Multimedia Tools and Applications (2023) 82:28589–28615
https://doi.org/10.1007/s11042-023-14392-3

A multimodal facial cues based engagement detection system in e-learning context using deep learning approach

Swadha Gupta[1] · **Parteek Kumar**[1] · **Rajkumar Tekchandani**[1]

Received: 13 August 2022 / Revised: 20 November 2022 /
Published online: 10 February 2023

Abstract

Due to the COVID-19 crisis, the education sector has been shifted to a virtual environment. Monitoring the engagement level and providing regular feedback during e-classes is one of the major concerns, as this facility lacks in the e-learning environment due to no physical observation of the teacher. According to present study, an engagement detection system to ensure that the students get immediate feedback during e-Learning. Our proposed engagement system analyses the student's behaviour throughout the e-Learning session. The proposed novel approach evaluates three modalities based on the student's behaviour, such as facial expression, eye blink count, and head movement, from the live video streams to predict student engagement in e-learning. The proposed system is implemented based on deep-learning approaches such as VGG-19 and ResNet-50 for facial emotion recognition and the facial landmark approach for eye-blinking and head movement detection. The results from different modalities (for which the algorithms are proposed) are combined to determine the EI (engagement index). Based on EI value, an engaged or disengaged state is predicted. The present study suggests that the proposed facial cues-based multimodal system accurately determines student engagement in real time. The experimental research achieved an accuracy of 92.58% and showed that the proposed engagement detection approach significantly outperforms the existing approaches.

Keywords Facial expressions · Engagement detection · Emotion detection · Deep learning · Eye-blinking · Head-movement · Real-time · Online learning

✉ Swadha Gupta
swadhagupta15@gmail.com

Parteek Kumar
parteek.bhatia@thapar.edu

Rajkumar Tekchandani
rtekchandani@thapar.edu

[1] Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, 147001, Punjab, India

1 Introduction

Digital learning is experiencing exponential growth post-COVID-19 crisis. E-Learning has become a necessity to ensure learning continuity during the COVID-19 outbreak. Directing student's attention and maintaining concentration during the process of listening is difficult, especially; when the teacher is not present physically to clear the doubts simultaneously. When students encounter unfamiliar terms, they lose focus and cannot catch up with the rest of the lecture. Due to this, problems like disengagement and lack of participation become common in e-learning [43, 73]. Most students leave the course mid-way due to no immediate support and feedback to their queries [45]. Student participation is highly influenced by the immediate support they receive from the teachers, which is lacking in an e-Learning environment [5]. Therefore, evaluating engagement is important to overcome the challenges faced by students in online education. Currently, most of the techniques used by teachers to evaluate students' learning performance are questionnaires and quizzes at the end of video lectures. These methods are inefficient as they do not guarantee that the student remains attentive throughout the video lecture and understands the whole concept. So, to evaluate engagement in the e-Learning environment, the human face plays an important role in communicating the individual's inner feelings and intentions. Observing one of the facial cues, i.e., the facial expression, is the direct way to recognise student focus during learning [72]. Analysis of Facial expression determines the interest and behaviour of the students during the learning process [10] same as what happens in the physical classroom environment. The teacher can quickly determine whether the student is engaged or disengaged in a physical classroom environment by observing their activities and reaction to a particular topic. So, besides facial expressions, there are other ways to observe the student classroom behaviour in e-learning. In addition to facial emotion recognition, eye-blinking and head movement can also be monitored by observing human facial features [20, 64, 67]. Eye-blinking and head movement detection are powerful methods to interpret an individual's intention while using a laptop/computer/tablet/mobile phone. Moving eye and head could be considered the student response to the delivered content in an e-Learning environment [4, 7, 57]. Although humans can naturally and intuitively accomplish it by observing, it is still a challenging problem for current computer systems. Thus, there is a need to have an interactive intelligent learning system that analyses facial expressions, eye-blinking, and head movements in a real-time learning environment.

The present work aims to detect student engagement based on multimodal information (face, eye, and head). A built-in web camera is proposed to grab real-time details of the face, eye, and head. We propose to use Deep CNN models for facial emotion recognition and 68-landmark points for eye-blinking patterns and head-movement monitoring. This information will help the teachers to make the e-Learning environment more interactive and adaptable according to the needs of the students.

The main contributions of this study are as follows:

- A real-time student engagement system is proposed to detect engagement levels on multimodal data, which combines appearance-based (facial emotion) and geometric-based (eye and head movement) information.
- The facial emotion algorithm is proposed based on the facial expression analysis. The facial expression is classified using the Deep CNN models such as ResNet-50 and VGG19 and implemented on the benchmarked datasets and collected datasets. The implementation of deep CNN results in different facial emotions. For real-time eye and head movement monitoring, facial landmarks are detected using Dlib.

- For eye-movement detection, eye-blinking and long eye-closure frequencies are analysed. EAR (Eye Aspect Ratio) is computed using the eye's landmarks information to get the count of eye-blinking and long eye closure in a specific period.
- For head-movement detection, a head position estimation algorithm is proposed. Based on the Euler angle calculation, the position of the head is estimated at any given time.
- Engagement evaluation algorithm is proposed to calculate the engagement index (EI) based on the output information from different modalities: facial emotion, eye-blinking, and head movement. The engagement state is predicted based on EI value to help teachers know how well students are engaged while studying online.
- At last, a comparative analysis with the existing systems is performed to show the effectiveness of the proposed engagement detection system.

