial expression. These temporary signals cause a volatile psychological disorder from which a person can recover soon. The following psychological disorder stage is stress, which is long-term anxiety caused due to various negative thoughts. A series of anxiety over multiple days can convert into stress that stays long-term. The subsequent and most extreme psychological disorder is depression, which lasts long and is a non-volatile

Abbreviations: AI, artifical intelligence; KNN, K-nearest neighbors algorithm; PCA, pricipal component analysis; CNN, convolutional neural network; SVM, support vector machine; VGG, visual geometry group; SAD, stress anxiety depression; ResNet, residual neural network; LPQ-TOP, local phase quantization from three orthogonal planes; AVEC, audio/visual emotion challenge; RNN, recurrent neural network; HDRS, Hamilton depression rating scale; LR, linear regression; DASS-42, depression anxiety and stress scale -42; ROC, receiver operating characteristic; KLT, Kanade–Lucas–Tomasi; IEMOCAP, interactive emotional dyadic motion capture; FFT, fast Fourier transform.

* Corresponding author.
E-mail addresses: astha@mnnit.ac.in (A. Singh), divyak@mnnit.ac.in (D. Kumar).

https://doi.org/10.1016/j.micpro.2022.104681
Received 29 April 2022; Received in revised form 8 August 2022; Accepted 6 September 2022
Available online 15 September 2022

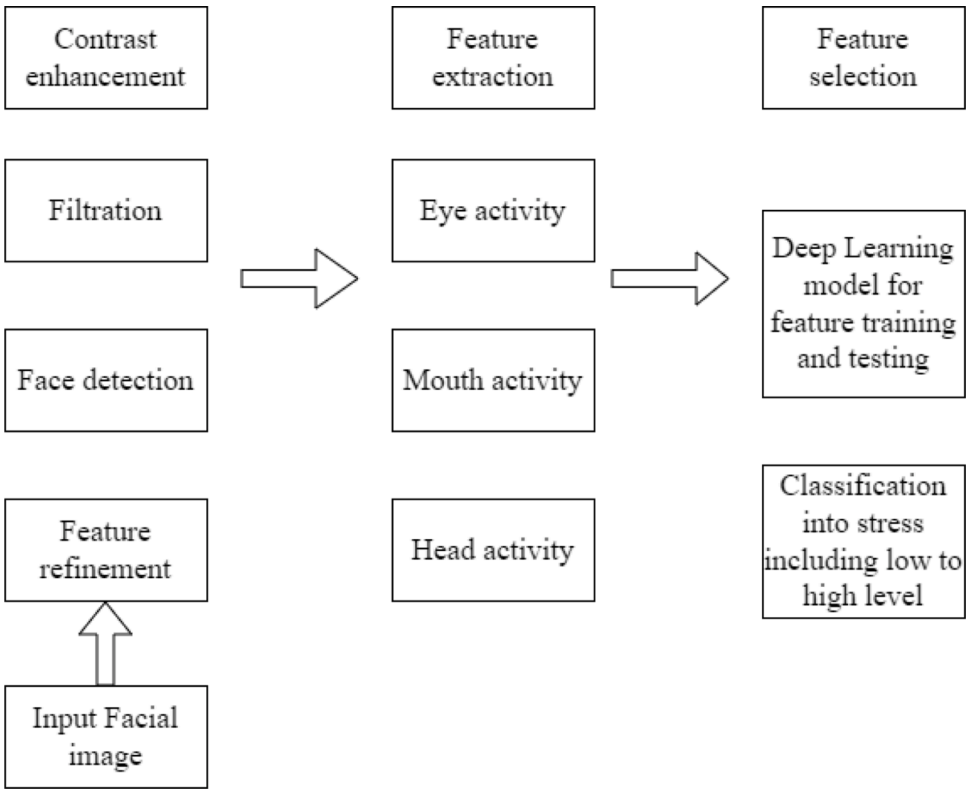


Fig.1. . General Architecture of depression detection scheme.

psychological disturbance. This stage can occur when multiple stress factors combine in which a person starts overthinking. These three stages of mental disorder are challenging to recognize accurately. These stages of cognitive disorder produce symptoms in the person in which a negative facial expression and body gesture/posture can easily be noticed. The earlier study investigated the facial expression and body movement using machine learning/deep learning to recognize mental disorders. Besides the earlier studies, it is still challenging to accurately identify and distinguish between stress, anxiety, and depression. Acute psychological disorder is a mental disorder that remains for a short period. Chronic psychological disorder [6] is the type in which stress remains for an extended period. It occurs due to a series of continuous minor stress and failure. It makes a person think more profound for an extended period.

The current study is based on features analysis from images and visuals in which non-verbal signs and expressions are classified using machine learning/deep learning techniques. The proposed method detects stress, anxiety, and depression from recorded video surveillance in which a person's activities are secretly recorded without acknowledging the patient. This system gives simpler user input and expressions. The proposed model performs feature extraction from the captured video clips and explores the non-verbal signs and movements. The proposed system uses a deep neural network model for feature training and testing.

This study aims to leverage contact-free video cameras for disorder detection. The study will find the facial signs and expressions that can provide insights into the identification of stress, anxiety, depression, and symptoms that are usually linked with fluctuations in physiological disorder and physical activities. The study will focus on extracting facial signs such as mouth activity, head motion, heart rate, blink rate, spatial gaze distribution, pupil dilation, and eye movements from different facial regions or using the Facial Action Coding System [7] and extracted Action Units from the face frames for stress detection. The proposed study leverage and integrates user action cues to enhance video-based stress detection.

The objective is to demodulate audio signals from a video clip if video data is inappropriate to distinguish between stress, depression, and anxiety. The demodulation audio will be separately analyzed into feature levels to determine if a person has stress, depression, or anxiety. In the earlier times, computer vision-based applied stress analysis had been widely adopted as it is fast and less expensive than any psychological test in medical terms. Depression detection started with emotion classification into anger, sadness, happiness, etc., from facial expressions. The features analysis was measured under various pre-processing and machine learning techniques to classify a person's emotional subject. Wen et al. [8] proposed depression detection using a sequence of facial images divided into 60 facial regions from which LPQ-TOP features are extracted. The components were trained using the support vector regression (SVR) model to obtain depression recognition accurately. The cognitive symptoms of an individual are those systems that can be recognized easily. It contains changing behavior, shouting, irregular body movement, facial expression, improper talking, etc. These symptoms are easily captured from the outside. The features of these symptoms may overlap as they are very closely related to psychological activities. Stress, depression, and anxiety features resemble each other as their psychological activities are correlated. Therefore, the severity level must be analyzed to distinguish between stress, depression, and anxiety. The level of severity can be measured through various tools.

The outline of the chapter scheme is as follows: Section 2 will discuss the related work occurred in this field. Section 3 contains methodology used in the proposed system. Section 4 shows the results achieved in the proposed system. Conclusion of the entire study section is presented by Section 5. The last section is reserve for the references taken from various studies.

Table 1
Distribution of categories of facial features types with stress or depression.

Head	Eyes	Mouth	Gaza	Pupil
head nod	eye blink	movement of mouth	direction of gaze	variation of the size of pupil
color of skin	movement of eyebrows	teeth depression	sharp gaze	pupil movement
Facial PPG	Eyelid variation	Swallow rate	Low gaze	Other variation

2. Related work

Cohn et. at. [1] Introduce an automatic assessment of depression using feature analysis from facial images. The study achieved 79% accuracy in distinguishing between depressed and non-depressed classes. The study was explored over the Pitts-burgh static image dataset in which an active appearance model and support vector machine was applied to classify depressed and non-depressed people. The experiment was not robust as it neither justifies the stress level nor can distinguish between short-term and long-term disorders. Alghowinem et al. [2] performed stress detection on the BlackDog dataset [6], in which a collection of facial images of various expressions are stored. This study applied over 128 images from which statistical features such as eye movement, eyelid changing, etc., had been studied under the SVM classifier. This study achieved about 88.3% in recognizing stress over the BlackDog dataset [6]. The model was further tested over the AVEC dataset, which analyzed images based on visual geometry group (VGG) features and obtained an 87.4% F1 score.

Wingenbach et al. [3] implemented three levels of emotional expression detection from a video-based dataset called ADFES-BIV (Amsterdam dynamic facial expression). The emotional expression was encoded from low to high intensity of stress level in which six numbers of emotions, such as anger, embarrassment, etc., are studied with three different intensity levels of stress. This study achieved about 69% accuracy in recognizing the intensity level of pressure on the ADFES-BIV dataset. These three intensity levels are stratified into low order stress, intermediate order stress, and high order stress based on emotional expression from facial images. Sonmez et al. [4] challenged proposing a classification model on the ADFES-BIV dataset in which the author performed the classification based on sparse representation of features by considering local information on facial data. The study was able to achieve 80% of accuracy for the recognition of those three intensity levels of stress.

