IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 December 2023, accepted 4 January 2024, date of publication 11 January 2024, date of current version 22 February 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3353053*

RESEARCH ARTICLE

# Multimodal Engagement Recognition From Image Traits Using Deep Learning Techniques

**AJITHA SUKUMARAN**[1], **(Senior Member, IEEE), AND**
**ARUN MANOHARAN**[2], **(Senior Member, IEEE)**
[1]School of Electronics Engineering, Vellore Institute of Technology, Vellore 632014, India
[2]Embedded Technology, School of Electronics Engineering, Vellore Institute of Technology, Vellore 632014, India

Corresponding author: Arun Manoharan (arunm@vit.ac.in)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

**ABSTRACT** Learner engagement is a significant factor determining the success of implementing an intelligent educational network. Currently the use of Massive Open Online Courses has increased because of the flexibility offered by such online learning systems. The COVID period has encouraged practitioners to continue to engage in new ways of online and hybrid teaching. However, monitoring student engagement and keeping the right level of interaction in an online classroom is challenging for teachers. In this paper we propose an engagement recognition model by combining the image traits obtained from a camera, such as facial emotions, gaze tracking with head pose estimation and eye blinking rate. In the first step, a face recognition model was implemented. The next stage involved training the facial emotion recognition model using deep learning convolutional neural network with the datasets FER 2013.The classified emotions were assigned weights corresponding to the academic affective states. Subsequently, by using the Dlib's face detector and shape predicting algorithm, the gaze direction with head pose estimation, eyes blinking rate and status of the eye (closed or open) were identified. Combining all these modalities obtained from the image traits, we propose an engagement recognition system. The experimental results of the proposed system were validated by the quiz score obtained at the end of each session. This model can be used for real time video processing of the student's affective state. The teacher can obtain a detailed analytics of engagement statics on a spreadsheet at the end of the session thus facilitating the necessary follow-up actions.

**INDEX TERMS** Deep learning, engagement recognition, facial expression, gaze direction, head-pose estimation.

## I. INTRODUCTION

The students of the 21st century are digital natives, and their skills to learn and adapt to the digital environment are remarkable. New ways of online and hybrid teaching have emerged in the post-COVID period. Both students and instructors are comfortable with the flexibility provided by online teaching and learning. The advantages of online sessions include saving time, flexibility, the possibility of recording classes and future access to these resources. The student cohorts of each session are usually from different socioeconomic backgrounds. Because of this diversity, they exhibit different learning styles, rate of learning speed and variation in motivation levels. In an online classroom, it is difficult for the instructors to keep track or monitor each student's activities or the level of their engagement during class. But for an offline classroom, this is possible to a certain extent with a smaller cohort strength. Engagement or student attention during the class refers to their active involvement in the learning activity [1]. Hence, student attentiveness is directly proportional to the acquired learning outcomes for the session.

Research indicates that the attention span of students reduces in the online environment compared to the traditional classroom [2]. In an online session, students cannot

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

physically interact with their peers and instructors. The literature has recorded several methods for measuring the engagement of students, including self-report, observational checklists, and automatic engagement recognition methods [3], [4], [5], [6]. Self-reports are survey questionnaires which help the instructor to understand the level of involvement, frustration, excitement, confusion, and boredom of students towards the lesson content, but the accuracy of the survey depends on the respondent's positive and negative attitude towards the teacher or the class. Observational checklists are questionnaires required to be filled in by external reviewers based on the subjective opinion of teachers on the behaviour and responses of students during class. Reviewers could also consider samples of student submissions. Both the self-report and observational checklists are time-based solutions, as they require extra time and effort from the students or reviewers; the findings may not always accurately indicate the engagement level of students.

Buntins et al. [3], provided a comprehensive review on the available instruments for assessing the student engagement in different education contexts by identifying the trends, uniformity, and gaps. They highlight the requirement of developing a unified and consistent approach for measuring student engagement. Veiga et al.. [4]evaluated the reliability and validity of these instruments in measuring student's engagement in diverse learning environments incorporating the psychometric qualities. As per Mandernach [5], the engagement level of the student is dynamic in nature and hence it requires a multifaceted design of assessments which captures affective, behavioural and cognitive aspects of student engagement. Fredricks et al. analyzed the pros and cons of assessing student engagement using self-reports, observational check lists, teacher ratings, interviews, and experience sampling techniques. They evaluated eleven self-report survey measures of student engagement, for measuring the consistency and authenticity of the existing information on each measure. They concluded that the self-report survey methods are more effective for obtaining the cognitive or emotional engagement of student [6]. However, when internal experiences are not easily observable, employing biometric measures or physiological signal analysis can be an effective approach to capture the engagement status of students.

Artificial Intelligence is widely used in educational practices (AIEd) [7]. Considering the different pedagogical perspectives in teaching and learning, AI can play the role of helping instructors improve the student learning experience and motivation that leads to better achievement of learning outcomes [8]. The progress in technologies such as Internet of Things (IoT) and the availability of Big Data have strengthened this trend. The developments in deep learning algorithms yielded significant results in many of the social signal processing problems [9], [10], [11], [12]. Roust et al. [9] and Kumar [10] examined the automatic affect recognition of humans using deep neural networks. They reviewed around 950 studies on deep learning conducted from 2010 to 2017. They reveal the trend of using deep learning in this field. Affective computing of large volumes of multimedia data available from social networks can be used to understand the behaviours and actions of humans and potentially, possess wide applications [11]. The recognition of human emotions from text data by exploring deep learning techniques was proposed by Kratzwald et al. [12]. The authors used the transfer learning approach with the customized Recurrent Neural network to achieve the task of emotion recognition and demonstrated that the deep learning approaches outperform machine learning algorithms.

The automatic tracking of engagement is essential for increasing the learning efficiency of students. Nowadays, research attention is increasingly focusing on the use of automatic methods for the behavioural analysis of students in the classroom through engagement tracking models. The automatic engagement recognition methods are physiological or neurological signals-based and computer vision-based approaches. The computer vision approaches are based on the image traits obtained from a camera like facial expressions, yawning detection, gaze direction and gesture and posture analysis. Various eye tracking devices are used for tracking the visual attention of learners by analyzing saccades, fixation, and blinks. In neuroscience, engagement is evaluated by the level of valence and arousal. Physiological signals associated with engagement can be obtained from sensors such as Electroencephalogram (EEG), Electrocardiogram(ECG), Galvanic Skin Response(GSR) and Blood Pressure(BP). Sukumaran et al. [13] conducted a review on various automatic attention tracking approaches and algorithms, both in online and traditional classes based on computer vision, physiological and neurological signals. Their survey examines the effectiveness, challenges, and advancements of these methods, by providing substantial ideas for researchers and educators.

Current research on Engagement analysis has identified the following research gaps.

- Lack of comprehensive datasets for engagement recognition with natural expressions.
- In the estimation of the gaze direction, the case of occlusion of eyes because of wearing glasses makes it difficult to predict gaze.
- Lack of research on the fusion of physiological sensor data with facial expression as most of the research is based on the single modality of facial expression.
- The available multiple sensor gadgets lack mechanisms in analyzing or processing data, as they allow monitoring only.
- Limited datasets for EEG, and physiological signals for emotion recognition are available.
- Each researcher uses different performance evaluation metrics, making it difficult to benchmark the result analysis.

We have proposed a computer vision based automatic engagement tracking system for students in an online classroom with face recognition, by analyzing the facial expressions, eye gaze with head pose estimation and eye

blinks. We estimated that the gaze direction based on head pose removed the problem of predicting gaze from wearing glasses. The key highlights of the paper are.

- Implementing a face recognition system for identifying the name of the student by providing a database of the images of all students in the respective session.
- Developing a CNN based facial emotion recognition model with FER 2013 dataset with a train and test accuracy of 95.6 and 73.4 respectively.
- An 'Engagement Indictor' algorithm is proposed by combining the result of eye gaze with head pose estimation, eye blink and with the classified facial emotions.
- Validating the results of the proposed engagement recognition system with the formative assessment Quiz score of each student for a one-hour session conducted on three consecutive days.
- The detailed analytics of engagement is available on a spreadsheet at the end of the session; and hence the teacher can take the necessary follow up actions.

We have designed a deep learning model for facial emotion recognition. A convolution Neural network (CNN) with MobileNet V2 architecture was trained using FER 2013 dataset to get a rich facial emotion recognition model to achieve state-of-the-art performance. A new multimodal student engagement recognition method was implemented based on the expressions identified by the facial emotion recognition model along with outputs of eye gaze direction with head pose estimation and the eye status. The training and testing of the proposed work were done using TensorFlow 2. Metrics such as Accuracy, Precision, Recall, and F score are used to demonstrate the performance of the implemented facial emotion recognition model.

Our work is a novel technological frame for predicting engagement status of students by combining the multiple modalities from the image traits. The remaining part of the paper is organized as follows. Section II discusses the related works, Section III presents the materials and methods, Section IV explains the experimentation methodology with results and discussion. Section V concludes with a summary of the main findings, potential applications of our work, and future research directions.

## II. RELATED WORK

The human face usually reflects the emotions of a person. Constant monitoring of facial expressions can help in the detection of the mental state of the person. Recently various studies have been conducted on different used cases of FER such as identifying emotions of autism affected children and persons with intellectual disabilities [14], identification of customers' emotions in marketing [15], drowsiness alerting system for drivers in smart cars [16], and testing of video games [17]. Regarding educational applications, tracking the engagement of students is very important for an online and offline teaching and learning process.

The engagement tracking methods are classified as manual, semi-automatic and automatic [18]. Due to the development of computer vision techniques and wide acceptance of Artificial Intelligence effects in all areas, the automatic engagement recognition techniques using machine learning [19]and deep learning [20] techniques have been gaining attention in recent years. There are many research studies on automatic engagement tracking in online and offline class scenarios, either using single modality or by combining modalities [20], [21], [22], [23], [24], [25]. One of the pioneering studies on Facial Emotion based engagement recognition of students in a computer-based environment which was conducted by Whitehill et al. [19] to assess the effectiveness of the existing state of the art computer vision architectures. Classifiers such as Gentle Boost, Support vector machines (SVM) and Multinominal Regression (MLR) based on the appearance-based features of the human face such as Gabor filter, Computer Expression Recognition Toolbox (CERT), Facial Action Coding system (FACS) and box filters were used. They used manual rating for the annotation of video recordings and assigned a single label to the 3-minute video. For more precise recognition of engagement, a continuous labelling is required. The conclusions drawn from their study were that the SVM classifier performed well compared to other classifiers for their generalized model across ethnicity.