The rest of the paper is organised as Section 2 gives a brief overview of the related work done for engagement detection in an e-Learning environment. Section 3 discusses the datasets used for the experiment. In Section 4, the proposed system is discussed. Section 5 presents the experimental results achieved after implementing the proposed approach. Section 6 illustrates the comparison of the proposed system with the state-of-the-art models. Section 7 concludes the paper with future work.

2 Related work

In recent years, the problem of determining student engagement in digital learning platforms has been gaining attention [44]. Engagement detection has become an important subject for research as it can help monitor the student's mental state without face-to-face supervision by the teacher. This section will discuss distinct approaches that help to detect student engagement. Authors in [30, 36, 48] utilised the moving pattern of the head and eyes to understand the online student's state. In this study, the system is proposed to detect emotions and classify learner's interests and participation while engaged in an offline classroom environment [58]. Authors in [41] proposed an approach based on facial expressions, eye gazes and mouse movement to measure the engagement of the user during a reading task. The authors used keystrokes and a web camera to measure attention. The geometric-based features were included rather than appearance-based features. The proposed approach achieved an accuracy of 75.5% using an SVM classifier. In [35], authors experimented with investigating the correlation between blinking eyes, facial surface temperature and students psychological behaviours during classroom learning. During the you-tube lesson video, the frequency of blinking eyes was measured, and the student's feedback response was considered in the e-Learning session. In [3], the authors proposed eye status detection using a deep learning methodology. The proposed approach is based on Convolutional Neural Network (CNN) architecture. Authors in [69] proposed a prediction framework to monitor the classroom based on camera technology. The student's face and body posture were captured using a Fly-eye lens camera placed in front of the classroom. The proposed approach used vision-based techniques to detect head motion using multiple high-resolution cameras to monitor engagement. In [11], an ensemble model is proposed based on face and body tracking for engagement detection. The student's engagement level was predicted using a cluster-based conventional model framework boosted with heuristic rules. Authors in [24] used student's facial expressions to propose the affective content analysis method. The test results showed

promising results with an accuracy of 87.65%. In [46], authors have analysed the relationship between eye movement and different emotional states. The results showed that the pupil diameter is larger during the negative emotion than the positive emotion. The eye blink frequency is higher in the negative than in positive emotion. In [39], authors have proposed a novel approach by fusing both the face and body gestures to detect student engagement automatically.

The previous research studies [8, 9, 61] showed the different approaches for student engagement measurement. According to previous studies, the different approaches are face features, eyeball tracking, eye gaze, body gestures, head motions, body surface temperature, mouse movements, and keystrokes. And also, the previous work was implemented on the students' stored videos and stored images; and for the in-person classroom setting. But the present work has been done for a real-time learning environment rather than an in-person classroom setting [15, 16, 56]. Thus, the present work proposes a real-time student Engagement detection system using deep learning techniques. The present work focuses on three modalities: eye-blinking, head movement and facial emotion recognition.

3 Dataset

The proposed system is evaluated using the WIDER face dataset for face detection; and CK+, FER-2013 and own created datasets for facial emotion recognition. These are publicly available and benchmarked datasets. The description of these datasets is given below.

3.1 Wider face

Wider Face dataset is used for training and testing the face images in the proposed system for the face detection step [71]. It is one of the largest datasets with 32,203 coloured images. The bounding boxes are used for labelling the images, and there are 393,703 labelled faces in the available images, as shown in Fig. 1(a). The summary of the Wider Face dataset is given in Table 1.

3.2 FER-2013

FER-2013 dataset is used to train and test the face images in the proposed system for facial emotion recognition. It's a publicly available dataset from the Wolfram data repository [18]. Images in the FER-2013 dataset are registered automatically in such a manner to allow equal space to the face area in all images. The sample images are shown in Fig. 1(b). The summary of the FER-2013 dataset is given in Table 1.

3.3 CK+ (extended Cohn-Kanade dataset)

CK+ (Extended Cohn-Kanade Dataset) dataset is also used to train and test the face images in the proposed system for facial emotion recognition purposes [23]. It's a publicly available laboratory-controlled dataset. Researchers have widely used it for facial expression classification purposes. CK+ consists of 593 sequences of images from 123 subjects. The sample images are shown in Fig. 1(c). The summary of CK+ dataset is given in Table 1.

justification=centering

Fig. 1 Sample images from (a) Wider Face (b) FER-2013 (c) CK+ (d) Own datasets

Table 1 Summary of various datasets

Face Detection DATASET					
Name	Year	Size (No of images)	Face Labels	Gray/Color	Types of Face Variations
Wider Face	2016	32,203	393,703	Coloured	Scale, Pose, Occlusion, Expression, Makeup, Illumination
Facial Expression Recognition DATASETS					
Name	Year	Size (No of images)	Image Size (in pixel)	Gray/Color	Types of Face Variations
FER-2013	2013	35,877	48 X 48	Grey Scale	Angry, Disgust, Frustration, Happy, Sad, Surprise, Neutral
CK+	2010	593	640 X 480	Mostly Gray / Coloured	Neutral, Sadness, Surprise, Happiness, Fear, Anger, Contempt, Disgust
OWN	2021	1800	48 X 48	Coloured	Happy, Neutral, Surprise, Sad, Angry, Fear

3.4 Own dataset

The new dataset has been created by collecting images from 45 subjects (undergraduate students). The collected images are labelled with six basic emotions for facial emotion classification purposes for the present study. The images are collected while keeping in mind the depth and RGB aspects of each subject. The sample images are shown in Fig. 1(d). The summary of the own dataset is given in Table 1.