Deep learning-based convolution neural networks (CNN) have been frequently applied to extract features from video-based facial datasets for depression analysis. Al jazaery et. at. [9] employed a recurrent neural network (RNN) to represent features of video-based input into deep learning neural network for better high-quality depression detection. The RNN is used to obtain temporal information encoded in the feature space sequence. Zhou et al. [10] introduced a neural network for a visual dataset to learn depression relation-rich features from facial expressions. The methodology identifies salient regions of the image and transforms them into histogram plots to study the variation of pixels and relate them to changing expressions. Jan et al. [11] proposed a histogram technique to represent feature space in a histogram to dawn an analysis of variation of features from input facial image. This technique used the partial least square (PLS) method and the linear regression (LR) model to conduct a depression detection experiment. The method was applied over AVEC 2014 [11] video-based dataset. Giannakakos et al. [12] introduced a methodology for facial cues to classify emotional stages by observing a person's eye, mouth, and head movement. This work implemented a model that distinguishes between short-term and

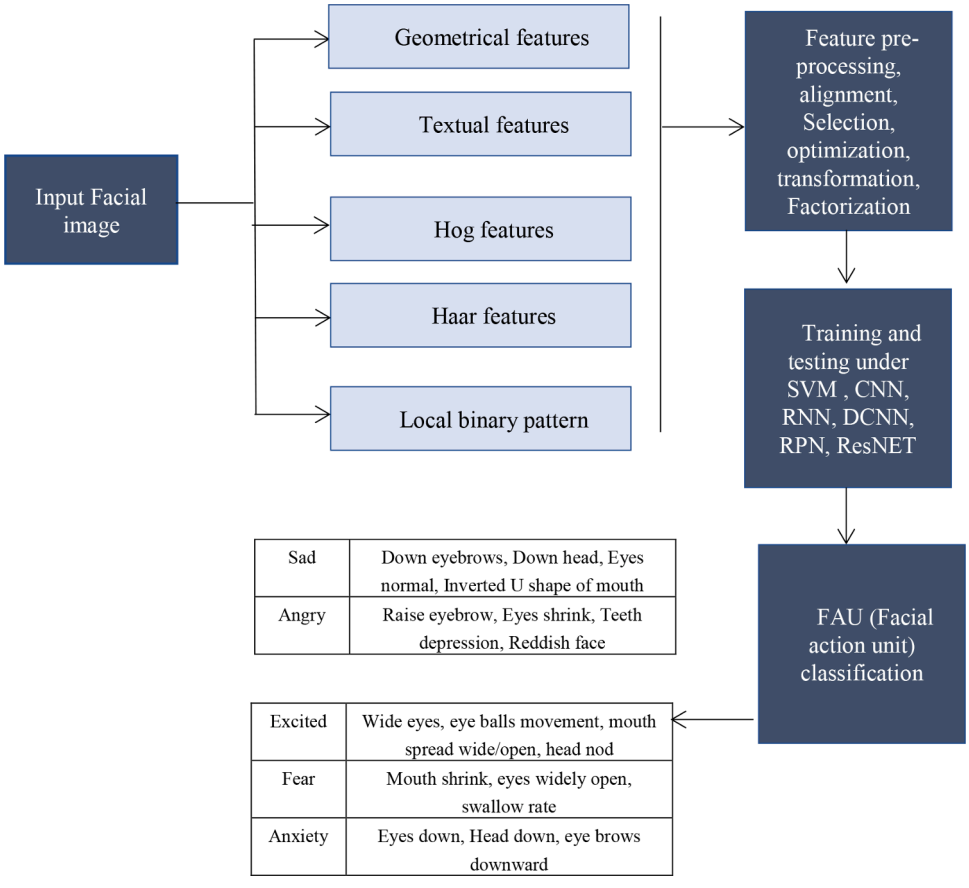


Fig. 2. Various features extraction on facial expression.

Table 2
Table representing stress/depression recognition using various scaling methodology and the related model.

Author	Description	Methodology
[15]	Support system in a student life under his stress	Psychological assessment
[16]	Examine exposure to stressors among student-employees	Stress level has been checked under various parameters like time pressure, social mistreatment, friendship problem, academics alienation etc.
[17]	Study of stress exposure in fresh student specially females	Sample based survey on Stress level in graduate female students enrolled in marriage under family pressure.
[18]	Explored the impact of financial, emotional, and academic on graduate student psychology that leads to generate stress problems.	Psychology student stress Questionnaire
[19]	Explored the organizational role in building stress in women	Organizational role stress scale
[20]	Facial expression tracked using features like smooth energy, MFCC, mean, standard deviation geometry of face	Classification algorithm like KNN, SVM and GMM has been applied to classify FAU. 92% of accuracy has been obtained.
[20]	GSR and ECG accelerometer data sources are used to extract spectral and time domain features.	Decision tree, naïve bayes and SVM algorithms has been applied and obtained 92.4% of recognition accuracy.
[21]	Questionnaire and skin conductance data sources are used to features like mean, standard deviation and mobility radius.	Applied SVM in which radial based function and linear kernel is used for the classification. The study also tested under PCA and KNN. The study obtained 75% of recognition accuracy.
[22]	The experiment performed stress detection using speech signal processing. BVP,ST and GSR data sources are used in which features like mean IBI, BV amplitude, GSR mean value etc. are extracted.	The model used naïve bayes, decision tree and SVM for the experiment and obtained 90.10% of recognition rate for stress. The model is totally based on speech recognition not the facial expression.
[23]	The objection of the study is to identify stress/depression in student life and hence the study utilized 322 students of different institutes.	Perceived stress scale, modified ways of coping scale and COPE scale has been applied to find level of stress in students.
[24]	The author defines seven types of emotional classes such as natural, surprise, joy, sad, disgust, anger and fear. 121 different feature points in the face are generalized using the modeling tool	The study used Microsoft Kinect for 3-dimensional face modeling.
[25]	The study performs RCNN based classification to avoid poor graded features from facial extract. The system proposes real time recognition.	The facial portion has been detected by using region proposal networks to obtain high quality features. The study achieved 94% of accuracy using active shape model ad boost classifier.

long-term anxiety disorders. Also, it performs the detection of depression separately. The methodology was robust but failed to maintain a better accuracy rate on a different dataset. A study conducted by Cohn et al. [13] shows the extraction of mid-level facial features, i.e., facial action units (FAUs) [14], for depression analysis. The method shows excellent recognition over AVEC 2016 video-based dataset. Fig. 1 shows the general architecture of the depression detection scheme.

The general architecture shown in Fig. 1 is prevalent in the earlier studies where mostly machine learning and deep learning algorithms are applied to obtain classification from input facial images. Fig. 1 shows the basic flow of methodology to receive recognition of emotions from facial activity containing swallow rate, eye movement, etc. Table 1 offer various facial features based on which emotional states can be classified.

The action units are stored in terms of feature information. Each action unit contains patient actions such as expression, eye movement, body gesture, etc. These action units are analyzed and mapped with psychological events to conduct a classification of emotion categories. Emotion processing does involve various facial features and the noted expression that connects with stress/depression. The different variation of facial parts has been described in Table 1. The movement of the head, eyes, mouth, gaze, and pupil may reveal various measurements of thoughts and is connected to emotional variations.

Fig. 2 explains the extraction and correlation of emotional features with the psychological states of a person. The features contain facial geometry, texture, local binary information, haar points, etc. These features are encoded into various facial action units, and the correlation will be established with emotional states. Table 2 contains the description and methodology used in multiple recent research works.

Stress levels in patients can be measured using various sets of standard scales proposed and tested through multiple studies. The Holmes and Rahe Stress Scale [12] bases the sources of our mental stress on our life events and contains a list of forty-three life events based on which a relative score is calculated. This scale has a low accuracy leading to poor overall performance. The Depression and Anxiety Scale (DASS-42) has 42 questions to calculate individual stress, anxiety, and depression scores. A shorter version of this scale with 21 questions has also been constructed and verified [11]. The Hamilton Anxiety Scale [11] consists of 14 items and can measure psychic and somatic anxiety. There are various scales for measuring depression levels in patients. The Hamilton Depression Rating Scale [8], Montgomery-Aberg Depression Rating Scale (Hamilton, Schutte & Malouff, 2001), Raskin Depression Rating Scale [21] Beck Depression Inventor [3], Geriatric Depression Scale (GDS) [14], Zung Self-Rating Depression Scale [25] and the Patient Health Questionnaire (PHQ) [20] are all different scales and questionnaires that provide a score rating that gives a relative measure of depression level among patients taking the test. Gabriel Tsechpenakis et al. in 2005 [21] researched deception detection, where it is required to detect and track the regions of interest in the examined video, i.e., the head and hands, using the skin color-based method. Then, remove the movement descriptors used in the recognition from the extracted blob features, i.e., positions and orientations. Finally, the HMM-based (Hidden Markov Model) approach is used to detect and recognize two possible behavioral states, agitation, and over-control, which indicate possible deception. HMM, a method is helpful in gesture, gait, and sign language recognition.

Don Ardell stress test [15] is a separate robust stress test to find specific stress levels in a person's life. It offers a balanced assessment of varied stress sources. It finds the importance of including all aspects of life in understanding stress. The Ardell Wellness Stress tests analyze a person's physical, mental, emotional, spiritual, and social aspects to outline a balanced assessment. The test contains a six-point scale, including a neutral point in which no negative and no positive emotions reside in life. Ardell wellness stress test is also an effective way to analyze the stress in people. It has seven basic score points, which reflect the person's mood. To implement this test in a population, a set of questionnaires is prepared and asked a student that acts as a stress source. The responses of the student are recorded in terms of 7 fundamental scores i.e. +3, +2, +1, 0, 1, 2, 3. After the sum of all recorded scores from each questionnaire is calculated and based on the total score, the final stress level of a person has been decided. These techniques are based on a questionnaire approach in which the patient needs to answer an interrogator's questions. The decision on stress level in the patient is taken based on their answers. This technique is obsolete and quite fragile as it requires the physical intervention of the patient. The patient may not co-operate or provide false information, which leads to decreased stress recognition accuracy. Earlier techniques were applied mostly on static images taken from standard datasets such as ADFES-BIV [16], AVEC [17], DemntiaBank [18], Reach Out Triage shared task [19], etc. The dataset is purely symmetric and does not contain any dynamic variations. Some limitations are also in the existing emotional