Reliable models are very important for accurate recognition of engagement in a learning environment. Recent studies show that deep learning-based models show promising results. A real-time engagement recognition system based on the analysis of facial expressions using deep learning model was proposed by Gupta et al. [21]. Engagement index is calculated based on the facial expressions and the output is classified as ''Engaged ''or ''Disengaged''. The authors used Faster RCNN for the detection of face and MFACXTOR for the extraction of facial points. FER 2013, Ck+ and their own datasets were used for training using the deep learning models Inception-V3, ResNet50 and VGG-19. They tested the system on 20 learners in an online class. Their results confirmed that ResNet-50 outperformed all others and obtained a real time accuracy of 92.32% for their own dataset. Recent studies in Deep learning models using Inception V4 and ResNet showed improved classification accuracy in recognition problems. However, the light weight models can better resolve the efficacy issues with mobile or other embedded devices [20]

A few studies have adopted the approach of extracting the cognitive, affective, and behavioural dimensions of engagement by combining body movements with facial expressions in a gamified learning environment [22], [26]. Savchenko et al. [23], analyzed the behaviour of students in an online classroom by video processing the facial expressions using a single neural network. They used the MobileNet and EfficientNet based light weight FER architectures for the extraction of emotional features from facial images. In order to predict the emotions of static images, the model was trained by modifying the softmax loss function with the theory of robust data mining. The Affectnet dataset was used for their study. Their CNN model was able to recognize eight emotions

on static images from Affectnet dataset and 7 basic emotions from AFEW dataset, and three affects (Positive, Neutral and Negative) from the VGAF dataset. The behaviour analysis of learners was performed by an application installed in his or her device and the report was sent only to the teacher; and thereby ensuring data privacy. But their proposed approach has lower accuracy for multimodal methods using VGAF and AFEW datasets. For predicting and understanding the affective attitude of students, it is very important to predict and analyze the valence and arousal [27].

Buono et al. [24] conducted, a related study used EmotiW 2019 challenge datasets to develop a model for assessing the engagement level of the learners from facial expressions, gaze and head poses based on the Long Short-term memory (LSTM) networks. In their experiment, the videos of students attending an online class were analyzed for two different modalities: slides and video lecture. The intention of the work was to identify the inclination of engagement towards these two modalities. By taking perceived emotion Index in both the cases, the self-report on engagement and the study based on the automatic engagement detection revealed that the engagement level of students was comparatively lower during the slide share compared to the video lesson. The Pearson Correlation analysis on automatic engagement detection was weakly correlated to the self-report method; however, the correlation was stronger for the emotion dimension of engagement as facial action units and head pose were concurrent with it. But for the gaze, it was observed that more engaged students have less gaze movements. Typically, the response of a person to a direct gaze is quite high with shorter fixations [28]. The students who gazed more at the teacher's eye conveyed a feeling that they are more focused and active for in-class activities. Hence the frequency of gaze is associated with interest.

Human emotions are categorized as primary and secondary emotions. The four basic emotions of humans are sadness, happiness, anger, and fear. The secondary emotions are derived from the primary emotions [29]. Academic emotions such as engagement, boredom, frustration, and confusion are secondary emotions. A few datasets such as HBCU [19], Daisee ( [30], In-the-Wild dataset [31] and OL-SFED [32]) are available for academic emotions. Many of the existing research for engagement recognition is based on the publicly available databases HBSC, Daisee and in the wild datasets. A short summary of publicly available datasets for facial emotion recognition and academic emotion of affective states along with its features are provided in the review articles [13], [33]

Based on the review of the literature, it is evident that for determining the attention level of students, facial expression and gaze are critical modalities. The head pose has a substantial role in determining the gaze direction. CNN architectures attained the state-of-the-art performance in computer vision-based problems making them apt for classification of engagement recognition task. But the research in this area is limited due to the lack of availability of exclusive dataset for the purpose, difficulties in fusing multimodal data and executing experiments in real time environment. In this study, we aim to develop an efficient engagement recognition model by combining the cues obtained from facial expressions, eye gaze with head pose alignment and eye status. The lack of availability of exclusive datasets for academic emotions taken for the online classroom conditions will reduce the accuracy of training models for the purpose. Hence here we didn't directly take the academic emotion dataset, rather, we used FER 2013, a large dataset which classifies facial emotions accurately.

## III. MATERIALS AND METHODS

This study attempted to measure the emotional engagement of students in an online classroom session through computer vision techniques. This involved recognizing the student's face and detecting the engagement status by analyzing the facial features. Facial analysis is easier to implement compared with physiological signal analysis as it is low cost, since the image is captured using the webcam. The facial expression classification was implemented using the Convolutional Neural Network (CNN) architecture. CNN is a type of multilayer neural network used for deep learning algorithms to recognize visual patterns directly by processing the data. With the LeNet architecture introduced in the 1990s started the era of deep neural network in practice. Another architecture, Alexnet [34], with more filters in each layer, deeper and with stacked convolutional layers won the first prize in imagenet competition in 2012, by achieving the best results. Later other deep learning network architectures like VGGNet [35], [36], GoogleNet [37], [38], ResNet [39], [40], and Efficientnet [41], [42] were designed for improved accuracy. But all these traditional CNN require a large memory and considerable computational time. Hence it is difficult to run on mobile or embedded devices. In online classes, the end device used by students is not restricted to laptops, they could also use smart phones. We chose the lightweight CNN architecture Mobilenet for our proposed model. Mobilenet V1 replaced the standard convolution layer with a depthwise separable convolution operation which reduces the number of model parameters, leads to the reduction of computation cost and model size. Hence this architecture has become well suited for mobile devices. Mobilenet V2 is faster, uses fewer operations, applies fewer parameters, and has high accuracy, when compared to mobilenet V1. For training and testing of emotion recognition model, we used mobilenet V2 architecture.

### A. PROPOSED METHODOLOGY

The framework of the proposed system is shown in figure 1. The programming language used is Python and it was implemented in deep-learning architecture by Tensor Flow. Haar cascade from open cv is used for face detection and face cropping and Mobilenet V2 for the classification of emotions. The region of interest of the face is given to Mobilenet V2 to extract features through different layers and provide seven classifications of emotions. The gaze direction with head pose

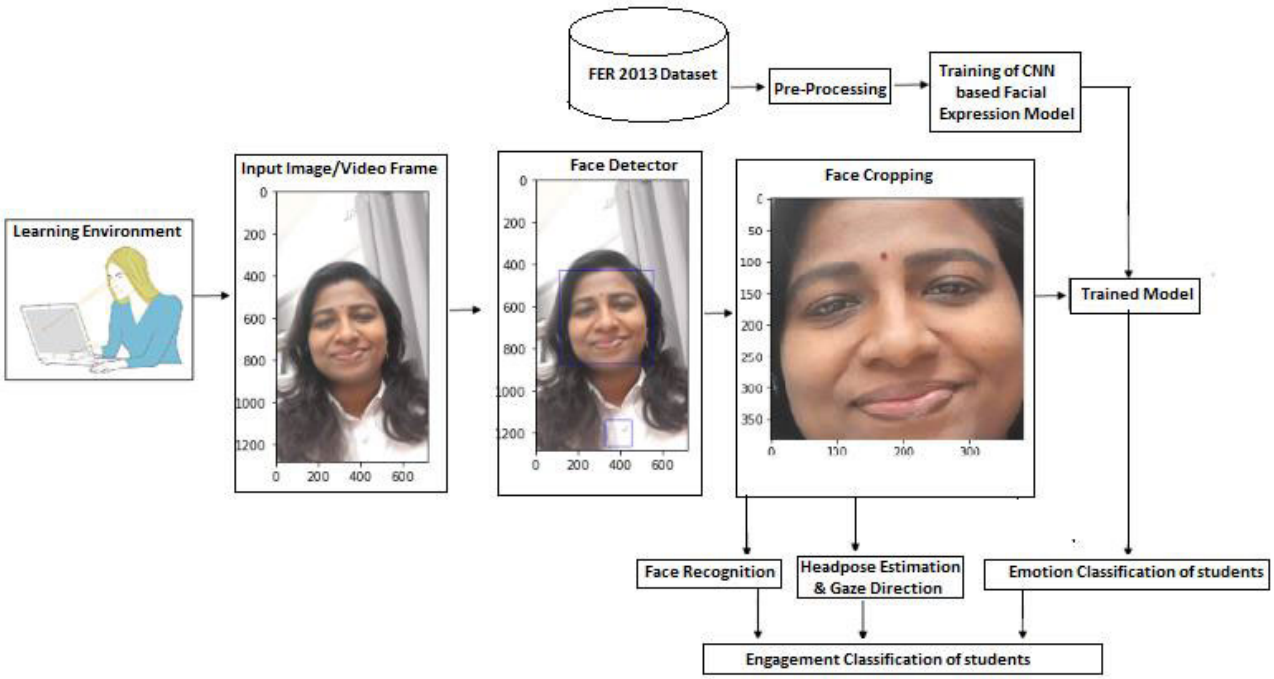


**FIGURE 1.** The proposed framework for student engagement detection.

estimation and eye aspect ratio for detecting eye status are implemented using Dlib model. The engagement classification unit combines the returned aggregated results of each modality for the video frames for the duration of 60 seconds.

### 1) DETECTION OF FACE USING HAAR CASCADE

Haar Cascade is a very popular face detection algorithm. This algorithm provides high accuracy and speed with less computation complexity and is evidenced in a lot of research works [43], [44], [45], [46]. The main four steps of haar-cascade algorithm are Haar Feature selection, Creation of Integral Images, AdaBoost Training and Cascade Classifier. Haar features are an arrangement of square shape functions (dark regions and light regions) introduced by Alfred Haar in 1909. As per Viola-Jonas algorithm [47], there are mainly three types of haar features - edge, line and four rectangle features.

The Haar features as shown in figure 2 are applied to the relevant areas of the face to detect the human face. For each feature calculation, the difference between the average pixel intensity values of dark regions and the average pixel intensity values of light regions are calculated; it is a single valued function.

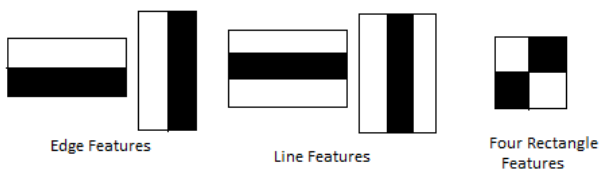


**FIGURE 2.** Examples of Haar features.

$$S = \frac{1}{n}\sum_{i=1}^{n} D(i) - \frac{1}{n}\sum_{i=1}^{n} W(i) \quad (1)$$

where $D(i)$ is the value of $i^{th}$ pixel in the dark region and $W(i)$ is the value of $i^{th}$ pixel in the light region, $n$ is the number of pixels and $S$ is the single valued function. Based on the value of $S$, which is compared with the threshold, the haar feature is detected.

In the image analysis, the best features from all the available features were taken by Adaboost training. The Haar cascade detection in open CV includes the pre-trained classifiers, including the face, eyes and smile. For our implementation the input image for face detection is converted to gray scale and loaded to haar cascade xml classifier file. The face in the gray scale image is detected by returning the boundary rectangles of the detected face as x,y w and h. and draw over the detected faces. For the case of real time videos, are divided into frames and each frame is converted to gray scale. We used 'detectMultiScale' function, as different frames may have different scales.