4 Methodology

This paper proposes a real-time engagement detection system based on multimodal facial-cues data to predict student engagement, i.e., engaged or disengaged. Figure 2 shows the architecture of the proposed system. The real-time data of facial cues are extracted to predict the engagement state in the e-learning environment. The data are collected from the live video streaming of the student during the e-Learning sessions. For analysis, different algorithms are proposed. The engagement state is predicted by calculating the EI (Engagement index) value. EI is calculated based on data from three modalities: eye, head and facial expressions. Eye blinking and head movement modalities are extracted from dlib's facial landmark approach. Facial expression recognition modality is extracted based on deep learning models, i.e., VGG-19 and ResNet-50. These models are trained and tested on FER-2013, CK+ and Own datasets to classify emotions into different classes such as 'Happy', 'Neutral', 'Surprise', 'Sad', 'Angry' and 'Fear'. Data from three modalities are combined to predict the engaged or disengaged state. This information will help the teachers to know the

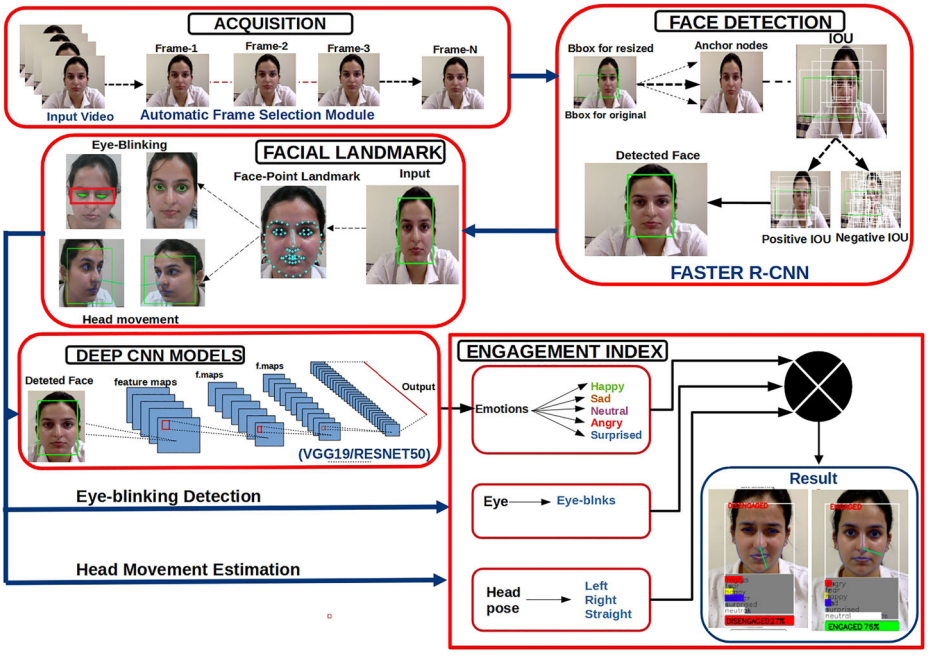

Fig. 2 Proposed framework for engagement detection system

different instances where most of the student's engagement level is falling. This will also help students to get immediate feedback in a real-time environment based on their engagement level. The detailed description of the work of each phase has been further discussed in the sections below.

4.1 Acquisition

The initial step is to capture real-time video streaming from the web-camera of the device that students are using for studying in an e-Learning environment. In this phase, frame-based processing is performed to acquire the useful frames from the real-time video streaming after a fixed time (say 20 seconds for our case). The processing of the selected frames is discussed in the next section.

4.2 Face detection using faster R-CNN

In this phase, the face is detected using the facial detection technique, i.e. Faster R-CNN [27, 60]. Tracking and detecting a face in the image is complicated due to various factors such as pixel values, skin colour, face orientation, occlusion, pose, expression, face position and other limiting factors present in real-time images. But recently, Faster R-CNN (Fast Regions with Convolutional Neural Network) has achieved great success in detecting faces during real-time scenarios by improving performance and computation efficiency [26]. Faster R-CNN is trained and tested on the WIDER face dataset for real-time face detection in the present work. The basic building block of Faster R-CNN is shown in Fig. 3.

The steps followed to detect the face in the image using Faster R-CNN are below.