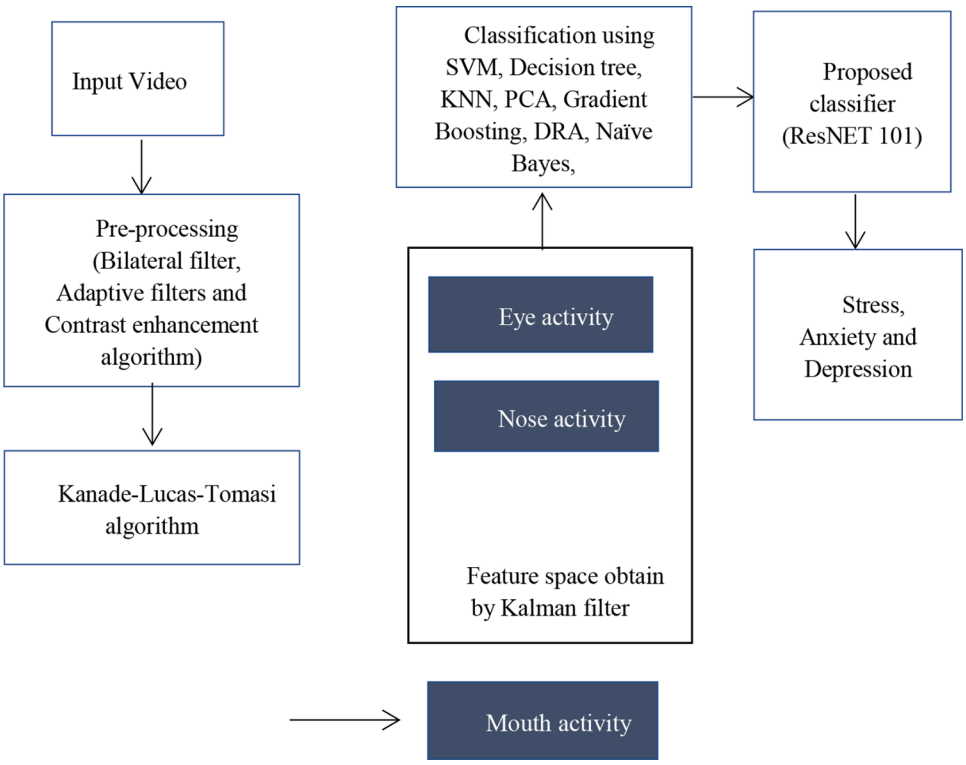


Fig.3. . Flow chart of the proposed scheme.

recognition models. Hardware deployment for the identification of emotional activities is one of the unavoidable challenges found in various studies. It contains many variations and complexes depending on the type of body signal. Another challenge in the current work is mapping the correlation between body signals and psychological events. In emotion identification, the challenging task is to design a model to measure the severity level of emotional symptoms. The symptom of one class of emotion resembles each other, and so it becomes challenging for a model to distinguish among the emotion categories. The earlier study found that the AI model remains convenient compared to medical investigation for detecting emotional activities. But the challenge is that the model must be able to process all body signals. The AI model is found to be effective for specific body signals in various existing works. Another challenge is to design a cost and time-efficient model for emotion category identification. The model must be robust so that it can handle any error. Some challenges exist in artificial intelligence techniques as this technique may fail to configure the dataset if it contains impurity. Model overfitting and underfitting are some issues observed in AI-based practices. Working with low-frequency user input data may be insufficient for emotion recognition analysis. Low-frequency data may contain insufficient features for making a correlation with emotional features.

The proposed technique has taken a standard surveillance video dataset [26] which contains about 300 people's surveillance streaming video data. The objective of the proposed model is to ensure the collection of natural facial expressions and features without using a person's interaction in any special interrogation. Unlike previous questionnaire-based techniques in which a person needs to answer or respond to a set of particular questions in front of a video recorder, the proposed method of video recording from a surveillance system captures a person's natural experiences that they may feel during different levels of stress. The model requires the person's interaction in real time to avoid capturing non-manipulated facial expressions. Fig. 3 shows the flow diagram of the proposed system. The model extracts real-time surveillance video input, then performs pre-processing of the frames to improve feature quality. The proposed method applies Kanade-Lucas-Tomasi (KLT) algorithm [27] for facial portion extraction, in which the algorithm plots a rectangular box around the face based on the local features. The KLT algorithm identifies the regional binary pattern on the facial way to avoid the detection of any noise or unwanted details. The algorithm performs local optimization on video frames to extract facial portions in a rectangular block. Here, the pre-processing task improves the visual quality of the video frames so that it becomes easier for the KLT algorithm to plot the track-able feature on the first frame and trace each feature in the subsequent frames with the help of displacement. The displacement of a particular feature is defined as the displacement that minimizes the sum of differences. The KLT algorithm produces output as extracted facial portion from video frames. The selected facial images are then undergone robust feature extraction that is performed by the Kalman filter algorithm. This algorithm is used to track the movements and variations in facial expression. The prediction of stress levels primarily depends on the interpretation of expressions based on which depression, anxiety, and stress will be classified. The Kalman filter [28] performs tracking of the geometry of eye movement, nose activity, head pose, and the movement of the mouth. All the movements are relatable to various levels of stress. Facial movement and expression are translated into local binary features, which are fed into ResNet 101 model [11]. The ResNet model performs the classification of features into stress, anxiety, and depression based on feature range, type, value, geometry, and orientation. Other classification algorithms include PCA, Gradient boosting algorithm, KNN, Decision tree, Naïve Bayes and SVM. These algorithms are also tested on the common feature space to perform a comparative study with the proposed ResNet 101 model.

3. Methodology

As discussed earlier, the proposed model uses a surveillance video dataset in which several video frames of 300 numbers of people are recorded and analyzed individually. The proposed model aims to distinguish between stress, anxiety, and depression correctly. Fig. 3 below shows a flow chart of the proposed scheme.

Fig.4. . Sample of video frames of IEMOCAP dataset.

Fig. 3 shows the overall flow diagram of the proposed system. The model first takes the input surveillance video and performs pre-processing. After, the model extracts the facial portion from the pre-processed video frame. The Kalman filter performs the feature extraction. Finally, the classification of features takes place using ResNet 101 and machine learning algorithms.

3.1. Dataset generation

IEMOCAP [26] (Interactive emotional dyadic motion capture) is a standard video dataset taken at the SAIL lab. It contains 12 h of surveillance video recordings. It captures various expressions, visuals, speech, movements, etc. The dataset also contains multiple dyadic sessions, which are based on detailed interrogation. The frames of the video dataset are labeled with various emotions such as sadness, anger, anxiety, stress, neutral, etc. The proposed methodology uses surveillance videos of about 300 people containing different emotions from the IEMOCAP dataset [26]. The sample of video dataset is shown below in Fig. 4.

The dataset is divided into training and testing sub-parts in 70 to 30 ratios. The 70% of data has been verified for the training in which each video is resized into equal dimensions first. Each video frame contains a size of 338 × 320 pixels. These frames are loaded, and the subject's facial portion is extracted with a dimension of 144 × 138 pixels. Few samples of extracted facial images of 5 subjects have been shown in Fig. 4. Each image is extracted from individual video frames.

There is a significant relevance of the study conducted for emotional recognition through video surveillance. Video dataset may increase complexity in feature analysis, but it is found to be very effective for recognizing patients' emotional activities. The video dataset contains the patient's body movement and other activities that record the person's behavior and body movement in various situations. It is very easy to find emotional categories through a video dataset as all the relevant features are available. This may not be effectively possible using other datasets such as audio and image. The audio dataset contains the patient's audio which the patient can fabricate. The audio dataset includes features that may overlap with other emotional activities. The accuracy of emotion recognition based on the audio dataset is low, and there may be higher false positive and false negative rates due to feature overlap cases. The image dataset is also not effective as compared to the video dataset. The image dataset contains static images, and its analysis is limited. The image of patients may not give rich information about patients' behavior as the video dataset does. In a video dataset, video input from the uses can be analyzed frame to frame, which contains various patient information such as expression, body movement, etc. Each frame can be analyzed separately with the correlation to the adjacent frames. The best algorithm, on average, that researchers recommend is the Kalman filter, which can extract features from video frames. Another best algorithm is based on a neural network with the highest capability to analyze video features. Kalman filter is a highly trained tool that enables the capture of feature information from a dynamic input dataset. In the case of video datasets, the tool can analyze detailed information frame-by-frame.

3.2. Pre-processing

The input video frames may contain various noises such as low contrast, asymmetric color variation, blurring effect, etc. These noises must be removed before the feature extraction to enhance the feature space that helps in classification. The model may fail to read features effectively, so the correlation of features with anxiety, stress, and depression also degrades. Data pre-processing, in which data interpretation and filtering has been made for the removal of any noise from the data. The data signals may carry unwanted signals and noise. And so, the pre-processing step is applied to extract the region of interest from the data. Data pre-processing is used to refine the feature vectors of the body signals. The pre-processing task must recognize stress, depression, and anxiety symptoms accurately. The pre-processing task makes the feature more visible to the model. It enhances the feature vector by applying various filters to the input dataset. After a certain threshold of the pre-processing task, information loss can occur. To avoid any information loss, the filtration process must be kept under the specific unit of a threshold value that gives the sufficient intensity of the filtration process. Hence, the proposed model performs pre-processing of the video frames first to enhance the overall feature quality. The proposed model applied the following filter to enhance feature space.

a) *Bilateral filter*

This filter is used to obtain non-linear combination of nearby pixels

Fig. 5. Sample of pre-processing task.