### 2) THE FACE IDENTIFICATION SYSTEM

The face identification system is implemented in three steps: initial database construction, facial feature extraction and face embedding, and finally the face identification. Images of every student in the session, each linked with their respective names, have been compiled in the database and stored in a designated folder. The extraction of facial region from the input image is performed by Haar Cascades in Open CV. The key points in the face such as the nose, mouth and corners

of eyes are identified with the help of the pretrained facial landmark detection model 'dlib' [48]. The model detects and locates the face and embeds the located face. As per the detected landmarks, normalizing the pose and scaling of face alignment for face is attained by using the eye position. The facial features or facial embeddings are obtained from the dlib pre-trained face recognition model which is able compute 128-dimensional face embedding vector from the aligned face image [49]. The network generates almost the same values when comparing the frame of the student with the picture of him/her from the database else, it generates a different value when it is not matching. Finally in the face recognition stage, the extracted face embedding vectors are compared with known face embeddings of the initially created database. This is done by using the distance metric, Euclidean distance to measure the distance between the feature vectors. Based on the threshold value, the matching face is identified.

The images of the students for the experimental study are saved as .jpg file with file name as the 'student name' (student1, student 2) in a folder. Initially, these images from the folder read as list and encode the images for facial landmarks. Then a csv file with ''name'', ''date'' and ''time'' is created. The split up of the line is done by a comma, which separates each data by column. 'Date-time class' provides the current date and time while taking the class. The webcam captures the student's face, and the face detection and cropping are done by haar-cascade. After that, the face compare function finds the match for the registered users by lowest face distance calculation. If the detected face matches with the registered images, the name of the student is recorded to the CSV file, with the date and time.

### 3) GAZE TRACKING AND HEADPOSE ESTIMATION

We developed the eye tracking system integrated with head pose estimation which provides the correct location of pupils and hence the direction of gaze in real time. The Dlib's shape Predictor is used to identify facial landmarks to track the eyes and locate pupils, and to understand the gaze direction. The implementation of gaze tracking system is done for each of the following modules written in different classes.

- Isolating eye and initiating pupil detection
- Calibrating the pupil detection algorithm from the optimum binary threshold value obtained for the webcam image.
- Detecting the iris and estimating the pupil position.
- Tracking user's gaze for understanding the position of eyes, pupils and allows to know whether the eyes are closed or not(blinking).

Figure 3 provides the details of gaze tracking based on the pupil and iris position.

The head pose of a person is obtained from the translation and rotation vector. From our two-dimensional image frame, the pose is estimated by Perspective-n-Point (PnP) solution.


**FIGURE 3.** Gaze direction based on pupil's detection.

The PnP problem equation is of the form.

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (2)$$

where, $f_x$ and $f_y$ are the focal point from the width of the image, $\gamma$ is the skew parameter, $u_0$ is the image width and $v_0$ is the image height.

From (2), it is possible to get back the rotational and translation matrices. But this equation has inputs of 2D co-ordinates in the image space, 3D co-ordinates in the world space and the camera parameters such the focal point, center parameter and skew parameter. For the estimation of head pose, we took six 3D model points from the face landmark detection algorithm. The points are the left and right edges of eye, left and right of mouth corner, the nose tip and the chin.

The conversion of rotational vector to matrix format is done by using Rodrigues rotation formula [50]. The Euler-Rodrigues formula provides the rotation matrix for the circular movement of a vector around the specified axis.

$$v_{rot} = \bar{v}cos\theta + \langle \bar{v}\ , \bar{k} \rangle \bar{k}.\,(1 - cos\theta) + \left(\bar{k} \times \bar{v}\right).sin\theta \quad (3)$$

where $v_{rot}$ is the resulting new vector, $\bar{v}$ is the original vector, $\bar{k}$ is the unit vector describing the axis of rotation and $\theta$ is the counterclockwise rotation angle.

The equation (3) can be extended for transforming all the three basis vectors to obtain the rotation matrix in the 3D rotation group (SO(3)).

$$R = cos\theta.I + sin\theta \begin{bmatrix} 0 & -k_z & k_y \\ k_z & 0 & -k_x \\ -k_y & k_x & 0 \end{bmatrix} + (1 - cos\theta)\,\bar{k}.\bar{k}^T \quad (4)$$

where $R$ is the rotational matrix and $I$ is the Identity matrix of order $3 \times 3$.

The extraction of the three Euler angle of rotation is done by the function cv2.decomposePrjectionMatrix. It computes the RQ decomposition from the obtained rotations and decompose the left $3 \times 3$ *submatrix* of the project matrix to the camera and rotation matrix. Finally, it projects a Jacobian matrix point with respect to rotation and translation direction to draw a line of direction.

Figure 4 provides the test results of the module, gaze direction with the estimated head pose and Figure 5 provides the respective camera matrix, rotation and translation vector details. Table 1 shows the weights assigned for gaze direction

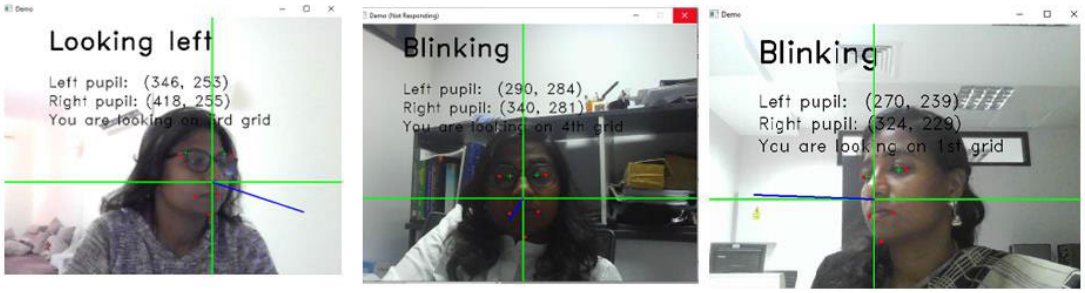


**FIGURE 4.** The gaze tracking with head pose estimation with quadrant information.

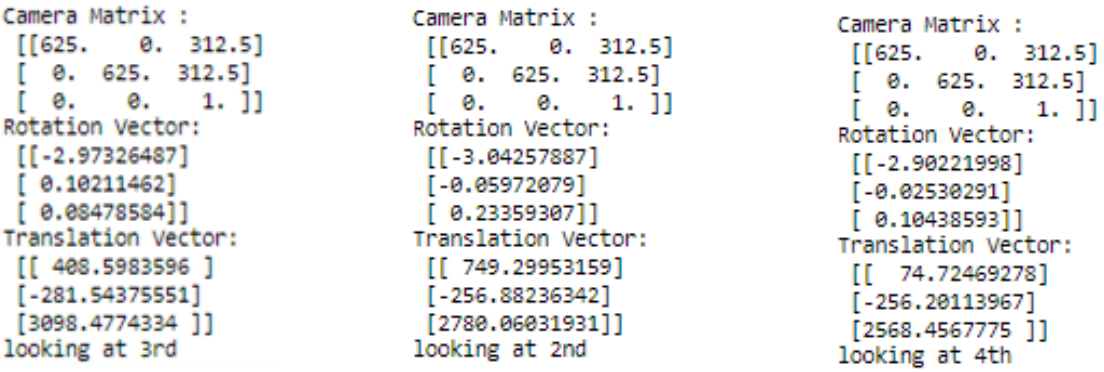


**FIGURE 5.** The camera matrix, rotation and translation vector associated with the head pose in different quadrants.

with head pose estimation and returned for the calculation of engagement indicator.

**TABLE 1.** Weights of Gaze direction with head pose estimation for engagement recognition.

| Look Center | Left head pose/Left gaze | Right Head pose/Right Gaze |
|---|---|---|
| 1 | 0 | 0 |

#### 4) EYE OPEN AND CLOSE STATUS WITH EYE BLINK

Eye Aspect Ratio is the scalar quantity, which provides the status of eye, whether it is opened or closed, by calculating the Euclidean distance of the respective eye coordinates. EAR formula does not depend on the gaze direction.

$$EAR = \frac{|P2 - P6| + |P3 - P5|}{2 \times |P1 - P4|} \tag{5}$$

$$Average\ EAR = \frac{EAR_L + EAR_R}{2} \tag{6}$$

Equation (5) and (6) are the EAR equations, where $P_1, P_2, P_3, P_4, P_5$ and $P_6$ are the 2D landmarks points as shown in figure 6. Distance between $P_1$ and $P_4$ measure as eye width; and distance between $P_2$ & $P_3$, $P_5$ & $P_6$ are used to provide the eye height. When the eye is open, the EAR

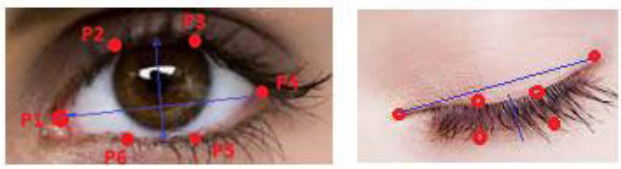


**FIGURE 6.** 2D landmark points of an open and closed Eye.

value is a constant and it is nearly equal to zero, when the eye is closed. As always blinking will occur simultaneously for both eyes; as per equation 2, we have averaged the left and right eye EAR.

In our study we selected the EAR threshold as 0.25. If $EAR > 0.25$, eye is open,else eye is close. Table 2 provides the weights assigned for eye status and returned for the calculation of engagement indicator.

**TABLE 2.** The weights of eye open/close for engagement recognition.

| Eye Close | Eye Open |
|---|---|
| 0 | 2.5 |

#### 5) FACIAL EMOTION RECOGNITION

The facial emotion recognition system consists of acquiring image/video, preprocessing, feature extraction and emotion classification. The captured faces by webcam undergo

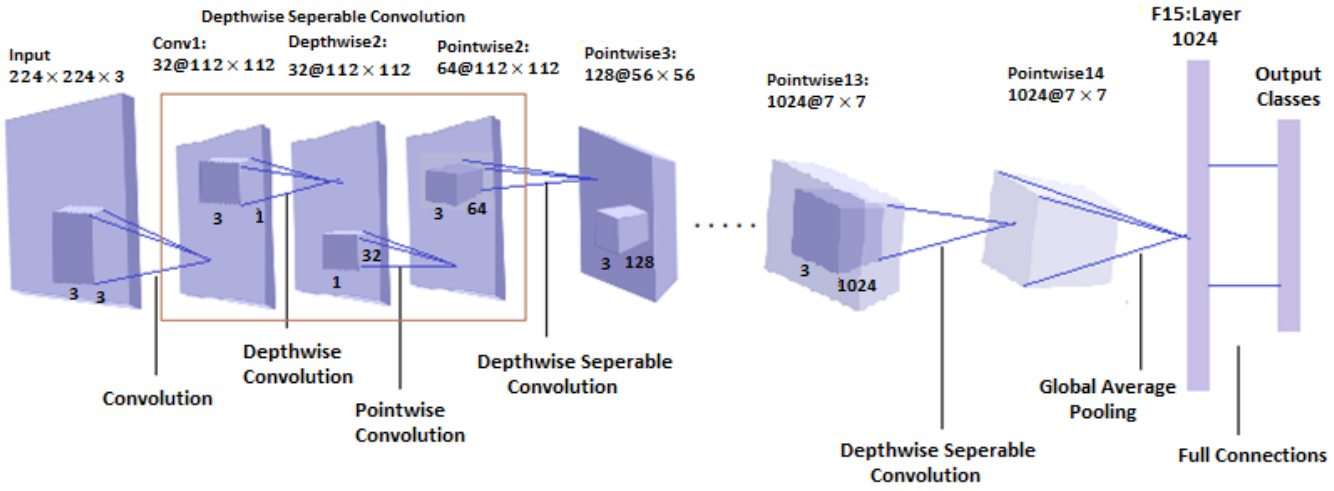


**FIGURE 7.** Mobilenet V2 architecture.

preprocessing stage. The main libraries used for our experiment are Tensorflow, Keras, numpy, Open CV, matplotlib and scikitlearn. Tensor flow was our system backend, although all the built-in functions like activation function, optmizers, layers etc. were provided by Keras. Open CV was mainly used for image preprocessing stage, Matplot lib for graph plotting and scikitlearn for generating the confusion matrix. From the perspective of model practicability and minimized computational complexity, we chose mobilenet V2, a lightweight CNN architecture for system training. Figure 7 depicts the architecture of mobilenet V2. It employs depth-wise separable convolutions, which reduce the number of parameters, compared to other CNN architectures. Therefore, it is a light weight deep neural network [51].