1. The input image is fed into the CNN (convolutional neural network) layer, which gives feature maps in the input image (224 x 224 x 3). For example, for human face images, the filters learn the shape and colour of the human face that exists through training.
2. The RPN (region proposal network) is applied to each point of the feature maps to generate object proposals (with objectness score given by IoU (Intersection-over-Union)). This is done by placing N-anchors boxes at each location of feature maps in the input image. The anchors indicate the presence of objects of different aspect ratios and sizes

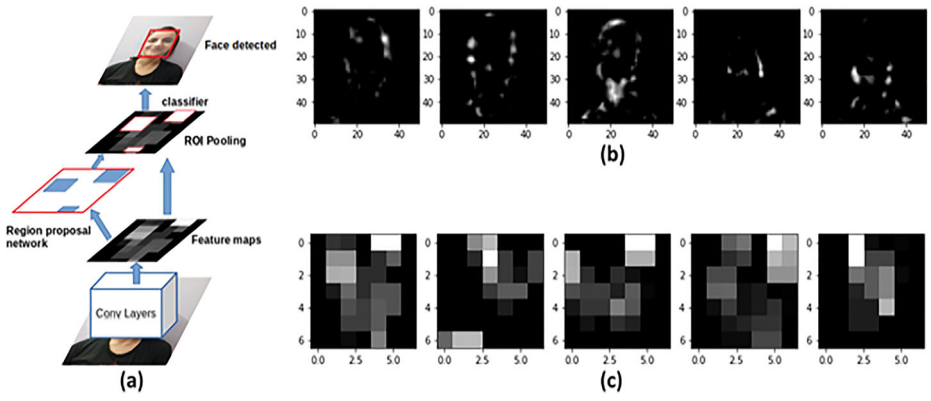

Fig. 3 (a) Building block of Faster R-CNN (b) Visualization of five channels extracted 50 x 50 x 512 feature maps for an input image (c) Visualization of first five ROI's feature maps after ROI pooling

at each location of the output feature maps of the input image. In the present study, the value of N-anchors used is 9, i.e. N=9 and three scales, i.e. 1282, 2562, 5122 pixels and three aspect ratios, i.e. 1:1, 1:2 and 2:1, are used. The visualisation of five channels with 50x50x512 features maps is shown in Fig. 3(b).

3. The ROI pooling layer is applied to make the object proposals (or region proposals) of similar size, and the present study proceeds with only positive IoU. The first five ROI's (Region of interest) feature maps after ROI pooling is shown in Fig. 3(c).
4. Finally, the object proposals are fed to the fully connected layer to get the object class using a softmax layer, and bounding boxes of objects, i.e. ROI (region of interest), are generated using the regression layer.

The loss function of an anchor box 'i' for the RPN is defined as:

$$L(\{P_i\},\{\tau_i\}) = \frac{1}{N_{class}} \sum_i L_{class}(\theta_i, \varphi_i^*) + \lambda \frac{1}{N_{regress}} \sum_i \varphi_i^* L_{regress}(\tau_i, \tau_i^*). \quad (1)$$

Here,

i= index of an anchor for a mini-batch, θ_i = predicted probability of anchor, φ_i^* = ground-truth label, Where $\varphi_i^* = \left\{ \begin{matrix} anchor\ is\ positive\ \ 1 \\ anchor\ is\ negative\ \ 0 \end{matrix} \right\}$ τ_i = a vector having four parameterized coordinates of the predicted bounding box, τ_i^* = a ground-truth box associated with a positive anchor, L_{class} = log loss over two classes, $L_{regress}$ = regression loss, $\varphi_i^* L_{cls}$ = the regression loss is activated only for positive anchors.

4.3 Facial landmark

Facial landmark is the process of localising key points on the specific regions of the face [14, 34, 63]. Identifying landmarks is essential to solve complex image analysis problems such as facial expression analysis, face recognition, gender classification and age estimation. The specific face regions that facial detectors localises are Mouth, Jaw, Right eye, Left eye, Right eyebrow, Left eyebrow and Nose. Dlib's pre-trained landmark detector is used in this paper [29].

After detecting 68-landmark points, as shown in Fig. 4, we consider those facial features in this study that best defines student engagement. We choose eye blinking and head movement as the considerable attributes for the real-time engagement level detection (i.e. engaged or disengaged) are discussed below.

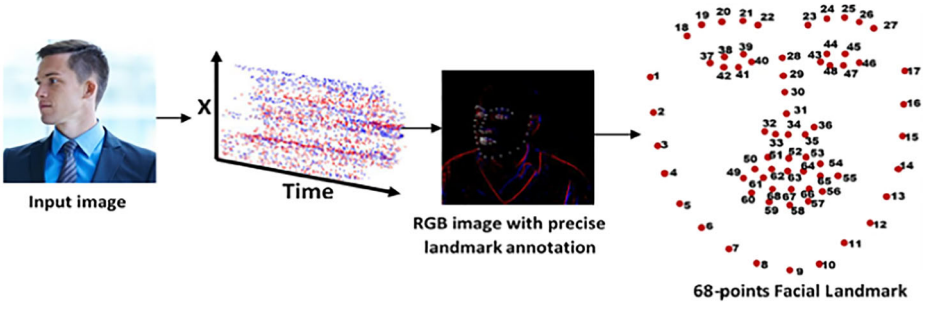

Fig. 4 Illustration of RGB image with precise landmark annotations and position and order of 68 facial landmark points

4.3.1 Eye blinking

The relationship between engagement and blinking pattern is well established in prior studies [21, 59, 70]. So, we propose to derive several features from the eye blink pattern, such as eye blinking frequency and frequency of long eye closure. After detecting 68-landmark points, the focus is on the two landmarks that involve the left and right eye in the real-time video stream [62]. The eyes are represented by 12-coordinate points with 6-(x-y) coordinate points for each eye. The 6-coordinate points (P1-P6) for an open and closed eye is shown in Fig. 5. The aim is to compute the EAR (Eye Aspect Ratio), which helps in detecting the eye-blinks [25, 66].