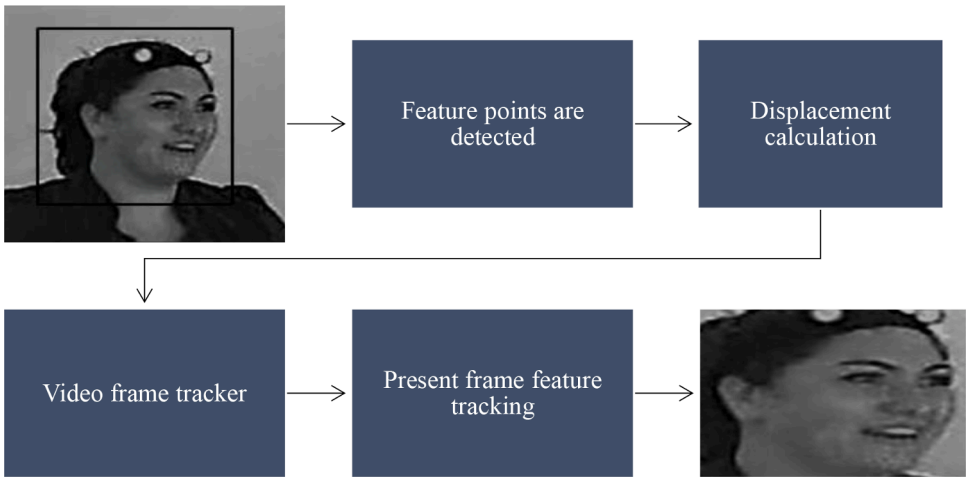


Fig. 6. Flow chart of Object Detection from Video Input.

in order to smooth the images in video frames while keeping the edges preserved. It collects the weighted average of local information that contains intensity information. The weighted average of a sample pixel (W(S)) is calculated over the intensity (I) is given as:-

$$W(S) = \frac{\sum_{s' \in S} I(s') S(S - s') T(I(S) - I(s'))}{\sum_{s' \in S} S(s' - S) I(I(S) - I(s'))}$$

Here $S(S - s')$ *and* $T(I(S) - I(s'))$ are the spatial and tonal weights of s' pixel information. Then Gaussian function is applied on pixel and its intensity to smooth its properties.

$$s'(S) = \frac{1}{\sigma_S \sqrt{2\pi}} e^{-S^2/2\sigma^2}$$

$$I(S) = \frac{1}{\sigma_I \sqrt{2\pi}} e^{-S^2/2\sigma^2}$$

a) *Adaptive filter*

In adaptive filter, the noise removal process depends on number of pixels associated. It works both as high pass filer and low pass filter based on number of associate pixel. It applies adaptation based on filter level automatically that ranges from 0 to max. The filter level is achieved as follows:-

$$f(I) \begin{cases} \max(0, f(I-1) - f_{dec}), & q(I-1) < q_f; \\ \min(0, f(I-1) + f_{inc}), & q(I-1) > q_f \end{cases}$$

f_{dec} and f_{inc} are the decrement and increment units based on level of filtration process changes by adaptive filter. $q(I-1)$ is the quantization parameter that is set to surpass the error estimation up-to few threshold. This makes the model flexible and adaptive to carry out pixel enhancement.

a) *Contrast enhancement algorithm*

Convert the frame into fuzzy domain:-

$$F(x_{ab}) = (1 - \cos(\pi x_{ab} / 255) / 2)$$

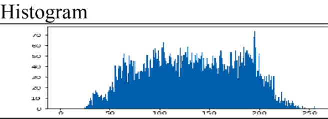
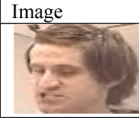
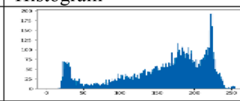
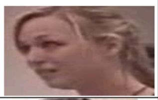
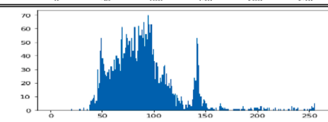
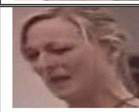
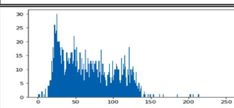
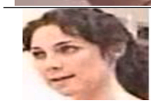
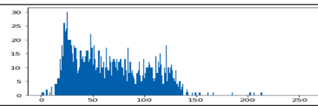
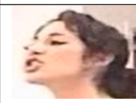
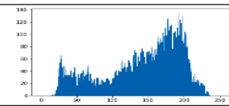
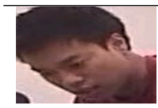
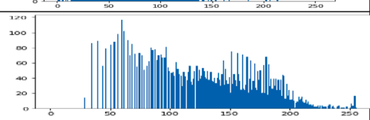
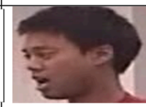
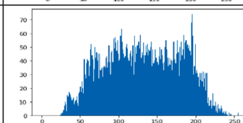
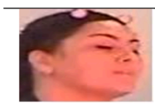
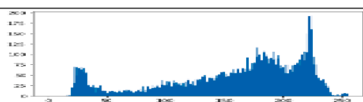
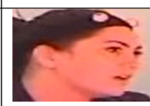
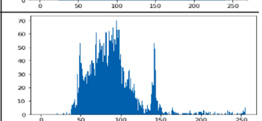
(x_{ab}) is the member function of the grey scale frames. The apply contrast enhancement equation for pixel value x_{ab} from a frame:-

$$F(x_{ab}) \begin{cases} \dfrac{x_{ab}^2}{x_{threshold}} & 0 \le x_{ab} \le x_{threshold} \\ 1 - \dfrac{(1 - x_{ab})^2}{1 - x_{threshold}} & x_{threshold} \le x_{ab} \le 1 \end{cases}$$

Fig. 5 shows the sample of input image from video frame, pre-processed image (after applying all the proposed pre-processing algorithms) and the histogram representation of pre-processed image.

The proposed model applies preprocessing techniques, including bilateral filter, adaptive filter, and contrast enhancement algorithm. The proposed preprocessing methods are working best to enhance feature vectors of action units of the input dataset. The proposed preprocessing algorithms are robust and consume less time to improve the feature vector. Other preprocessing algorithms are also existing Pixel brightness transformations, Brightness corrections, Geometric Transformations, Image Filtering and Segmentation, Fourier transformation, Image restauration, Laplacian Filtering, and Directional Filtering, etc. These preprocessing algorithms are complex and may trigger the loss of information from the image that may adversely affect the recognition rate. These preprocessing algorithms are time-consuming and show the best performance on specific datasets. Therefore, the proposed model utilizes bilateral, adaptive, and contrast enhancement algorithms as pre-processing algorithms to enhance feature quality in the video dataset. The proposed preprocessing technique does not cause any information loss.

Table 3
Sample of histogram obtained from video frames.

Image	Histogram	Image	Histogram
			
			
			
			
			

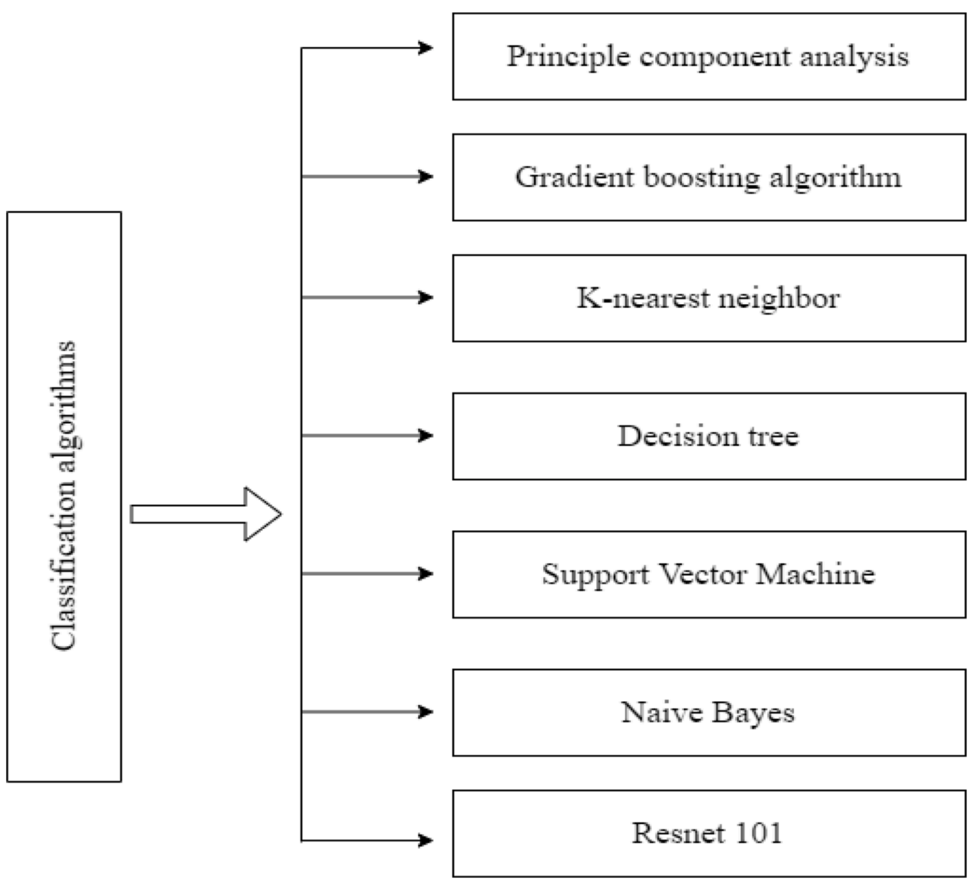


Fig. 7. Types of classification Algorithm.

3.3. Object detection from video input using Kanade-Lucas-Tomasi (KLT) algorithm

The KLT algorithm is used to extract the facial portion from a pre-processed video input frame by making a plot of features on the facial portion and transforming the extraction into a rectangular block. The steps of the KLT algorithm have been shown below :-

Step1: Take the input video containing sequence of captured frames.

Step2: Plot the region of interest on the facial portion to detect face components.

Step3: The localization is made in a rectangular block for one frame image. The same localization is made in other frames based on plotting

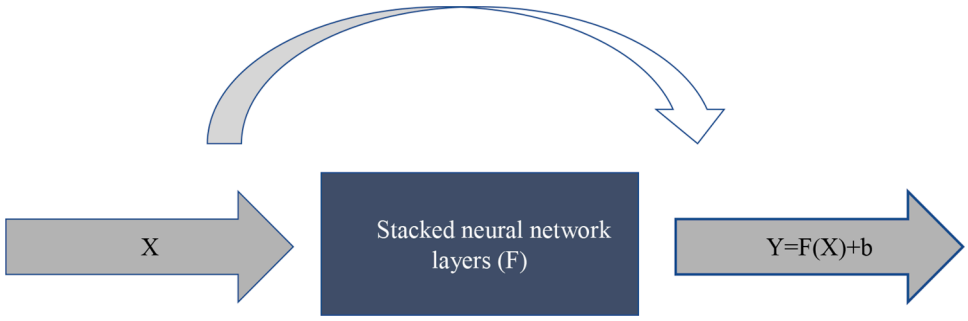


Fig.8. . Basic flow diagram of ResNet101 model.