Dataset:

The dataset selected is expected to be realistic and for our implementation, we used FER 2013 dataset, downloaded from Kaggle. In FER 2013, there were 35887 grayscale images of size 48 × 48 in seven emotion categories: angry, fear, disgust, happy, sad, neutral and surprise.

As a few of the images did not correspond to the respective class, a manual filtering of images was done at the beginning and the images with occlusion, contrast variation, imbalance problem, eye glass problem and intra class variations were removed. Finally for implementation, we considered 19706 images together in all 7 classes.

The total number of images were divided to the ratio 7.5 : 2.5 for training and testing purposes. In the preprocessing step the images in the dataset were resized (upsampled using bilinear interpolation) to 224 × 224. As we used transfer learning, the input image size of the deep learning algorithm is 224 × 224 pixel. All the images of the training dataset folder were read and converted to an array format. Then the images were shuffled, as our deep learning model should not learn the sequence and must be dynamic and robust. Later the array of images was transformed to

**TABLE 3.** The number images for different emotions for training and testing.

| Emotions | No: of Images Taken for Training | No: of Images Taken for Testing |
|---|---|---|
| Angry | 2442 | 650 |
| Disgust | 319 | 111 |
| Fear | 1979 | 900 |
| Happy | 3202 | 1014 |
| Sad | 2551 | 810 |
| Surprise | 1917 | 612 |
| Neutral | 2394 | 805 |
| Total: | 14804 | 4902 |

**TABLE 4.** Final classifier model layer details.

| Layer(Type) | OUTPUT SHAPE |
|---|---|
| Global Average Pooling 2D | (None,1280) |
| Dense | (None,128) |
| Activation | (None,128) |
| Dense1 | (None,64) |
| *Activation1* | (None,64) |
| *Dense2* | (None,7) |

4-dimensional array with total number of images, image size and number of channels (14804 × 224 × 224 × 3). Then the normalization of data was done by dividing it by 255 (The maximum pixel value of any image in the array). One hot encoding is used to generate a layered label for each image. We labelled 0,1,2,3,4,5 and 6 for the emotion categories starting from angry to Neutral as per the order in Table 3.

The model was trained using transfer learning. We downloaded the existing trained model from mobilenet V2 and this was already trained for 1000 classes. The last layer was −1, then we removed the last layer and output layer is changed to -2. We added a new layer after the output of global

**TABLE 5.** The weights of facial emotions for engagement recognition.

| Emotions | Angry | Sad | Fear | Surprise | Disgust | Happy | Neutral |
|---|---|---|---|---|---|---|---|
| Weight of Emotions towards Engagement | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 1.1 | 1.4 |

pooling layer and three fully connected layers. The activation function of the last fully connected layer is softmax and previously it was RELU. The final model layer information is provided in Table 4. Since the adopted approach was transfer learning, our training accuracy started from 45%.

The training process using Mobilenet V2 took 12 hours with the computational time per epoch being 1400 seconds. We used 14804 images for training. The epochs taken are 25 with a batch size of 32 and learning rate of 0.0001. After training the mode, an accuracy of 95.16% was achieved. The whole training process to build the model took 9.72 hours. The training accuracy and loss versus the number of epochs is provided in Figure 8.

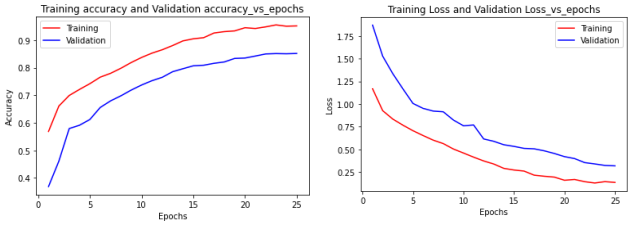


**FIGURE 8.** Plot of training and validation accuracy/loss vs epochs.

Figure 9 provides the individual test results of emotion recognition from webcam video. With FER 2013 dataset, we achieved the maximum test accuracy of 73.4%. As our main aim was to build an engagement recognition model, we did not focus on increasing the accuracy of detecting emotions.

Emotions influence the attention level of students. Hence, it is possible to establish a link between different types of primary emotions and academic affective states. Engaging with the literature [52], [53], and in consultation with the experienced faculty members from academic departments and student counsellors from Middle East College, we established a link between the various categories of student emotions during classroom sessions with their engagement level. This was done by assigning a certain weight to the detected emotions as per the attention level of each student at that point of time. Table 5 provides the weights assigned for each facial emotions of the class towards academic affective state.

### 6) ENGAGEMENT RECOGNITION

The engagement recognition block combines the weights of different modalities (eye status, gaze direction with head pose and facial emotions). The recorded video lesson of the online class is fed to face detection and cropping blocks as discussed in facial emotion recognition session. However, here it is done for multiple faces, for the faces of leaners $x \geq 1$.The detection of facial emotion, eye status and detection of gaze for every $x^{th}$ face is obtained along with corresponding weights. The returned weights of faces from the sequential frames are aggregated for the duration of 60 seconds using STAT function. Engagement status classification is performed based on the aggregate score obtained for the "engagement indicator". which is the sum of the weighted average of each of modalities gaze direction with head pose estimation, eye aspect ratio and facial emotion from the image traits over 60 seconds calculated.

$$EI = \sum_{i=1}^{60} [g\,(i) + e\,(i) + f\,(i)] \tag{7}$$

where $EI$ is the Engagement Indicator, $g\,(i)$ is the weight of gaze direction with head pose for i$^{th}$ second, $e\,(i)$ is the average ESR for i$^{th}$ second and $f\,(i)$ is the weight of the facial emotion obtained for $i^{th}$ second.

Figure 10 shows the continuous scale for our proposed $EI$ calculation algorithm for the classification of engagement status. The scale for the engagement indicator is developed through collaboration with literature [25] and discussions involving highly experienced teachers, each possessing over 15 years of expertise, as well as staff members from the Student Success Center at Middle East College, which offers support in student counselling.

Table 6 provides the range of Engagement Indicator (EI) score for deciding the engagement status. Figure 11 shows the engagement status predicted as per the EI score ranges obtained in Table 6.

**TABLE 6.** Engagement classification based on EI value.

| Engagement status | EI |
|---|---|
| Highly Engaged | $EI \geq 4.5$ |
| Confused | $4 \leq EI < 4.5$ |
| Boredom | $2.5 \leq EI < 4$ |
| Sleepy | $EI < 2.5$ |

### 7) EVALUATION METHOD

We used *Accuracy, TP, TN, FP, FN, Precision, Recall* and *F score* for evaluating the performance of our facial emotion recognition model.

True Positive (TP) : Model identifies correctly the positive class

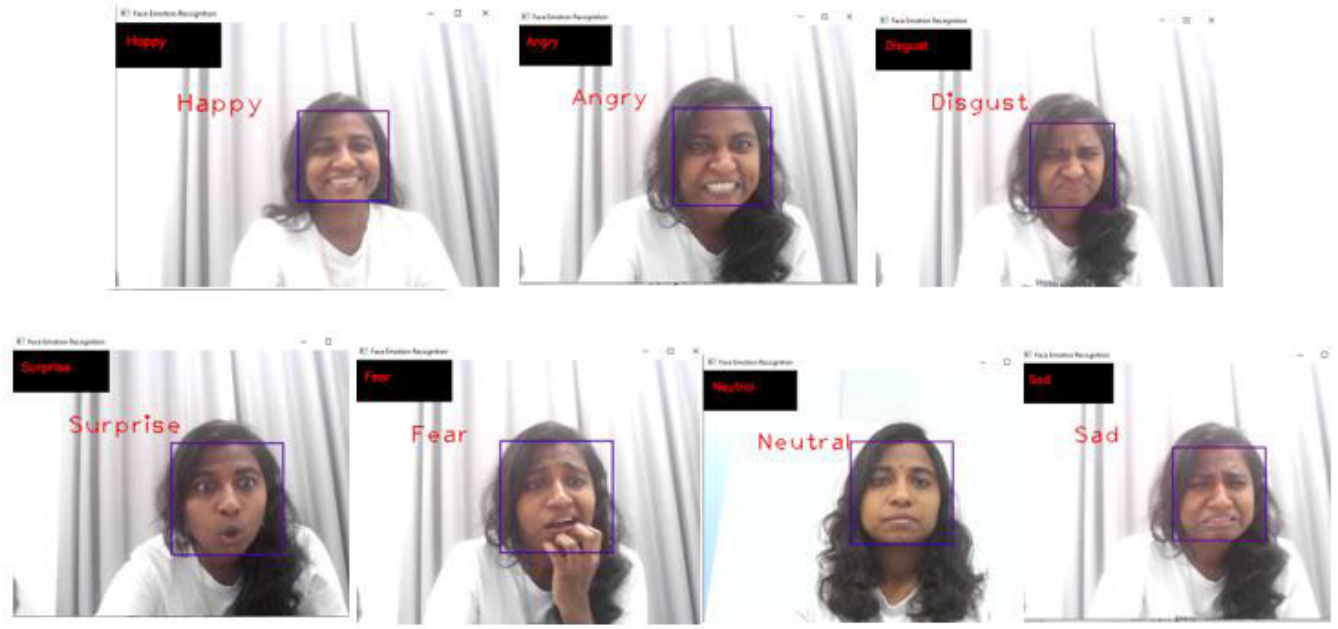


**FIGURE 9.** Result on webcam video: With all seven emotions predicted.

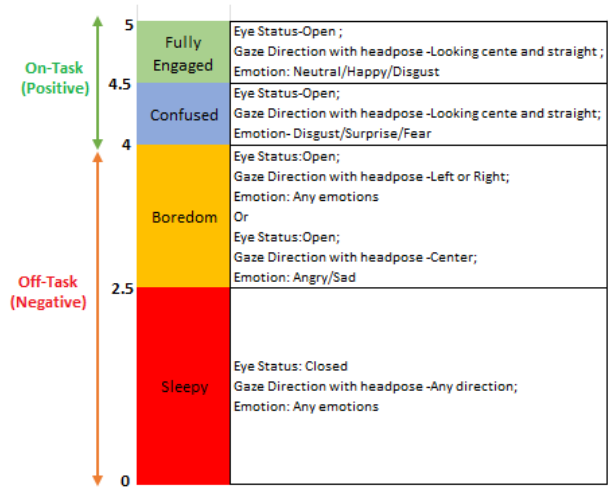


**FIGURE 10.** Continuous scale manual rating of the proposed EI calculation algorithm.

True Negative(TN): Model identifies correctly the negative class

False Positive (FP): Model identifies incorrectly the positive class

False Negative (FN): Model identifies incorrectly the negative class

Accuracy:

Accuracy is defined as the ratio of correct predictions of the output to the total number of outputs. It works well if FP and FN have similar costs.

$$Accuracy\,(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \qquad (8)$$

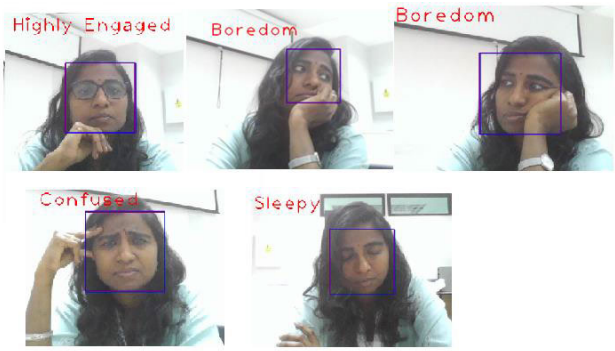


**FIGURE 11.** Result on webcam video: The engagement status predicted as per EI score.

Precision:

It is the ratio of the correct positive labelled predictions to total positive prediction. Precision is a good measure when it has more importance or value for false positive cases. A good precision value is recommended.

$$Precision = \frac{TP}{TP + FP} \times 100 \qquad (9)$$

Recall

It is the ratio of the correct positive labelled predictions to the sum of correct positive and incorrect negative predictions. Hence ''Recall'', will help to select the best model when there is a high cost associated with false negatives.

$$Recall = \frac{TP}{TP + FN} \times 100 \qquad (10)$$

Specificity

It is the ratio of the correct prediction of negative category to the sum of correct negative and incorrect positive

predictions.

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{11}$$

F-score

When it is required the balance between the Precision and Recall or in the case of uneven class of distribution (In our case disgust emotion is very less compared to others)

$$F = \frac{(2 \times Precision \times Recall)}{Precision + Recall} \tag{12}$$

The performance of the Engagement recognition model is evaluated by analyzing Quiz scores and their respective engagement status obtained for each student based on the one-hour session conducted for three consecutive days.

## IV. THE STUDY

The study was conducted at the Department of Computing and Electronic Engineering at Middle East College, Oman with 10 student participants undertaking the fourth semester of Bachelor of Engineering with Electronics and Telecommunication specialization. The main purpose of the study was to test the quality of the engagement detection model in predicting learner's engagement. The participants were attending the class of ''Communication Systems and Networks'' subject, with the topic of ''Amplitude Modulation''. The topic was delivered online on MS Teams and problems were solved with the help of Wacom tablet. The students are aged between 20 and 30 years. At the beginning, to evaluate their knowledge of the prerequisite topic, a formative quiz was conducted for each student before the commencement of the class sessions.

### A. SET UP

The proposed system was implemented using our laptop with specifications, Intel® Core™ i7-10510U CPU@1.8GHz 2.3GHz with RAM 24GB, windows 10,64-bit operating system.

### B. PROCEDURE

We conducted the experiment on 10 students, by offering one-hour sessions through MS Teams. Each session was recorded and after each the session, we downloaded the recording of the interactive session for the engagement analysis. The video was processed at 20 frames per second. All the 10 students' images in .jpg extension were saved in a folder. This was done for the face recognition purpose of the student in the class. The engagement indicator (EI) score was calculated from the aggregate weight values returned for the modules-gaze direction with head pose, eye status and facial emotions. Based on the obtained EI value, the engagement recognition was performed and was recorded in the csv file at the end of every 60 seconds. The CSV file created includes ''name'', ''date'','' time'' and ''Engagement Status'' for recording the information. If the EI value is less than 2.5, eyes are closed and logged as ''Sleepy'', as the highest weight assigned compared to all other modalities, is for eye open/ closed condition. Gaze direction with head pose alignment is assigned next highest weight, as looking away from the screen could mostly indicate 'Boredom' attitude.

At the end of the class, we conducted a multiple-choice quiz based on the discussed topic in class through Moodle to verify the correctness of our model. For the proposed model, the detailed analytics is made available to the teacher in spread sheet.

### C. RESULTS AND DISCUSSION

As discussed in the previous section, we applied the engagement classification to a cohort of 10 undergraduate students. The facial emotion recognition model was trained by FER 2013 dataset, with mobilenet V2 CNN architecture. Figure 12 shows the confusion matrix obtained for our facial emotion recognition model for FER 2013 dataset. Table 7 provides the performance metric of the Emotion recognition model. We achieved 73.4% accuracy for the model.

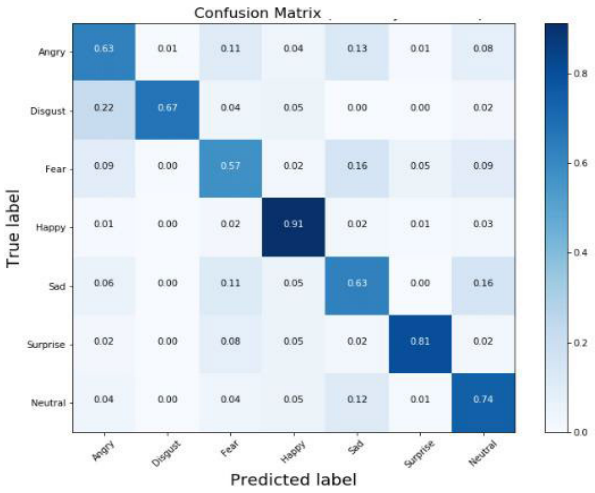


**FIGURE 12.** Confusion matrix for individual emotion recognition using FER 2013.

**TABLE 7.** Facial emotion performance evaluations.

| | Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Accuracy | 0.88 | 0.95 | 0.87 | 0.95 | 0.88 | 0.96 | 0.91 |
| Misclassification rate | 0.12 | 0.05 | 0.13 | 0.05 | 0.12 | 0.04 | 0.09 |
| Precision | 0.62 | 0.67 | 0.58 | 0.91 | 0.62 | 0.81 | 0.74 |
| Recall | 0.59 | 0.99 | 0.59 | 0.78 | 0.58 | 0.91 | 0.65 |
| Specificity | 0.94 | 0.95 | 0.92 | 0.98 | 0.94 | 0.97 | 0.96 |
| F1 | 0.61 | 0.80 | 0.58 | 0.84 | 0.60 | 0.86 | 0.69 |

The Engagement recognition is performed based on the proposed equation (7), by calculating the Engagement Indicator score from the output weight values of different modalities from the image traits.

The figure 13 shows the screenshot of detected engagement status of the available students in the online class. The

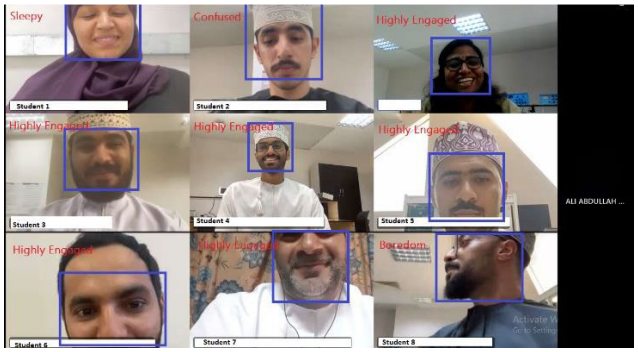


**FIGURE 13.** Detected engagement status in the online class.

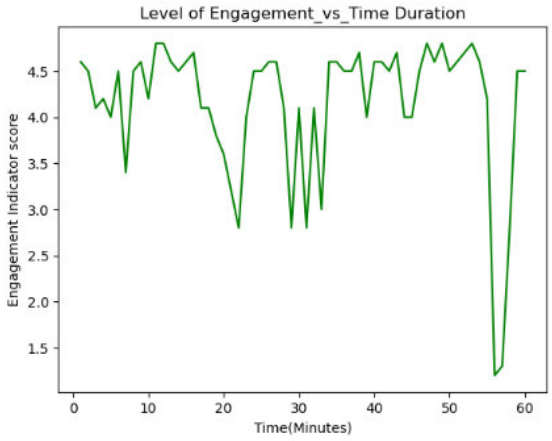


**FIGURE 14.** Level of engagement of student 1 for Day 1 session.

**TABLE 8.** Average engagement percentage for each student for the 1 hour on day 1.

| Student Name | Engagement Status | | | |
|---|---|---|---|---|
| | Highly Engaged | Confused | Boredom | Sleepy |
| Student 1 | 60% | 21.67% | 15% | 3.33% |
| Student 2 | 35% | 28.33% | 28.33% | 8.33% |
| Student 3 | 78.33% | 11.67% | 10% | 0% |
| Student 4 | 76.67% | 13.33% | 8.33% | 1.67% |
| Student 5 | 81.67% | 5% | 13.33% | 0% |
| Student 6 | 86.66% | 3.33% | 10% | 0% |
| Student 7 | 58.33% | 25% | 3.33% | 13.33% |
| Student 8 | 25% | 10% | 40% | 25% |
| Student 9 | 75% | 10% | 11.67% | 3.33% |
| Student 10 | 68.33% | 18.33% | 13.33% | 0% |

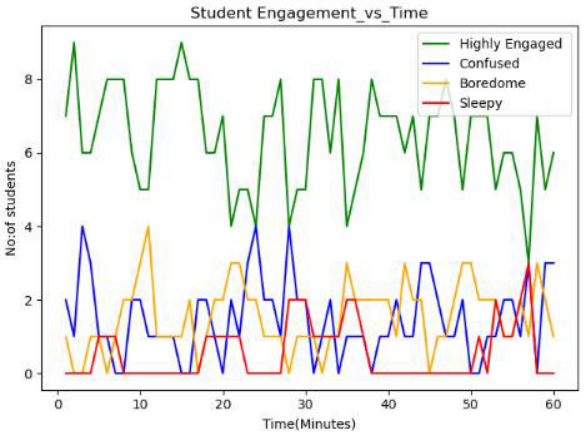


**FIGURE 15.** Time series details of engagement statistics.

time series details of engagement recognition statistics of our experiment are provided in figures 14 and 15. Figure 14, plots the engagement indicator value obtained for student 1 for the 60-minute session conducted on Day 1. Figure 15 provides the analysis statistics of the number of students and their predicted involvement towards 60 minutes session for day 1. Table 8 and figure 16 provide the details of the classified engagement status for each student in a 60-minute session.