Using the coordinate points of the eye, the height and width of an eye are computed. The width of an eye is represented by a horizontal line as shown in Fig. 5 and computed by measuring the distance between points P1 and P4. The height of an eye is represented by a vertical line as shown in Fig. 5 and computed by measuring the distance between middle points P2, P3, and P5, P6. The length of these two lines is computed, and then their ratio is computed, referred to as EAR (Eye Aspect Ratio) and its (2) is as follows:

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

EAR for both the eyes is computed in the similar way to (2).

$$EAR = \begin{cases} Eye = Open; \ a > 0.20 \\ Eye = Close; \ a < 0.20 \end{cases} \tag{3}$$

In (3), EAR's interger value is referred by 'a' which is greater than '0'. (3) shows that EAR gives constant value when eye is open and gives close to '0' value for closing an eye. For both the eyes, EAR is averaged as both eyes blink in an synchronous way and its (4) is as follows:

$$Avg_{Ratio} = \frac{EAR_L + EAR_R}{2} \tag{4}$$

The threshold value after computing average EAR is set to 0.20 in this study. To detect engagement, we count the frequency of eye blink in an minute using (5) and also count the

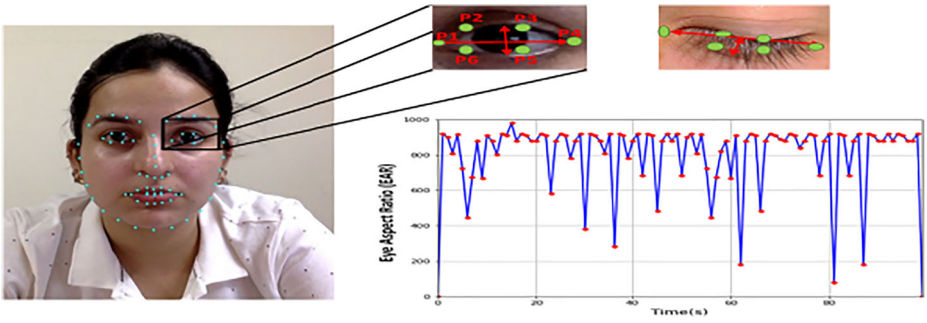

Fig. 5 Schematic representation of the EAR over time frame (in secs) is shown in right lower graph along with 68 facial landmarks on the face image. Eye region is represented by 6 discrete landmark points (P1-P6) for both open and close eye

long eye closure in an minute (i.e. how many times the eye is closed for longer duration in an minute) using an (6).

$$\theta_{Ecount}(Blink_{count}) = \begin{cases} Avg_{Ratio} < 0.20 \\ \&\& \qquad Count = 1 \\ Previous\_Avg_{Ratio} > 0.20 \\ Otherwise \qquad Count = 0 \end{cases} \tag{5}$$

$$\theta_{Lcount}(Eye\_ClosedLong_{count}) = \begin{cases} Avg_{Ratio} < 0.20 \\ \&\& \qquad Count = 1 \\ Conse\_Time_{Frame} > 10secs \\ Otherwise \qquad Count = 0 \end{cases} \tag{6}$$

In (5), both the Average aspect ratio and Previous ratio are considered to calculate the frequency of eye blinks in a minute. It is done because if the Average EAR goes below the threshold, then the state of the eye is assumed to be closed, and two blinks are considered rather than a single blink. Hence, both the present and past ratios are taken to avoid errors. Therefore, the higher blinking rate and higher frequency of long eye closure contribute to the emergence of disengagement.

4.3.2 Head-movements

In this study, we propose an algorithm to estimate the head position. The position of the student's head, i.e. where the student is facing during the ongoing video learning session, is monitored in the present work for detecting engagement level [19, 22, 31, 40, 68]. To do so, standard 3D world coordinates of selected facial landmarks are taken, i.e. tip of the nose, the right corner of the right eye and left corner of the left eye, extreme right and left points of the lips and chin, as shown in Fig. 6.

For better understanding, the algorithm to detect head movement is discussed below:

Steps for Algorithm 1 are:

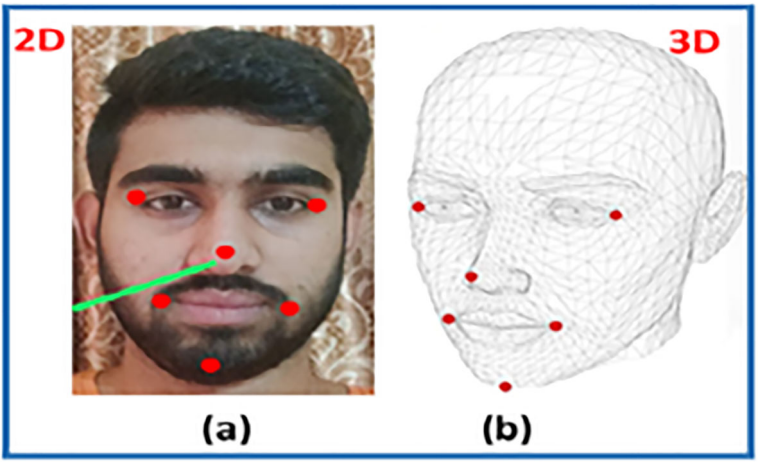

Fig. 6 3D points and its corresponding 2D image

Initialization: head_model ← model of head eight 3D points i.e. P_i,
where, i = {1, 2, ..., 8 }
head_model represents as H_{model}
landmarks denotes as ζ;
R denotes rotation; T denotes translation,
K represents camera intrinsic parameters matrix;

BEGIN:

1. Detect facial landmarks:
$\zeta_i \leftarrow$ detected face landmarks

2. R ← calculated rotation for this frame;
$T \leftarrow$calculated translation for this frame;

3. R, T ← **Solve** PnP (landmarks ζ_i, head_model P_i);
where, PnP problem is solved with the help of Levenberg-Marquardt optimization.