Fig 9. The image showing sample of people in stress, depression and anxiety.

Table 4
Score calculation of Anxiety.

Image 1	Image 5	Image 8	Image 16	Image 20	Image 32	Image 36	Score	Anxiety
3	1	6	2	0	0	0	4	Extremely severe
4	2	4	4	6	5	8	6	Moderate severe
1	0	6	5	0	7	9	3	Normal
3	0	0	4	3	6	0	0	Mild

Table 5
Score calculation of Depression.

Image 3	Image 7	Image 18	Image 29	Image 29	Image 34	Image 42	Score	Depression
3	1	6	2	0	0	4	4	Extremely severe
4	2	3	4	6	3	4	6	Moderate severe
1	0	6	5	3	7	9	3	Normal
3	0	0	4	3	6	0	0	Mild

Table 6
Score calculation of Stress.

Image 3	Image 7	Image 18	Image 29	Image 29	Image 34	Image 42	Score	Stress
3	1	6	2	0	0	4	4	Extremely severe
4	2	3	4	6	3	4	6	Moderate severe
1	0	6	5	3	7	9	3	Normal
3	0	0	4	3	6	0	0	Mild

the tracking points.

Step4: The tracker is used to estimate the scale, rotation and translation between previous and new-points.

Feature analysis from each video frame is based on the study of the frequency of every pixel information. The frequency analysis can be seen by histogram representation which contains pixel information. The fast Fourier transform (FFT) algorithm has been applied in various existing models to fetch frequency components. Each such component contains pixel information of an image. The proposed model also analyzes the frequency variation of region of interest of the input video dataset. This analysis helps to map or correlate the features with the emotion category. Recognition of stress, anxiety, and depression has been performed based on features containing pixel information in frequency components.

Fig. 6 shows the flow diagram of the functioning of the KLT algorithm in which feature points are firstly located in each video frame. The feature displacement has been calculated to track the facial portion of the subject. Fig. 4 shows one of the video frames of a subject from which

Table 7
Confusion Matrix obtained by different ML methods on Anxiety, Depression and Stress.

Method Name	Anxiety	Depression	Stress
PCA	[22 0 0 0 0]	[22 0 0 0 0]	[24 0 0 0 0]
	[0 39 18 0 0]	[0 39 18 0 0]	[0 29 17 0 0]
	[0 16 1 0 0]	[0 16 60 0 0]	[0 16 59 0 0]
	[0 5 0 20 0]	[0 5 0 1 0]	[0 6 0 1 0]
	[0 0 0 8 1]	[0 0 0 8 1]	[0 0 0 8 1]
Gradient boosting algorithm	[39 0 0 0 0]	[31 0 0 0 0]	[29 0 0 0 0]
	[7 21 10 0 0]	[7 21 3 0 0]	[0 20 3 0 0]
	[0 16 35 6 0]	[0 6 27 2 0]	[0 6 26 5 0]
	[0 0 0 1 0]	[0 0 0 10 0]	[0 6 0 7 0]
	[0 0 8 0 25]	[0 0 8 0 17]	[0 0 0 0 13]
ResNet 101 algorithms	[38 0 0 2 0]	[20 0 0 0 0]	[24 0 0 2 0]
	[0 30 3 0 0]	[0 21 0 0 0]	[0 16 0 0 0]
	[1 6 12 0 0]	[0 6 11 2 0]	[6 0 2 0 0]
	[0 0 0 4 3]	[2 2 0 5 1]	[0 2 0 12 0]
	[0 0 0 0 12]	[0 0 0 0 10]	[0 0 0 0 8]
KNN	[29 0 0 0 4]	[18 0 0 0 4]	[20 3 8 0 0]
	[0 17 0 0 0]	[0 15 0 0 0]	[0 13 0 0 0]
	[0 20 10 0 0]	[21 0 7 0 6]	[22 0 20 0 0]
	[5 0 0 15 0]	[5 0 0 31 0]	[0 0 0 26 0]
	[0 0 0 0 24]	[0 0 0 0 16]	[0 0 0 0 19]
Decision Tree	[20 0 8 0 4]	[18 0 0 0 4]	[28 0 2 0 6]
	[0 19 0 0 0]	[0 15 0 0 0]	[4 15 0 0 0]
	[3 0 13 6 0]	[21 0 7 0 6]	[0 0 22 0 0]
	[0 0 0 20 0]	[5 0 0 31 0]	[0 2 0 12 0]
	[0 0 0 0 2]	[0 0 0 0 16]	[0 0 0 0 2]
Naïve bayes	[12 0 2 0 2]	[18 0 2 2 0]	[20 0 2 2 0]
	[3 15 0 0 0]	[3 17 0 0 0]	[3 17 0 0 0]
	[0 5 19 0 0]	[0 5 21 0 0]	[0 4 21 0 0]
	[0 0 0 0 2]	[0 5 1 12 0]	[2 5 0 17 0]
	[0 0 0 0 2]	[0 0 0 0 2]	[0 0 0 0 5]
Support Vector Machine	[23 0 2 0 0]	[15 0 2 3 0]	[19 0 3 3 0]
	[3 8 0 0 0]	[5 2 0 0 0]	[4 7 0 3 0]
	[1 0 20 6 0]	[3 0 15 0 0]	[0 0 2 1 0]
	[0 5 0 12 0]	[5 0 0 18 0]	[4 0 0 20 0]
	[0 0 0 0 2]	[0 0 0 0 4]	[0 0 0 0 8]

the KLT algorithm has extracted the facial portion.

3.4. Feature space

Head pose: The movement of the head correlates with emotions containing happiness, confidence, fear, stress, etc. The proposed model performs head movement tracking using a Kalman filter in which X-Y-Z coordinates of various head pose has been evaluated. The X coordinate shows movement at a horizontal level. The Y coordinate reflects the head pose in the vertical direction, and the Z coordinate contains the head pose in the depth direction that contains pitch, yaw, and roll movements. These head poses are encoded into axes by the Kalman filter. The normalization of these actions has been done in the model to suppress variation caused by head pose. The head movement (M) can be calculated using the equation below.

$$\mathrm{M} = \frac{1}{N}\sum_{j=1}^{M}\frac{1}{6}\sum_{L=1}^{6}||X_L - Y_L^{tef}||$$

Here, N is the number of frames in video input. L is the number of head movement tracked by the model in X and Y coordinates. The speed (v) of head movement is computed as:-

$$\mathrm{v} = \frac{1}{N}\sum_{j=1}^{M}\frac{1}{6}\sum_{L=1}^{6}||X_L(t) - Y_L(t-1)||$$

Here, t is the angle at which head movement is recorded.

Eye gaze: The psychological emotions are also reflected by eye gaze that has been tracked using the Kalman filter. The Kalman filter is used here to define the angles at which eye movement happens. The model encoded the movement into the X and Y axis and computed the variation of eye gaze. The model normalized each person's gaze by taking the difference in median values calculated over the entire video sequence.

$$\mathrm{E} = \frac{1}{2}\sum_{j=1}^{M}\left|X_jY_{j+1} - Y_jX_{j+1}\right|$$

Here, M is the number of equal segmentation applied over X and Y coordinates by Kalman filter in order to track eye movement.

Facial Action Units (AUs): The facial action unit is encoded in terms of various facial expression movements termed as facial action units. These action units are movement/actions of localized facial muscles. The Kalman filter is used to track and localize the movement of facial muscles. It converts these movements into action units. Each action unit defines the cue of the emotional state of a person. The proposed model identifies 15 action units such as AU1, AU2, AU3, AU4, AU5, AU6, AU7, AU8, AU9, AU10, AU11, AU12, AU13, AU14, and AU15. Action units contain a variety of action intensities in the form of changing co-ordinates. These action units contain valuable features for determining

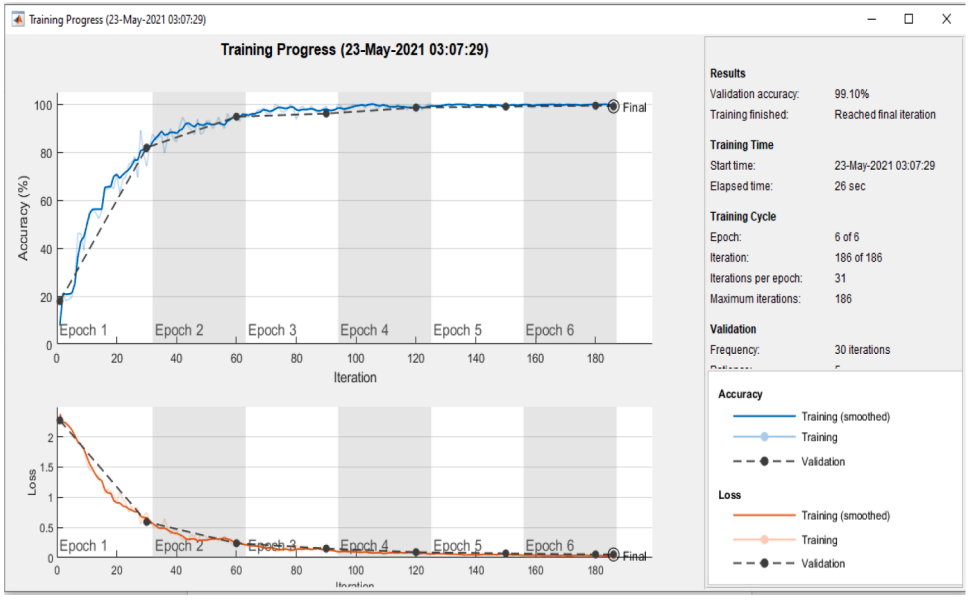


Fig. 10. Basic GUI (graphical user interface) of working model.

Table 8
Statistical measures of different classification methods.