For each student, we compared the percentage of 'highly engaged' category and their respective quiz score for validation. The details for the three days session and their percentage for highly engagement category and quiz score obtained for each day's session is given in Table 9. A visual representation of the correlation between students' quiz score performance and percentage of their engagement over three consecutive day sessions are depicted in figure 17,18 and 19.

At the end of the session, the statistics regarding the engagement details are available for the teacher in spreadsheet. Table 10 provides the name of each student, with the details of time and the respective engagement status from the downloaded csv file. This facilitates the analysis of the engagement status of each student and teachers can take the necessary action according to the results.

In summary the integration of facial emotions, gaze direction, head pose alignment and eye status in our engagement recognition model represents a multifaceted approach for capturing the degree of the student's attention level. The inclusion of facial emotions allows the model to understand the emotional aspects of engagement, facilitating the perception of the student's affective states. Meanwhile, gaze direction with head-pose alignment and eye status offers supporting cues, enhancing the depth of understanding by considering where students are directing their attention. Combining these features increases the classification accuracy of the model and enables a more comprehensive evaluation of engagement, considering both emotional and cognitive dimensions.

The proposed multimodal engagement recognition model mitigates the challenges associated with the recognition of the individual's affective states. The images of faces are affected by lighting conditions; moreover, the prediction of gaze might vary when students are wearing glasses. The gaze direction with head pose alignment can compensate for these potential inaccuracies, promoting a more reliable engagement recognition. The validation of our model involved the implementation of a quiz at the end of each session, providing a robust method to assess the accuracy of the model in

**TABLE 9.** The percentage of "Highly Engaged" category and the respective Quiz score for three days.

| Student | Day 1 | | *Day 2* | | Day 3 | |
|---|---|---|---|---|---|---|
| | % of Highly Engaged category | Quiz Score Percentage | % of Highly Engaged category | Quiz Score Percentage | % of Highly Engaged category | Quiz Score Percentage |
| Student 1 | 60% | 80% | 80% | 90% | 55% | 60% |
| Student 2 | 35% | 30% | 55% | 40% | 60% | 70% |
| Student 3 | 78.33% | 80% | 65% | 80% | 75% | 60% |
| Student 4 | 76.67% | 90% | 76.67% | 80% | 73.33% | 60% |
| Student 5 | 81.67% | 90% | 85% | 90% | 78.33% | 100% |
| Student 6 | 86.67% | 100% | 73.33% | 80% | 81.67% | 100% |
| Student 7 | 58.33% | 70% | 70% | 80% | 68.33% | 80% |
| Student 8 | 25% | 30% | 40% | 60% | 30% | 40% |
| Student 9 | 75% | 70% | 68.33% | 70% | 65% | 80% |
| Student 10 | 68.33% | 60% | 80% | 70% | 75% | 80% |

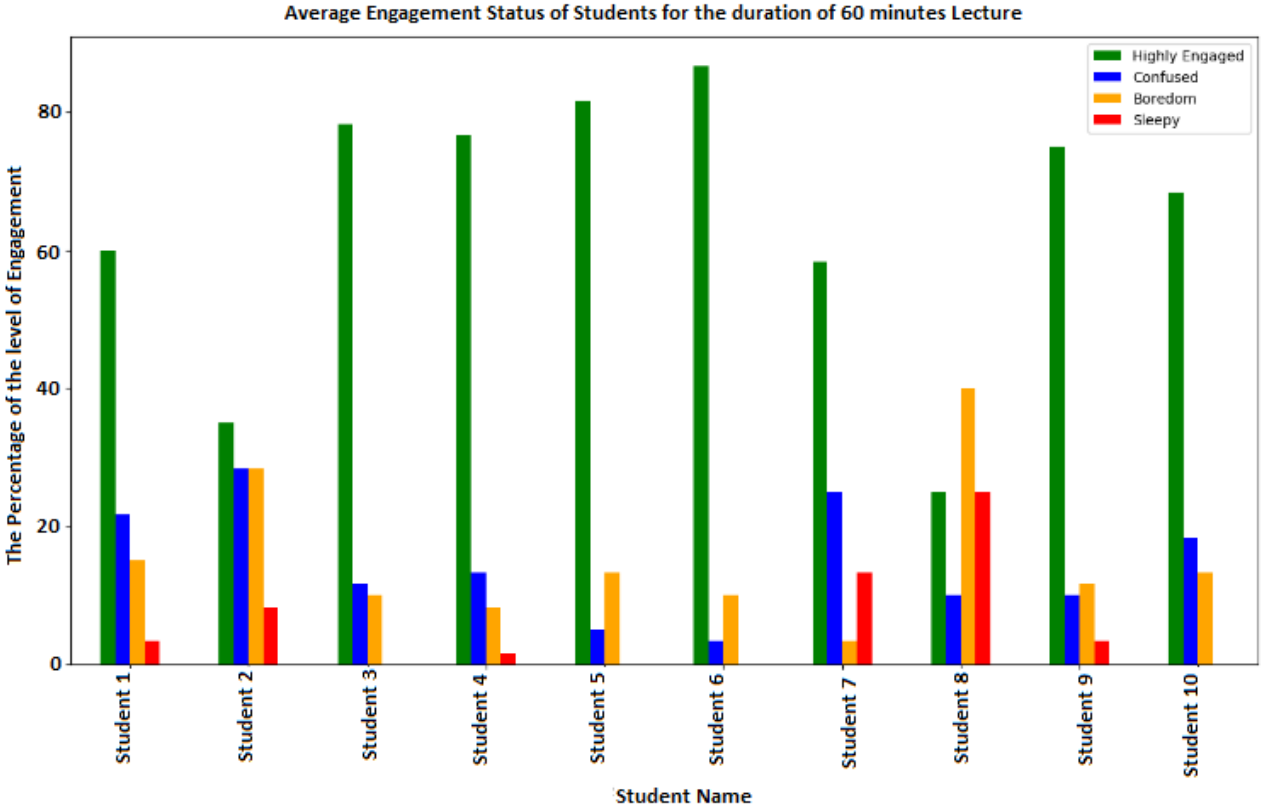


**FIGURE 16.** The average engagement statics of student for 60-minute session on Day 1.

predicting the student's engagement. The findings affirm the effectiveness of the model and imply the positive relationship between engagement and academic outcomes.

In contrast to the existing engagement recognition models that primarily focus on predicting the engagement status, our work introduces a significant improvement by not only predicting engagement, but also providing the analytics in a spreadsheet to teachers. This innovation adds a practical dimension to our model's utility by enabling teachers to get deeper understandings into the dynamics of student engagement, promoting a more updated and actionable approach to instructional strategies.

In the field of AI in Education, our research on engagement recognition of students places a significant concern on ethical considerations. We have collected the signed consent from all participants for conducting the research. By obtaining consent, we prioritize the ethical principles of privacy, acknowledging the sensitive nature of facial emotion

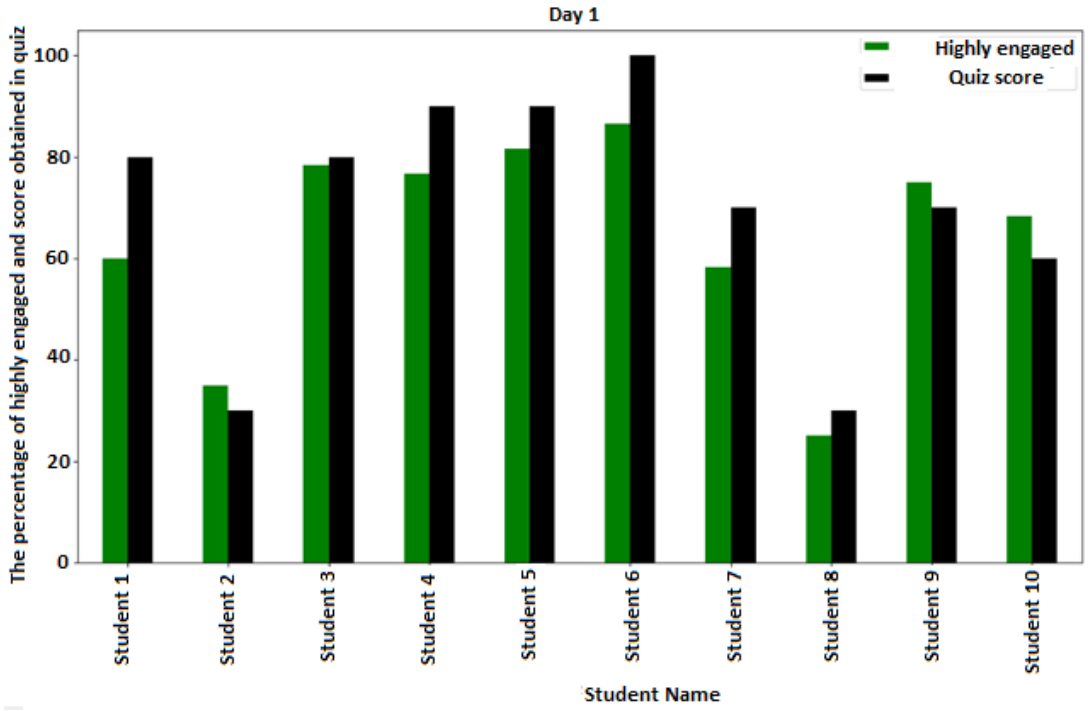


**FIGURE 17.** The percentage of participants classified under "Highly Engaged" category and their corresponding Quiz score for Day 1 across the three days sessions.