4. Evaluate reprojection ζ_i' of each point P_i of H_{model}:
$\zeta_i' \leftarrow$K[RT$]$P_i$

5.Evaluate$error of each reprojection$:
e←distance(ζ_i', ζ_i)

6. Search which point P_k, where k ∈ (1, 2, ..., 8) generate e, ζ_k' represents facial landmarks point
$H_{model} \leftarrow$all points from H_{model} except P_k'
ζ' ←all points from landmarks except ζ_k

7. Solve PnP using ζ' and H_{model}' R, T = solve PnP (ζ', H_{model}');

8. OUTPUT: R', T'

9. Position=
$$\begin{cases} Right; & angle > 30 \\ Left; & angle < -15 \\ Straight; & otherwise \end{cases}$$

END

Algorithm 1 Head position estimation algorithm.

- First of all, Face-landmarks (ζ_i) were detected as evaluated in the previous phase. The two important parameters are calculated, i.e. rotation (R) and translation (T).
- From (7), the PnP problem is identified, for which Levenberg-Marquardt optimisation is used to solve it and get the final values of rotation and translation (R', T'). The PnP problem equation is as follows:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{7}$$

The left side of the equations denotes the 2D image representation. On the right-hand side of the equation, the first part represents the camera matrix, the middle part represents the rotation and translation, and the last part belongs to the 3D model of the face.

- Head model points (H_{model}), reprojections (ζ_i') are computed every time after calculating the initial head position.
- Next, the reprojection error (e) is computed for each point.
- The calculation of rotation and translation is done repeatedly for all points (P_k) except the biggest error point (ζ_k').
- The calculated values are final rotation and translation values and are represented as (R', T').
- In our proposed system, we assumed the value of the Euler angle for different head positions, i.e. left, right and straight. If the angle is greater than 30, then the position is right, and if the angle is less than -15, then the position is left; otherwise, the position of the head is straight.

4.4 Facial emotion recognition based on deep CNN models

In this paper, deep CNN models are implemented to recognise the facial emotion expressed by students during the e-Learning session [17, 38, 42, 52]. Deep neural networks, which is a prominent classifier to analyse and extract features from images and videos in a lesser computation time with higher accuracy [1, 2, 6, 33]. We have used VGG19 and ResNet-50 as deep CNN models for facial emotion recognition. These models are trained and tested on publicly available standard and benchmarked datasets, i.e., FER-2013, CK+ and own collected images dataset. The deep CNN models classify the facial emotion into six categories based on the facial expression analysis as 'Happy', 'Neutral', 'Surprise', 'Sad', 'Angry' and 'Fear'. The output emotion data is further used to evaluate the student engagement level. An algorithm for facial emotion recognition is proposed, and its main steps are explained below (Fig. 7).

Steps for Algorithm 2 are:

- S_{ID} and δ_E represents the change occurred in the proposed approach for facial emotion recognition each time the weights are updated.
- $WT_{Student}$ (weight matrix corresponding to every student) and $WT_{Emotion}$ (weight matrix correspondents to six emotions) generates random number between 0 and 1 whenever it is called using the random function.
- First of all, face is detected as evaluated in the previous phase.
- Next, emotions ($\delta\_emotion$) are predicted based on the facial expression analysis by fedding the ROI of face to the deep CNN models.

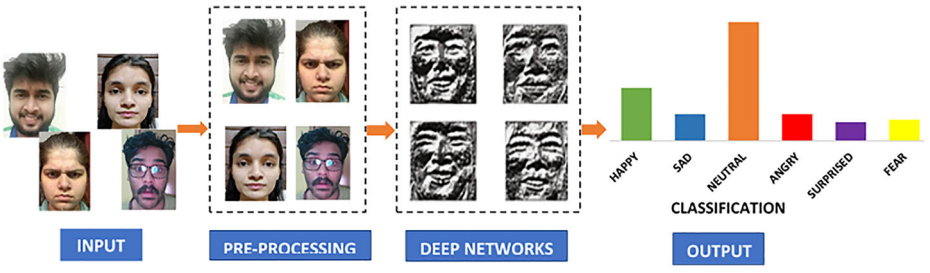

Fig. 7 Steps of facial expression recognition

Input : $S_{ID} \leftarrow$ unique student id
$\delta_E \leftarrow$ label for facial emotions
E $\in \{happy, sad, angry, neutral, surprised\}$
Output : $WT_{Student} \leftarrow$ Student Weights
$\text{WT}_{Emotion} \leftarrow$ Emotion Weights
$\text{MT}_{emotion} \leftarrow$ Emotion Weights Matrix
$\text{WT}_{prediction} \leftarrow$ prediction weights
$\delta_{emotion} \leftarrow$Predicted Emotion
BEGIN:
Random Initialization $\rightarrow WT_{Student}$ and $WT_{Emotion}$