Classifier	Mental illness	Accuracy	Error Rate	Precision	Recall	F1	ROC Area
PCA	Anxiety	0.638	0.37	0.531	0.638	0.649	0.849
	Depression	0.724	0.2744	0.550	0.724	0.730	0.850
	Stress	0.708	0.292	0.534	0.708	0.716	0.876
Gradient boosting algorithm	Anxiety	0.720	0.28	0.785	0.720	0.714	0.949
	Depression	0.803	0.21	0.826	0.803	0.807	0.930
	Stress	0.826	0.244	0.822	0.826	0.825	0.916
ResNet 101 algorithms	Anxiety	0.865	0.135	0.816	0.865	0.869	0.992
	Depression	0.838	0.139	0.806	0.838	0.850	0.980
	Stress	0.861	0.267	0.785	0.861	0.881	0.996
KNN	Anxiety	0.733	0.267	0.785	0.733	0.72	0.939
	Depression	0.707	0.293	0.7777	0.707	0.731	0.970
	Stress	0.707	0.293	0.777	0.707	0.731	0.926
Decision Tree	Anxiety	0.779	0.221	0.843	0.779	0.779	0.959
	Depression	0.867	0.133	0.914	0.8657	0.865	0.960
	Stress	0.865	0.151	0.885	0.849	0.836	0.936
Naïve bayes	Anxiety	0.769	0.231	0.808	0.769	0.760	0.899
	Depression	0.795	0.205	0.829	0.795	0.791	0.940
	Stress	0.816	0.184	0.846	0.846	0.814	0.956
Support Vector Machine	Anxiety	0.793	0.207	0.819	0.793	0.792	0.969
	Depression	0.750	0.259	0.730	0.750	0.752	0.970
	stress	0.757	0.243	0.752	0.757	0.758	0.946

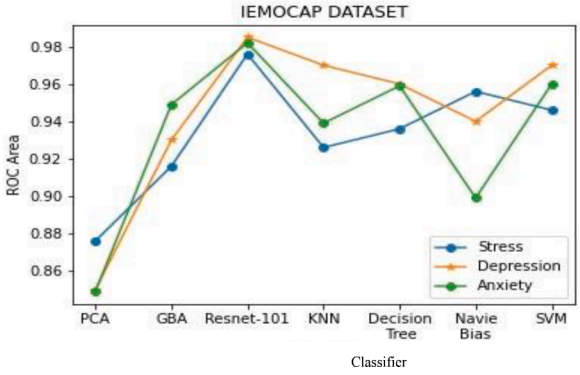


Fig.11. . Accuracy of the proposed result with the comparison of Different classifier.

stress, anxiety, and depression. The power of movement is tacked using local feature points plotted over the facial expression by the Kalman filter. Each action unit is normalized by taking the difference of 1st quartile over the entire video frame. The facial action units are stored by means of finding correlations in similar actions. The other action units that contain body heating, conductance, vibrations, small gestures, etc., are not considered in this experiment. These action units are not easily visible in the video dataset, or if they are visible, they contain much less information. The feature information in these action units may not be able to analyze through a video dataset as video frames are not very sensitive to detecting sensitive movement. Various body sensors can be used to measure action units containing sensitive features. Such action units may have overlapping features also, e.g., skin conductance can be due to various emotional factors, and so it may not be classified in specific emotion categories due to which they are not considered in the proposed scheme. The correlation (R) on features (a) is computed by following equation.

$$\mathrm{R(a)} = \sum_{J=1}^{N} a_i / N$$

The features of facial signs are the action units for which correlation has been found. The facial sign (s) is computed using binomial distribution as given below.

$$\mathrm{s(r,\ n,\ p)} = \binom{n}{r} p^r (1-p)^{n-r}$$

Here, n is the number facial sign reported the algorithm. r is the number of agreement and p is the prior probability.

Histograms: All the facial features are tracked, then transformed

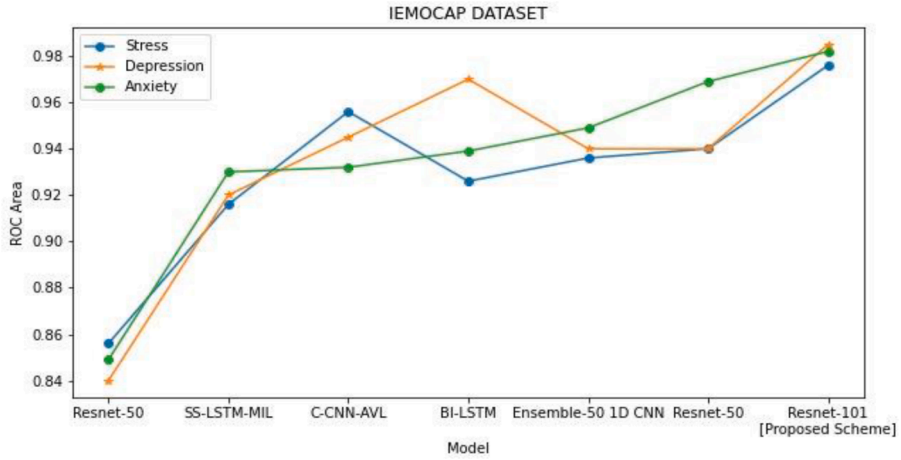


Fig. 12. ROC Area results of the proposed and several existing methods in the form of IEMOCAP Dataset.

Table 9
Shows the comparison of the proposed model (ResNet101) with other recent works.

Refs.	Dataset	Analysis	Model	Result (Average)	
Carneiro de Melo [30r]	AVEC 2013	Depression	ResNet-50	Error rate is 7.97	
Wang et al. [29]	DAIC-WOZ dataset	Depression and Anxiety	SS-LSTM-MIL	F1 score is 0.783	
Haque et al. [26]	DAIC-WOZ dataset	Depression	C-CNN-AVL	F1 score is 0.769	
	Uddin [30]	AVEC 2014	Depression and Anxiety	Bi-LSTM	Error rate is 0.74
Vázquez-Romero et al. [31]	Oz (DAIC-WOZ) database	Depression	Ensemble-50 1D CNN	Accuracy is 72%	
de Melo et al. [32]	AVEC 2013, AVEC 2014	Depression	ResNet-50	Error rate is 8.23	
Proposed Scheme	IEMOCAP [26]	Depression, anxiety and stress	ResNet-101	Accuracy is 99.4% and F1 score is 0.875	

Algorithm 1
Algorithm for feature extraction using Kalman filter.

Step 1: - Input extracted facial components
Step 2: - Time update:-
Subject state ($X_k^{'}$)
$X_k = (x_k y_k x_k^{'} y_k^{'})^T$
x_k and y_k are the series of x, y axis location in the frames of extracted facial subject from video input. $x_k^{'}$ and $y_k^{'}$ are the x_k, y_k axis speed.
$X_k^{'} = MX_{k-1} + Nw_k$
M and N are the system parameter that is in matrix form. w_k is the kalman weight used in tracking the features.
Subject error covariance ($P_k^{'}$)
$P_k^{'} = MP_{k-1}M^T + Q$
Step 3:- Measurement update:-
Calculate Kalman gain
$G_k = \frac{P_k^{'} t^T}{tP_k t^T + R}$
Here, t is the parameter in multi-measurement system.
Update estimation using measurement model (Z_k). Z_k is the observation vector.
$X_k = X_k^{'} + K_k(Z_k - tX_k^{'})$
$Z_k = tX_k + V_k$
Update error covariance
$P_k = (1 - K_k t)P_k^{'}$
P_k is the aposteriori estimate
Step 4:- Feature estimate for each pixel
$X_t = X_{t-1} + w_t(X_{t-} - L_t)$
X_t is the feature estimation for each pixel at average time t, w_t is the kalman weight for time t and L_t is the luminance intensity value of the pixel.

into histogram representation which shows the frequency distribution of the entire feature space for a person. The histogram representation contains the change in frequency based on expression changes. Histogram computation has been done for eye movement, head movement, mouth movement, and expression action units. The intensities of features are already encoded in facial action units that are transformed into histogram representations ranging from 90 to 90 in the X and Y axes. The histogram is computed using three equal-spaced bins, which contain features of stress, anxiety, and depression. Table 3 shows the histogram formation of some of the video frame that is used as a dataset in the proposed scheme.

3.4.1. Feature space extraction from images using Kalman filter

Tracking the region of interest from the sequence of the facial portion is a most sensitive task since it requires displacement mathematics for feature extraction. The Kalman filter has been used in the proposed system, which is used to track the movement and speed of the subject, such as head, mouth, eyes, etc. It first locates the region of interest on the extracted facial subject and then performs tracking in a sequence of frames where the subject's movement is effectively traced. Tracing movements of the region of interest also depends on the frame rate and the search region. Tracking is the localization of features across the frame sequence that is well performed by the Kalman filter. Kalman filter takes account of subject representation and its association for tracking purposes. Kalman filter initialized with estimating apriori parameter for the tracking of features. This parameter shows the patterns' location so that the Kalman filter can update the predicted stage. Kalman filter again performs prediction of the patterns in the next frame based on the difference of movement of information from the previous frame. The Kalman filter performs a time update reflecting the current state's forward movement with respect to time. It also estimates covariance error to compute the apriori parameter in the next frame. The model also updates the measurement in feedback that includes improved posterior values. Algorithm 1

3.5. Feature classification algorithms

The proposed system applies ResNet 101 model for the classification of the extracted feature space into stress, depression, and anxiety classes. The proposed method also uses other classification algorithms on the same feature space to compare the results and prove the efficiency of the ResNet 101. These classifiers are PCA, Gradient boosting algorithm, Dimensional reduction algorithm, KNN, Decision tree, Naïve Bayes and SVM. Fig. 7 shows several types of classification algorithms used in this paper.

A) *PCA (Principal component analysis)*

Principle component analysis is a dimensionality reduction approach in which the dimension of the feature space has been reduced into various components. Each component is individually processing for the classification be PCA algorithm. The algorithm of PCA has been described as below.

Suppose $a_1, a_2, a_3 \ldots a_n$ are the feature vectors.