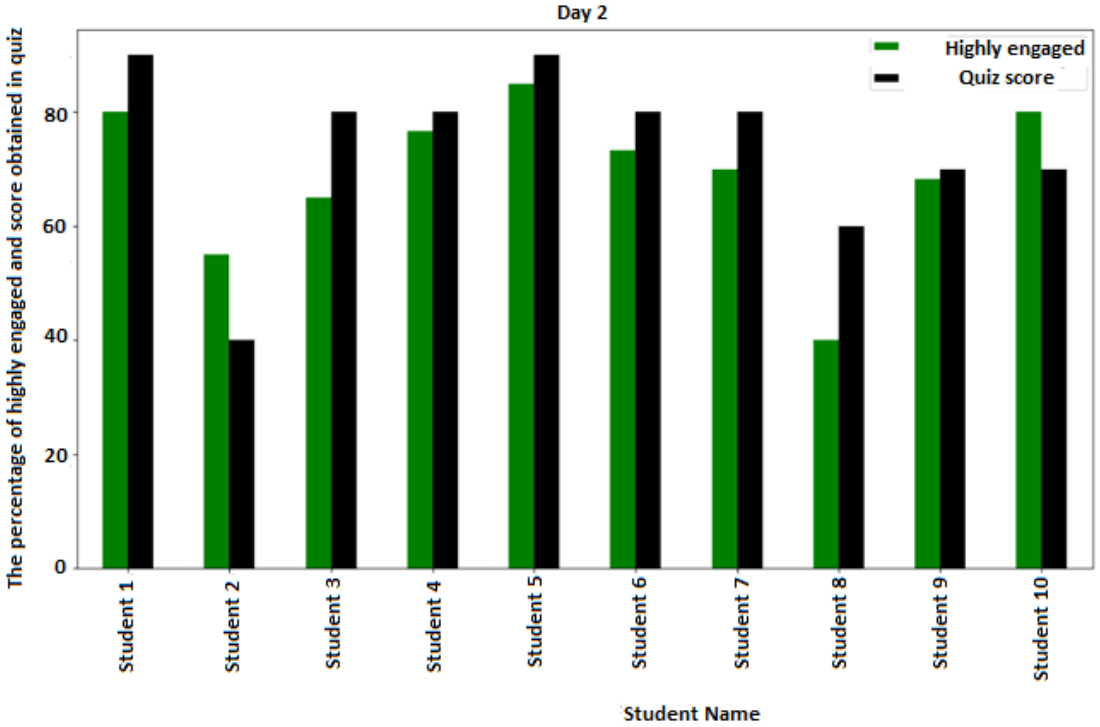


**FIGURE 18.** The percentage of participants classified under "Highly Engaged" category and their corresponding Quiz score for Day 2 across the three days sessions.

analysis. Our commitment to ethical considerations aligns with legal and moral standards and contributes to promoting trust and transparency in the applications of AI technologies for educational settings.

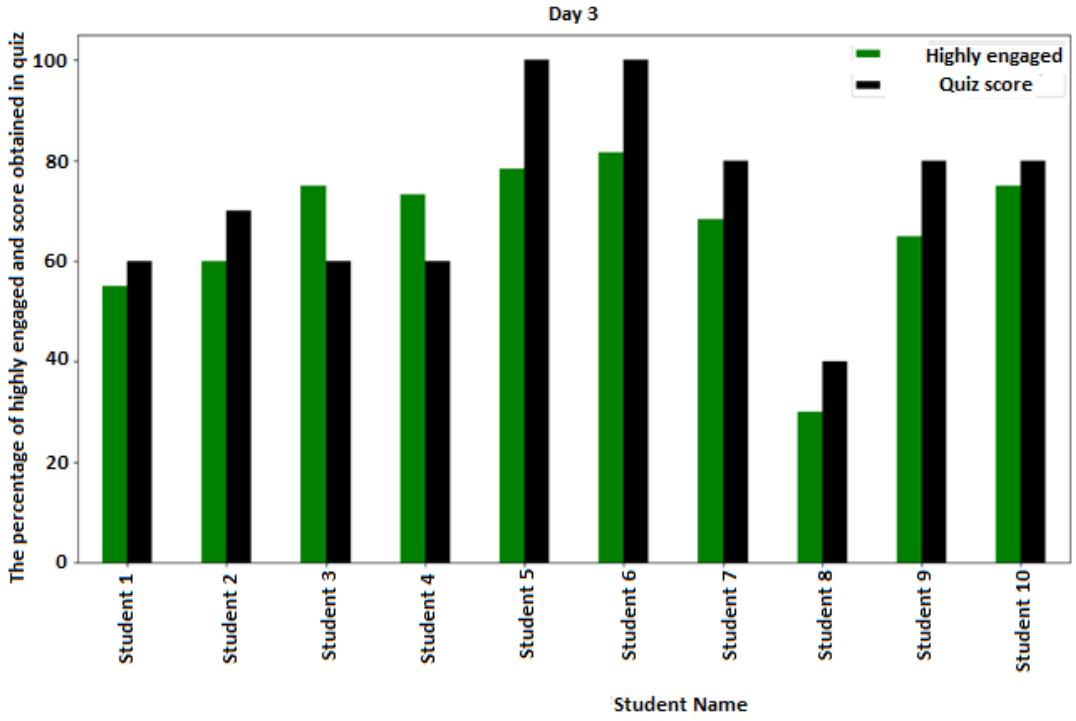


**FIGURE 19.** The percentage of participants classified under "Highly Engaged" category and their corresponding Quiz score for Day 3 across the three days sessions.

**TABLE 10.** Data from the downloaded CSV file.

| Name | Date | Time | Engagement Status |
|---|---|---|---|
| Student 1 | 16/05/23 | 14:10:12 | Highly Engaged |
| Student 1 | 16/05/23 | 14:11:12 | Highly Engaged |
| Student 1 | 16/05/23 | 14:12:12 | Confused |
| Student 1 | 16/05/23 | 14:13:12 | Highly Engaged |
| Student 1 | 16/05/23 | 14:14:12 | Confused |
| Student 1 | 16/05/23 | 14:15:12 | Highly Engaged |
| Student 1 | 16/05/23 | 14:16:12 | Bored |
| Student 1 | 16/05/23 | 14:17:12 | Highly Engaged |
| Student 1 | 16/05/23 | 14:18:12 | Highly Engaged |
| Student 1 | 16/05/23 | 14:19:12 | Confused |
| Student 1 | 16/05/23 | 14:20:12 | Highly Engaged |
| Student 1 | 16/05/23 | 14:21:12 | Highly Engaged |
| Student 1 | 16/05/23 | 14:22:12 | Highly Engaged |

## V. CONCLUSION

The increased research focus on the automatic monitoring of students' attention during lecture sessions is the key element in the broader movement to adaptive and personalized learning. A review of the literature reveals that automatic engagement recognition methods are mostly based on computer vision or sensors. This study focused on engagement detection with computer vision-based approach. The main contributions of our work include designing a model that can correctly predict the engagement status of students based on the analysis of their gaze direction, eye status and emotions. We propose an engagement indicator (EI) algorithm calculated from image traits. Based on the continuous scale of the engagement indicator, the engagement status is labelled into four categories: Highly Engaged, Confused, Boredom and Sleepy. The proposed model was tested successfully for multiple students from the recorded lecture sessions. The details of engagement statistics are available in the spread sheet for the teacher to plan follow up activities. This work is an added feature to the ongoing development of AI-driven solutions in education and to provide better educational insights for both learners and teachers.

The applications of engagement recognition are not limited to the education field, and this research has several economic and commercial impacts. Engagement recognition can be integrated into the productivity tools used in various professions and industries; they can be used in platforms designed for professional training and upskilling of employees. It has potential applications in conferences and other professional events to gauge participants' engagement and to monitor and enhance the virtual collaboration and engagement of remote work teams.

The limitations of our work are that for the experiment we considered only undergraduate students (students of 4th semester) as participants. The experiment was conducted in the Middle East region, and due to cultural reasons, many students didn't provide consent for recording the session with their cameras switched on; hence our cohort size was limited to only 10. Hence, generalizing these results might not be possible. Future work in this area can consider students at different levels of education, starting from primary school to tertiary education and beyond.

During the facial emotion recognition process, discrepancy was observed between the train and test accuracy (95.6% and 73.4%, respectively), indicating a potential issue of overfitting in our model. Due to this, the generalizability of the model might be limited to the diverse facial expressions. The overfitting issues can be addressed by regularization, data augmentation, normalization layers, and dropout. Future work could explore these strategies to enhance the model's robustness and applicability.

Because of the lack of a dedicated dataset for engagement recognition, in our experiment the engagement status was identified from facial expression associated with a limited set of emotions: happiness, neutral, surprise, disgust, anger, fear and sadness. We are in the process of constructing a dedicated dataset of affective state of emotions of students while they are engaged in learning. Further work will be to use customized datasets for engagement detection dedicated to online and offline classroom set ups. Another important aspect is that facial emotions may not always reflect the person's mental state. It is better to include the analysis of mental state based on physiological and neurological signals. Therefore, a fusion of the modalities considered from image traits along with analysis of physiological and neurological signals will improve the accuracy of the proposed engagement recognition model. However, the proposed system displayed reliable results in detecting human faces, gaze direction, emotion recognition and the status of engagement of the participants. This research has the potential to positively impact education and society by fostering personalized learning experiences and contributing to the development of innovative educational technologies.

## ACKNOWLEDGMENT

The authors would like to thank Middle East College, faculties, and students for providing the necessary support for the implementation stage, and also would like to thank Dr. Priya Mathew, Head of Center for Academic Writing, Middle East College for reviewing the article.

## REFERENCES

[1] C. M. Amerstorfer and C. F. von Münster-Kistner, "Student perceptions of academic engagement and student-teacher relationships in problem-based learning," *Frontiers Psychol.*, vol. 12, p. 4978, Oct. 2021, doi: 10.3389/fpsyg.2021.713057.

[2] S. Fabriz, J. Mendzheritskaya, and S. Stehle, "Impact of synchronous and asynchronous settings of online teaching and learning in higher education on students' learning experience during COVID-19," *Frontiers Psychol.*, vol. 12, p. 4544, Oct. 2021, doi: 10.3389/fpsyg.2021.733554.

[3] K. Buntins, M. Kerres, and A. Heinemann, "A scoping review of research instruments for measuring student engagement: In need for convergence," *Int. J. Educ. Res. Open*, vol. 2, Jan. 2021, Art. no. 100099, doi: 10.1016/j.ijedro.2021.100099.

[4] F. H. Veiga, J. Reeve, K. Wentzel, and V. Robu, "Assessing students' engagement: A review of instruments with psychometric qualities," Tech. Rep., 2014.

[5] B. J. Mandernach, "All rights reserved," Tech. Rep., 2015.

[6] J. A. Fredricks and W. McColskey, "The measurement of student engagement: A comparative analysis of various methods and student self-report instruments," in *Handbook of Research on Student Engagement*. Berlin, Germany: Springer, 2012, pp. 763–782, doi: 10.1007/978-1-4614-2018-7_37.

[7] F. Ouyang and P. Jiao, "Artificial intelligence in education: The three paradigms," *Comput. Educ., Artif. Intell.*, vol. 2, Jan. 2021, Art. no. 100020, doi: 10.1016/j.caeai.2021.100020.

[8] X. Chen, H. Xie, D. Zou, and G.-J. Hwang, "Application and theory gaps during the rise of artificial intelligence in education," *Comput. Educ., Artif. Intell.*, vol. 1, Jan. 2020, Art. no. 100002, doi: 10.1016/j.caeai.2020.100002.

[9] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 524–543, Apr. 2021, doi: 10.1109/TAFFC.2018.2890471.

[10] S. Kumar, "Deep learning based affective computing," *J. Enterprise Inf. Manag.*, vol. 34, no. 5, pp. 1551–1575, Nov. 2021, doi: 10.1108/jeim-12-2020-0536.

[11] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 15, no. 3, pp. 1–32, 2019, doi: 10.1145/3363560.

[12] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, pp. 24–35, Nov. 2018, doi: 10.1016/j.dss.2018.09.002.