for a=1 to len(S_{ID}) **do**

for b = 1 to len(δ_E) **do**
1. $Detect\_face \leftarrow$ Fast R-CNN models
2. CLASSIFY : $Detect_{face} \rightarrow \delta_E$
where E $\in$ {happy, sad, angry, neutral, surprised}
3. $\delta\_emotion \leftarrow$ Apply Deep CNN models on $Detect_{face}$
4. $\text{WT}_{prediction}[a][b] = \sum_{a=0}^{len(S_{ID})} \sum_{b=0}^{len(\delta)} \text{WT}_{Student}\ WT_{Emotion}$
5. $S_{ID}(\delta_{emotion}) = MT_{emotion}^{[\delta} \times WT_{prediction}$

OUTPUT: $\delta\_emotion$ and $\text{WT}_{Emotion} \rightarrow$ recognised emotion information of an student to help generate the EI (Engagement index) for the engagement state prediction.
END

Algorithm 2 Facial emotion recognition algorithm.

- To calculate the $WT_{prediction}$, all matrix cells of $WT_{Student}$ and $WT_{Emotion}$ are multiplied.
- Based on the values of weight matrix of emotions, online student's emotion are classified $\delta\_emotion$ into six categories by multiplying $MT_{emotion}[\delta]$ and $\times\ WT_{prediction}$
- The calculated values of $WT_{prediction}$ and $\delta\_emotion$ are then used as input to Algorithm 3.

The deep CNN models used in facial emotion recognition are given below.

-VGG19 Deep learning model The network of VGG19 consists of 19 layers [13], and its detailed network architecture is shown in Fig. 8. In the current study, the input layer holds an image of 224x224x3 pixel value. 16 convolution layers are used, with each layer having a kernel size of 3x3x1. When the input image passes through these convolution layers, it computes the low-level features. There are five max-pooling layers in VGG-19 [32, 54]. Then, the last layers comprise fully connected layers, which give input on the probabilities of being in a particular class. The last layer utilises the SoftMax activation function to know

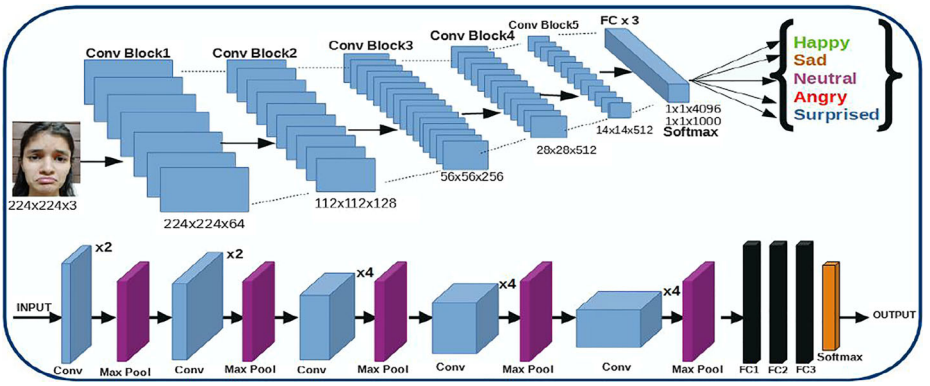

Fig. 8 Network architecture of VGG19

the emotion class of the input image based on the probabilities calculation. We have the probabilities of different emotion classes, based on which the EI will be calculated (Fig. 9).

-ResNet-50 (Residual Network) deep learning model ResNet-50 consists of 50 layers, and the convolution and pooling layers work similarly to the standard CNN [12, 37, 51]. The main part of this model is the residual block which makes the connection between the predictions and actual inputs, as shown in Fig. 10. It means that the model learns from the present input and the output of the previous layer. This is done because the problem of vanishing gradient (accuracy degrading) arises in the case of training the extremely deep networks [49, 50, 53]. For this issue, skip connection is used as it allows the flow of information from lower to higher layer as shown in Fig. 9. The single residual block is represented