Step 1:-

$$a = \frac{1}{N}\sum_{i=1}^{N} a_i$$

Step 2:- Subtract the mean:-

$$\varnothing_n = a_i - \hat{a}$$

Step 3:- For feature matrix Y=[$\varnothing_1$, $\varnothing_2$,….. $\varnothing_n$]

$$D = \frac{1}{N}\sum_{N=1}^{N} \varnothing_n \varnothing_n{}^T = \frac{1}{N} YY^T$$

Step 4:- Compute eigenvalues

$$D : \mu_1 > \mu_2 > \ldots \mu_n$$

μ_1, $\mu_2 \ldots$ μ_n are orthogonal values of input image matrix. $a - a^{'}$ is the linear combination of eigenvectors.

$$a - a^{'} = b_1\mu_1 + b_2\mu_2 + \ldots b_n\mu_n = \sum_{i=1}^{N} b_n\mu_n$$

$$b_i = \frac{(a - a^{'}).\mu_n}{\mu_n^2}$$

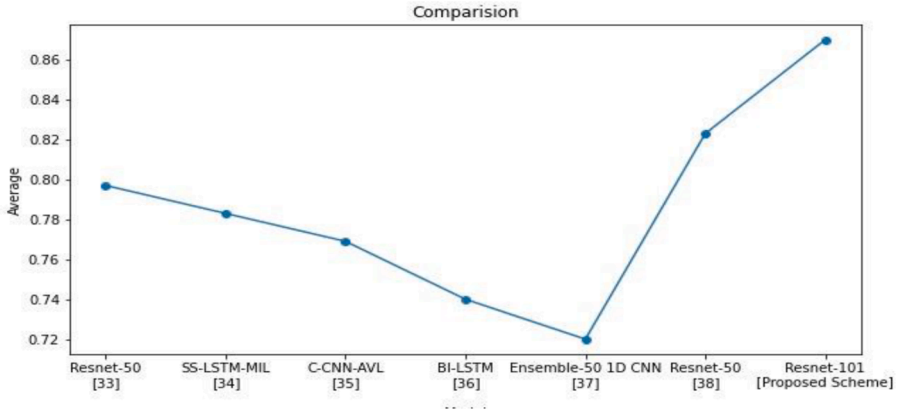


Fig. 13. Shows the comparison of the proposed model (ResNet101) with other existing works.

Step 5:- Feature reduction:-

$$a^{'} - a = \sum_{k=1}^{H} b_n \mu_n \; where \; H \gg M$$

Here, H is the highest eigenvalue.

Dimensionality reduction is performed by:-

$$\begin{matrix} b_1 \\ b_2 \\ b_k \end{matrix} = \mu^T (a - a^{'})$$

A) *Gradient boosting algorithm*

Gradient boosting algorithm is similar to adaboost algorithm. In gradient boosting algorithm, initial weights are assigned with low decision making capability. The weights are increase to covert a weak classifier into a strong classifier. The prediction capability of the model is increased gradually with the training information. Gradient boosting algorithm defines an error/loss function with is caused by the difference between actual predicted value $f(X_i)$ and the target value (Y_i). The algorithm aims to reduce this difference using weight updates. The equation of loss function is given as:-

$$L(X_i) = |f(X_i) - Y_i|$$

A) *KNN (K-nearest neighbor)*

KNN algorithm finds the nearest neighbor using similarity index generated by Euclidean distance algorithm between the feature spaces. The nearest features will belong to same class. The equation of KNN algorithm is shown below.

$$d(x,y) = \sqrt{\sum_{i=1}^{N} |x_i - y_i|}$$

Here, x and y are the two feature space.

A) *Decision tree*

A decision tree algorithm is used for classification using feature space which is translated into a tree structure in which each node contains the information gained from each feature space. The information gained from each node has been calculated to construct a decision tree. The leaf node contains the class, i.e., stress, anxiety, and depression.

Algorithm:-

Step 1:- Entropy of each feature attribute:-

$$E(X) = \sum_{i=1}^{n} - P_i \; log_2 P_i^{'}$$

Where $P_i$and $P_i^{'}$are the probabilities of occurrence and non-occurrence of features in the dataset (X).

Step 2:- Information gain (G)

$$G(X) = E(X)^{'} - E(X)$$

Here, E(X)' is the summation of entropy of the entire subset of actual dataset.

$$Split_{information}(X) = - \sum_{i=1}^{n} \frac{X_i}{X} log_2 \frac{X_i}{X}$$

$$Gain\_ratio = \frac{Gain}{Split_{information}}$$

Step 3:- After calculation of the information gain of each feature, the objective is to select the feature base that has maximum information gain to make it a root node. In this way, other internal nodes of the tree will be decided.

A) *Naïve bayes*

The naïve Bayes classifier uses the Bayes theorem to find the probability of belongingness of a feature space into a class that could be anxiety, stress, and depression.

The equation of the Bayes theorem is defined as:-

$$P(X/B) = \frac{P(B|X)P(X)}{P(B)}$$

Here, P(X|B) is the probability of X such that event B is already true. X and B are the two events. The naive bayes classifier compares the probabilities of belongingness of various feature space in a class. Feature that has maximum probability is classified into a specific class.

A) *SVM (Support Vector Machine)*

SVM is used in the proposed system to classify extracted feature space into classes like stress, anxiety, and depression. The SVM is used to plot all the training features into three-dimensional X-Y-Z planes. The SVM classifies the feature points in segregated classes using decision boundary/hyper-plane. The equation of linear decision boundary is given as:-

$$Y = A.B + C$$

Here Y is the predictor; A is the slope that decides the inclination of the best fit of the decision boundary. B is the training feature for which the predictor will be calculated. C is the intercept. SVM also draw non-linear decision boundary for the large and complex dataset in which the target value/class depends on more than one feature space. The equation for the non-linear decision boundary is given as:-

$$\sum_{i=1}^{N} \alpha_i Y_i K(s_i, z) + b = 0$$

Here α_i is the learning rate with is a constant value. s_i is the number of support vector in SVM and z is the training pattern.

A) *Proposed classification method using ResNet 101*

ResNet 101 is a neural network that contains various perceptrons and forms a learning base for the incoming feature space. The network includes an input layer at which the features are fed to the network. The hidden layer comes that performs black-box processing of the features in which the model trained itself. Then, the output layer resides where the

output of classification of features will generate that is further fed to the input layer for any error resolution. Fig. 8 shows the basic functioning of the ResNet 101 model. The Y predictor variable has been generated as output for the X input feature using the F perceptron function and b bias input. Fig. 8 shows the basic flow diagram of the ResNet 101 model.

4. Experimental results

The proposed scheme classified the obtained feature into anxiety, stress, and depression using various classification models. The confusion matrix of the respective models, including the proposed one, is described below.

Fig. 9 shows a few samples of images of people who have stress, depression, and anxiety. The proposed model's objective is to classify each individual's psychological state. The psychological conditions of an individual depend on facial features that rely on facial action units and movements.

Tables 4–6 are the score calculation of anxiety, depression, and stress, respectively. These score points are calculated on random extracted images taken from the dataset. Each category is also divided into extremely severe, moderate-severe, regular, and mild. These categories are based on score points taken based on the feature space of images.

Fig. 10 shows the GUI of a working model. The figure is divided into two sections. The upper section shows the accuracy rate. The lower section shows the error rate. The other details of the experiment are mentioned on the right side of the graph.

The confusion matrix in various algorithms reflects necessary information such as accuracy rate, error rate, precision, recall, and F1 score. These different parameters are evaluated as defined below.

$$Accuracy\ rate = \frac{sum\ of\ diagonals\ of\ confusion\ matrix}{total\ number\ of\ instances\ classify}$$

$$ErrorRate = 1 - Accuray\ rate$$

$$Precision = \frac{TP}{TP + FN}$$

$$Kappa = \frac{Total\ Accuracy - random\ Accuracy}{1 - random\ Accuracy}$$

$$F1\ Score = \frac{2 * precision * recall}{precision + recall}$$

Here TP is true positive = Diagonals of matrix
FN is false negative = sum of the corresponding row for class
FP is false positive = sum of the all corresponding column for class
TN is true negative = sum of the all row and column

Table 8 *contains the accuracy-related parameters of the various classification algori*thms. Table 2 has accuracy, error rate, precision, recall, F1 scope, and ROC area. The table shows that the average accuracy of the ResNet101 is highest compared to other classification algorithms. The proposed model has applied algorithms such as principle component analysis, gradient boosting algorithm, Naïve Bayes, K-nearest neighbor, decision tree, support vector machine, and Resnet- 101. Table 8 in the manuscript contains performance measures of all the applied algorithms in terms of accuracy, error rate, precision, recall, F1 score, and ROC area. From Table 8, the best suits algorithm in terms of robustness and accuracy is found to be Resnet-101. Various statistical measures are presented in Table 8 with the help of the confusion matrix obtained at each classification technique. It is evident from Tables 1 and 2 that ResNet 101 network is efficient for classifying the emotional status of various subjects chosen from the surveillance video dataset. ResNet 101 performs best in the neural network category and can distinguish each class at a high dimension. The performance of naïve Bayes and KNN are almost similar. The SVM algorithm also performs well as its non-linear kernel efficiently distinguishes classes in high dimensions. The performance of the decision tree is also better, but the formulation of information gained for each decision node is time-consuming. PCA classifier shows poor accuracy results in the given feature space, but it consumes less computation time.

Fig. 11 shows the comparison of various classifiers in a graphical view. The graph contains three color bars in which the green color shows the accuracy for Anxiety, the blue color indicates the accuracy for stress, and the yellow colored bar offers the accuracy for depression.

Fig. 12 shows the ROC curve comparison of various classification models on the IEMOCAP dataset. The ROC curve shows three lines for stress, depression, and anxiety. It is concluded from the figure that the ResNet 101 model is showing better accuracy on the IEMOCAP dataset as compared to other classification modes. Table 9 shows the comparison of the proposed model (ResNet101) with other recent works.