[13] A. Sukumaran and A. Manoharan, "A survey on automatic engagement recognition methods: Online and traditional classroom," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 30, no. 2, pp. 1178–1191, May 2023, doi: 10.11591/ijeecs.v30.i2.pp1178-1191.

[14] M. A. Rashidan, S. N. Sidek, H. M. Yusof, M. Khalid, A. A. A. Dzulkarnain, A. S. Ghazali, S. A. M. Zabidi, and F. A. A. Sidique, "Technology-assisted emotion recognition for autism spectrum disorder (ASD) children: A systematic literature review," *IEEE Access*, vol. 9, pp. 33638–33653, 2021, doi: 10.1109/ACCESS.2021.3060753.

[15] H. T. Le and L. A. Vea, "A customer emotion recognition through facial expression using Kinect sensors v1 and v2: A comparative analysis," in *Proc. 10th Int. Conf. Ubiquitous Inf. Manage. Commun.*, New York, NY, USA, Jan. 2016, pp. 1–7.

[16] A. Altameem, A. Kumar, R. C. Poonia, S. Kumar, and A. K. J. Saudagar, "Early identification and detection of driver drowsiness by hybrid machine learning," *IEEE Access*, vol. 9, pp. 162805–162819, 2021.

[17] S. Kwon, J. Ahn, H. Choi, J. Jeon, D. Kim, H. Kim, and S. Kang, "Analytical framework for facial expression on game experience test," *IEEE Access*, vol. 10, pp. 104486–104497, 2022, doi: 10.1109/ACCESS.2022.3210712.

[18] M. A. A. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: A review," *Smart Learn. Environments*, vol. 6, no. 1, pp. 1–20, Dec. 2019, doi: 10.1186/s40561-018-0080-z.

[19] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan. 2014, doi: 10.1109/TAFFC.2014.2316163.

[20] Y. Hu, Z. Jiang, and K. Zhu, "An optimized CNN model for engagement recognition in an e-learning environment," *Appl. Sci.*, vol. 12, no. 16, p. 8007, Aug. 2022, doi: 10.3390/app12168007.

[21] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 11365–11394, Mar. 2023, doi: 10.1007/s11042-022-13558-9.

[22] A. Psaltis, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal student engagement recognition in prosocial games," *IEEE Trans. Games*, vol. 10, no. 3, pp. 292–303, Sep. 2018, doi: 10.1109/TCIAIG.2017.2743341.

[23] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2132–2143, Oct. 2022, doi: 10.1109/TAFFC.2022.3188390.

[24] P. Buono, B. De Carolis, F. D'Errico, N. Macchiarulo, and G. Palestra, "Assessing student engagement from facial behavior in on-line learning," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 12859–12877, Apr. 2023, doi: 10.1007/s11042-022-14048-8.

[25] O. Sumer, P. Goldberg, S. Dmello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1012–1027, Apr. 2023, doi: 10.1109/TAFFC.2021.3127692.

[26] C. J. Hellín, F. Calles-Esteban, A. Valledor, J. Gómez, S. Otón-Tortosa, and A. Tayebi, "Enhancing student motivation and engagement through a gamified learning environment," *Sustainability*, vol. 15, no. 19, p. 14119, Sep. 2023, doi: 10.3390/su151914119.

[27] V. Skaramagkas, E. Ktistakis, D. Manousos, N. S. Tachos, E. Kazantzaki, E. E. Tripoliti, D. I. Fotiadis, and M. Tsiknakis, "A machine learning approach to predict emotional arousal and valence from gaze extracted features," in *Proc. 21st IEEE Int. Conf. BioInformatics BioEngineering*, Oct. 2021, pp. 1–5, doi: 10.1109/BIBE52308.2021.9635346.

[28] D. Cazzato, M. Leo, C. Distante, and H. Voos, "When I look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking," *Sensors*, vol. 20, no. 13, pp. 1–42, Jul. 2020, doi: 10.3390/s20133739.

[29] S. Gu, F. Wang, N. P. Patel, J. A. Bourgeois, and J. H. Huang, "A model for basic emotions using observations of behavior in drosophila," *Frontiers Psychol.*, vol. 10, p. 781, Apr. 2019, doi: 10.3389/fpsyg.2019.00781.

[30] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," 2016, *arXiv:1609.01885*.

[31] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2018, pp. 1–8, doi: 10.1109/DICTA.2018.8615851.

[32] C. Bian, Y. Zhang, F. Yang, W. Bi, and W. Lu, "Spontaneous facial expression database for academic emotion inference in online learning," *IET Computer Vision*, vol. 13, no. 3, pp. 329–337, 2019, doi: 10.1049/iet-cvi.2018.5281.

[33] A. Sukumaran, "A brief review of conventional and deep learning approaches in facial emotion recognition," Tech. Rep., 2019.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[35] Z.-Y. Huang, C.-C. Chiang, J.-H. Chen, Y.-C. Chen, H.-L. Chung, Y.-P. Cai, and H.-C. Hsu, "A study on computer vision for facial emotion recognition," *Sci. Rep.*, vol. 13, no. 1, p. 8425, May 2023, doi: 10.1038/s41598-023-35446-4.

[36] N. Abbassi, R. Helaly, M. A. Hajjaji, and A. Mtibaa, "A deep learning facial emotion classification system: A VGGNet-19 based approach," in *Proc. 20th Int. Conf. Sci. Techn. Autom. Control Comput. Eng. (STA)*, Dec. 2020, pp. 271–276, doi: 10.1109/STA50679.2020.9329355.

[37] T. Ramu and A. Muthukumar, "A GoogleNet architecture based facial emotions recognition using EEG data for future applications," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2022, pp. 1–7, doi: 10.1109/ICCCI54379.2022.9740864.

[38] M. Aslan, "CNN based efficient approach for emotion recognition," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7335–7346, Oct. 2022, doi: 10.1016/j.jksuci.2021.08.021.

[39] Y. Huang and D. Bo, "Emotion classification and achievement of students in distance learning based on the knowledge state model," *Sustainability*, vol. 15, no. 3, p. 2367, 2023, doi: 10.3390/su15032367.

[40] B. K. Durga and V. Rajesh, "A ResNet deep learning based facial recognition design for future multimedia applications," *Comput. Electr. Eng.*, vol. 104, Dec. 2022, Art. no. 108384, doi: 10.1016/j.compeleceng.2022.108384.

[41] P. Utami, R. Hartanto, and I. Soesanti, "The EfficientNet performance for facial expressions recognition," in *Proc. 5th Int. Seminar Res. Inf. Technol. Intell. Syst.*, 2022, pp. 756–762, doi: 10.1109/ISRITI56927.2022.10053007.

[42] S. B. Punuri, S. K. Kuanar, M. Kolhar, T. K. Mishra, A. Alameen, H. Mohapatra, and S. R. Mishra, "Efficient net-XGBoost: An implementation for facial emotion recognition using transfer learning," *Mathematics*, vol. 11, no. 3, p. 776, Feb. 2023, doi: 10.3390/math11030776.

[43] G. Singh, I. Gupta, J. Singh, and N. Kaur, "Face recognition using open source computer vision library (OpenCV) with Python," in *Proc. 10th Int. Conf. Rel., INFOCOM Technol. Optim. (Trends Future Directions) (ICRITO)*, Oct. 2022, doi: 10.1109/icrito56286.2022.9964836.

[44] R. Andrie Asmara, M. Ridwan, and G. Budiprasetyo, "Haar cascade and convolutional neural network face detection in client-side for cloud computing face recognition," in *Proc. Int. Conf. Electr. Inf. Technol. (IEIT)*, Sep. 2021, pp. 1–5, doi: 10.1109/IEIT53149.2021.9587388.

[45] A. B. Shetty and J. Rebeiro, "Facial recognition using Haar cascade and LBP classifiers," *Global Transitions Proc.*, vol. 2, no. 2, pp. 330–335, Nov. 2021, doi: 10.1016/j.gltp.2021.08.044.

[46] L. Cuimei, Q. Zhiliang, J. Nan, and W. Jianhua, "Human face detection algorithm via Haar cascade classifier combined with three additional classifiers," in *Proc. 13th IEEE Int. Conf. Electron. Meas. Instrum. (ICEMI)*, Oct. 2017, doi: 10.1109/ICEMI.2017.8265863.

[47] P. Viola, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Society Conf. Comput. Vis. Pattern Recognit.*, Jun. 2001.

[48] D. E. King, "DLIB-ML: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[49] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 815–823.

[50] G. Gallego and A. Yezzi, "A compact formula for the derivative of a 3-D rotation in exponential coordinates," *J. Math. Imag. Vis.*, vol. 51, no. 3, pp. 378–384, Mar. 2015, doi: 10.1007/s10851-014-0528-x.

[51] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[52] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers Psychol.*, vol. 8, p. 1454, Aug. 2017, doi: 10.3389/fpsyg.2017.01454.

[53] R. Bachler, P. Segovia-Lagos, and C. Porras, "The role of emotions in educational processes: The conceptions of teacher educators," *Frontiers Psychol.*, vol. 14, Jun. 2023, Art. no. 1145294, doi: 10.3389/fpsyg.2023.1145294.

**AJITHA SUKUMARAN** (Senior Member, IEEE) received the B.Tech. degree in electronics and communication from the Mar Athanasius College of Engineering, Kerala, India, in 2002, and the M.Tech. degree in digital system and communication engineering from the National Institute of Technology, Kerala, in 2005. She is currently pursuing the Ph.D. degree in digital image processing with the Vellore Institute of Technology, Tamil Nadu, India.

She is also as a Senior Lecturer with the Department of Computing and Electronic Engineering, Middle East College. Her research interests include artificial intelligence, nonlinear chaotic communication, cryptography, coding theory, and multiuser detection. For the past 19 years, she has taught various subjects within the field of electronics and communication engineering programs to students in highly reputed higher education institutions, both India and Oman. She is a fellow of the Higher Education Academy in recognition of attainment against the U.K. professional standards framework for teaching and learning support in higher education.

**ARUN MANOHARAN** (Senior Member, IEEE) received the Ph.D. degree in high performance computer networks from Anna University, Tamil Nadu, India, in 2011. He was a Postdoctoral Researcher with the Institute of Electronics and Informatics Engineering, University of Aveiro, Aveiro, Portugal. He is currently a Professor with the Department of Embedded Technology and an Assistant Director of International Relations with the Vellore Institute Technology, India. His current research interest includes the edge level security standards for LoRaWAN networks. His research contributions are towards the development of high-performance heterogeneous computing algorithms for computer intensive use cases, such as CFD, DNA sequence search, and space datasets. His research results in the publication of 40 plus research articles indexed in Scopus and Web of Science. He has 20 years of engineering academic and research experience. He secured funded projects from Indian Government Agencies, such as the Department of Science and Technology (DST), Indian Centre for Medical Research (ICMR), and AICTE. He worked for industrial consultancy projects to provide vision-based solutions to textile and manufacturing industries problems in collaboration with national instruments, NVIDIA, and Titan India.