Fig. 13 shows the comparison graph of Table 9. It concludes that stress recognition-related works have been performed on various datasets in recent years. The earlier results show the detection of depression on AVEC, DAIC-WOZ, etc., datasets. Most works were accomplished using neural network models such as ResNet-50, SS-LSTM-MIL, CNN, Bi-LSTM, etc. The proposed work applied the ResNet-50 model to recognize depression, stress, and anxiety from video input. The proposed works are found to be more acceptable in terms of accuracy and precision as compared to the other techniques. The proposed model does the recognition of stress, anxiety, and depression. The stress analysis is significant as it improves real-time-based applications' efficiency and reduces medical expenses. The stress analysis is automated in the proposed model that uses artificial intelligence and its related algorithm. The proposed model applies an algorithm found to be robust for stress analysis. The significance of the proposed model is its robustness and applications that reduces space and time complexity.

5. Conclusion and future scope

. The proposed scheme is successfully able to distinguish features in anxiety, stress, and depression classes. The proposed algorithm efficiently applied pre-processing techniques to enhance the feature qualities of video frames. The proposed study performed pre-processing on surveillance video frames in which bilateral filer, adaptive filter, and contrast enhancement algorithm are applied to enhance the quality of video frames and feature statistics. These filters and techniques improve feature quality for non-symmetric feature space. The facial portion from the video frames is extracted using the Kanade-Lucas-Tomasi algorithm. The algorithm applied rectangular blocks around the facial subjects. Then Kalman filter is used for feature extraction. The algorithm applies a tracking system to capture the movement of the eye, head, mouth, and facial action units to draw out the feature points. The classification task has been carried out by various algorithms such as PCA, Gradient boosting algorithm, KNN, Decision tree, Naïve Bayes, and SVM. The proposed classification algorithm is ResNet 101, which uses a neural network to classify the features into stress, anxiety, and depression. The proposed algorithm shows an average accuracy of approximately 98.4% on the IEMOCAP [26] dataset, which is found to be higher than other classification algorithms. Hence, the proposed scheme concludes that the ResNet 101 model is very efficient in the given feature space and robust against any error. In the future, the proposed system can be extended to evaluate other psychological states, such as happiness, anger, excitement, etc., that correlate with facial expressions. In the future, other robust neural network algorithms can be applied to increase the robustness of the model. In the future, the work can be extended to apply recognition of various other emotions by using the proposed algorithms to carry out a better accuracy rate. Attacks can also be applied to video frames to check the robustness of the recognition model. In the future, the face recognition task can be performed using

thermal images from video input to minimize the noise effect. In real-time, the study for detecting stress, anxiety, and depression features significantly reduces suicidal and criminal cases. The study is also relevant to increasing business productivity by applying the model to employees to capture their moods. Automatic detection of stress, anxiety, and depression using a machine learning model is relevant to saving lives and reducing medical expenses. The study also helps improve the business strategy by identifying the customer's mood.The proposed research is relevant to assist the medical investigation in recognizing patients' psychological activities so that crimes and suicidal cases can be prevented.

Ethical approval

This paper has not submitted to anywhere and published anywhere. It does not contain any studies with human participants or animals performed by any one of the authors. The submitted work is original and not published elsewhere in any form or language

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Yang, M.H. Nguyen, M.T. Padilla, F. Zhou, F. De la Torre, Detecting depression from facial actions and vocal prosody, in: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, 2009, pp. 1–7.

[2] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, M. Breakspear, Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors, IEEE Trans. Affect. Comput. 9 (4) (2016) 478–490.

[3] T.S. Wingenbach, C. Ashwin, M. Brosnan, Validation of the Amsterdam Dynamic Facial Expression Set–Bath Intensity Variations (ADFES-BIV): a set of videos expressing low, intermediate, and high intensity emotions, PLoS One 11 (1) (2016), p.e0147112.

[4] E.B. SÖNMEZ, An automatic multilevel facial expression recognition system. Süleyman Demirel Universities Fen Bilimleri Enstitüsü, Dergisi 22 (1) (2018) 160–165.

[5] Afzali, A., Delavar, A., Borjali, A. and MIRZAMANI, M., 2007. Psychometric properties of DASS-42 as assessed in a sample of Kermanshah High School students.

[6] B. Armoon, Y. Mokhayeri, J. Haroni, M. Karimy, M. Noroozi, How is the quality of life of students? The role of depression, anxiety and stress, Polish Psycholog. Bull. (2019) 43–48.

[7] A.T. Beck, R.A. Steer, G. Brown, Beck depression inventory–II, Psychol. Assess. (1996).

[8] C. Carmassi, F. Pardini, V. Dell'Oste, A. Cordone, V. Pedrinelli, M. Simoncini, L. Dell'Osso, Suicidality and illness course worsening in a male patient with bipolar disorder during Tamoxifen treatment for ER+/HER2+ breast cancer, *Case Rep. Psychiatry, 2021* (2021).

[9] A. Ghaderi, M. Salehi, A study of the level of self-efficacy, depression and anxiety between accounting and management stu-dents: Iranian evidence, World Appl. Sci. 12 (9) (2011) 1299–1306.

[10] A. Singh, D. Kumar, Gauging stress among Indian engineering students, in: International Conference on Computational Intelligence in Communications and Business Analytics, Springer, Cham, 2021, pp. 175–186.

[11] M. ISLAM, A. KHAN, A.M.K. SHERWAN, Prevalence of iron deficiency anaemia among the reproductive age group women attending the Unani Hospital, Bangalore, Karnataka, India, J. Clini. Diagn. Res. (12) (2020) 14.

[12] J.C. Gillies, D.J. Dozois, The depression anxiety stress scale: features and applications. The Neuroscience of Depression, Academic Press, 2021, pp. 219–228.

[13] J.P. Maher, D.J. Hevel, E.J. Reifsteck, E.S. Drollette, Physical activity is positively associated with college students' positive affect regardless of stressful life events during the COVID-19 pandemic, Psychol. Sport. Exerc. 52 (2021) p.101826.

[14] H.K. Kim, Effects of stress, depression, self-efficacy, and social support on quality of life of community dwelling elderly with chronic diseases, Medico Legal Update 20 (4) (2020) 1234–1238.

[15] W.W. Zung, A self-rating depression scale, Arch. Gen. Psychiatry 12 (1) (1965) 63–70.

[16] K.J. Sher, P.K. Wood, H.J. Gotham, The course of psychological distress in college: a prospective high-risk study, J. Coll. Stud. Dev. (1996).

[17] G. Jogaratnam, P. Buchanan, Balancing the demands of school and work: stress and employed hospitality students, Int. J. Contemp. Hosp. Manag. (2004).

[18] M. Polson, R. Nida, Program and trainee lifestyle stress: a survey of AAMFT student members, J. Marital Fam. Ther. 24 (1) (1998) 95–112.

[19] N. Cahir, R.D. Morris, The psychology student stress questionnaire, J. Clin. Psychol. 47 (3) (1991), 414-4.

[20] L. Manea, S. Gilbody, D. McMillan, Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis, CMAJ 184 (3) (2012) E191–E196.

[21] S. Naveen, M. Swapna, K. Jayanthkumar, S. Manjunatha, Stress, anxiety and depression among students of selected medical and engineering colleges, Bangalore-a comparative study, Int. J. Public Ment. Health Neurosci. 2 (2) (2015) 25–28.

[22] A. Raskin, J. Schulterbrandt, N. Reatig, J.J. McKEON, Replication of factors of psychopathology in interview, ward behavior and self-report ratings of hospitalized depressives, J. Nerv. Mental Dis. (1969).

[23] M. Shah, S. Hasan, S. Malik, C.T. Sreeramareddy, Perceived stress, sources and severity of stress among medical undergraduates in a Pakistani medical school, BMC Med. Educ. 10 (1) (2010) 1–8.

[24] P. Svanborg, M. Åsberg, A comparison between the Beck Depression Inventory (BDI) and the self-rating version of the Montgomery Åsberg Depression Rating Scale (MADRS), J. Affect. Disord. 64 (2–3) (2001) 203–216.

[25] P. Vitasari, M.N.A. Wahab, A. Othman, T. Herawan, S.K. Sinnadurai, The relationship between study anxiety and academic performance among engineering students, Procedia-Soc. Behav. Sci. 8 (2010) 490–497, 18 above18 above18 above.

[26] Haque, A., Guo, M., Miner, A.S. and Fei-Fei, L., 2018. Measuring Depression Symptom Severity from Spoken Language and 3D Facial Exressions. arXiv Preprint. arXiv:1811.08592.

[27] Aziz, M., 2004. Role stress among women in the Indian information technology sector. Women in Management Review.

[28] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Langu. Res. Evalu. 42 (4) (2008) 335–359.

[29] Y. Wang, J. Ma, B. Hao, P. Hu, X. Wang, J. Mei, S. Li, Automatic depression detection via facial expressions using multiple instance learning, in: Proceedinngs of the IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 1933–1936.

[30] M.A. Uddin, J.B. Joolee, Y.K. Lee, Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm, IEEE Trans. Affect. Comput. (2020).

[31] A. Vázquez-Romero, A. Gallardo-Antolín, Automatic detection of depression in speech using ensemble convolutional neural networks, Entropy 22 (6) (2020) p.688.

[32] W.C. De Melo, E. Granger, A. Hadid, Depression detection based on deep distribution learning, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 4544–4548.

Divya Kumar is an Assistant Professor in the Department of Computer Science and Engineering at Motilal Nehru National Institute of Technology Allahabad, India. He received his PhD Motilal Nehru National Institute of Technology Allahabad, India His-research interest are software engineering, Soft computing, evolutionary optimization and reliability engineering and Machine Learning

Astha Singh is a PhD Candidate in the Department of Computer Science at Motilal Nehru National Institute of Technology Allahabad, India. She received his M.tech in Computer Science from Centre for Advanced Studies Lucknow, India. Her research interests include Machine Learning, Natural Language Processing