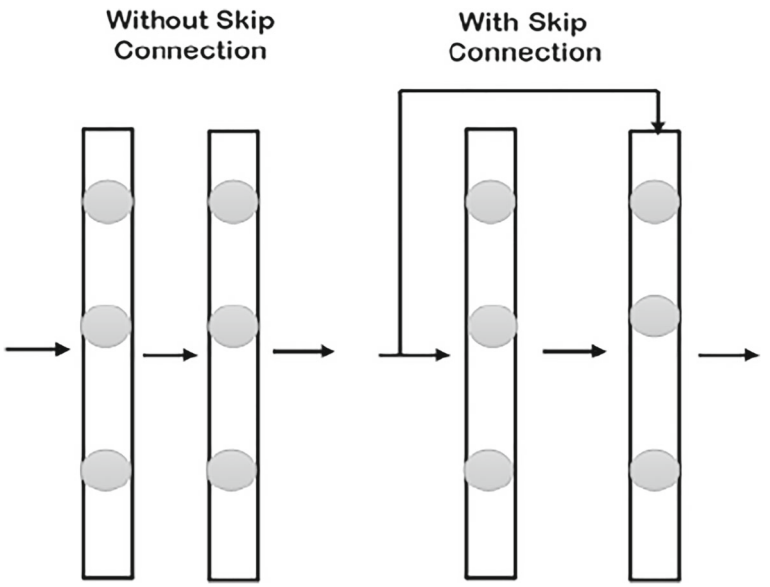

Fig. 9 Skip connection

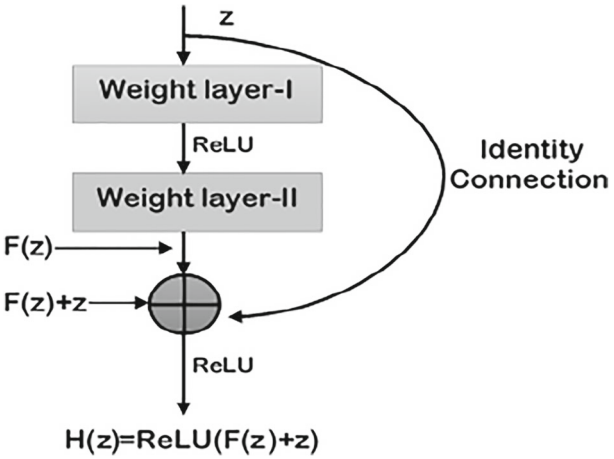

Fig. 10 Single residual block

in Fig. 10. There are five stages in this model, and each stage consists of convolution and identity blocks, and both blocks contain a convolution layer for each of them. The input layer holds images of 224x224x3 pixel value and kernel size of 7x7 and pooling's kernel size of 2x2 [65]. The detailed network architecture of ResNet-50 is shown in Fig. 11.

4.4.1 A multimodal real-time student engagement detection

To evaluate engagement, Algorithm 3 is proposed based on a multimodal approach. The predicted results from the three modalities discussed above are combined to detect the student engagement level. For calculating EI and detecting engagement, some assumptions considered are:

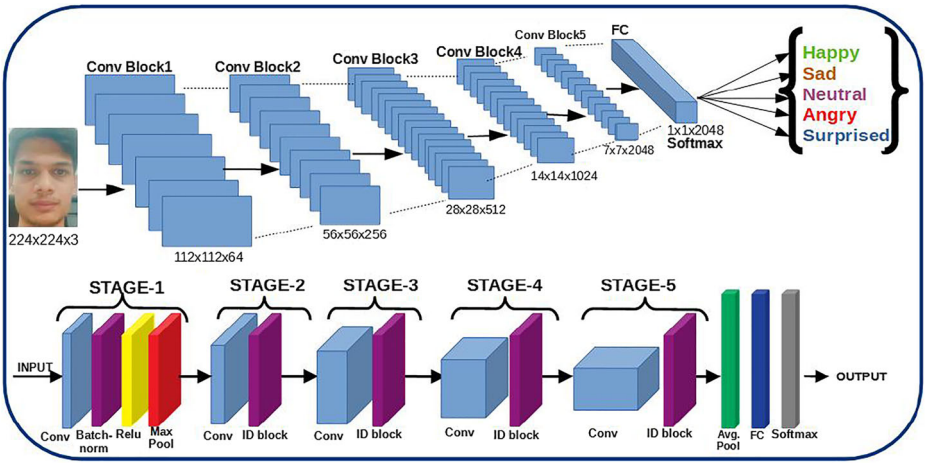

Fig. 11 Network architecture of ResNet-50

- If happy, surprised and neutral emotions are detected, then the engagement state is engaged else disengaged (Algorithm 2).
- If eye-blinking count (5) doesn't exceed 15 and long eye closure count (6) doesn't exceed 3; then the state is engaged else disengaged.
- If the head position is straight, the state is engaged or disengaged (Algorithm 1).
- Lastly, combining the results of three modalities gives the EI (engagement index).

  Combining these modalities gives improved engagement detection results and accuracy rate.

The calculated values of $\delta_{Emotion}, \theta_{Ecount}, \theta_{Lcount}, \zeta$ are then used as an input to Algorithm 3, and further their outcomes are discussed in next section.

Declaration: Emotion denotes as δ; **Eye blink count** denotes as θ; **Head movement** denotes as ζ; **Engagement Index (EI)**={Engaged, Disengaged}
Input: $\delta \in$ {Happy, Sad, Angry, Neutral, Surprised}
$\zeta \in$ {Straight, Left , Right}
$\theta_{Ecount} \in \{15, more\}, \theta_{Lcount} \in \{3, more\}$
Output: EI← Engagement Index
Assumptions:
1. $\delta_{happy}, \delta_{surprised}, \delta_{neutral}$ = Engaged
2. $\delta_{sad}, \delta_{angry}$ =Disengaged
3. $\theta_{Ecount} \leq 15$ = Engaged
4. $\theta_{Ecount} \geq 15$ = Disengaged
5. $\theta_{Lcount} \leq 3$ = Engaged
6. $\theta_{Lcount} \geq 3$ = Disengaged
7. $\zeta_{Straight}$ = Engaged
8. $\zeta_{Left}, \zeta_{Right}$ = Disengaged

for $\epsilon = 1$ to t **do** 1. $Detect_{face} \leftarrow$ Fast R-CNN models
2. $Detect_{face-point} \leftarrow$ landmark face-point extractor
3. $Detect_{eye} \leftarrow$ landmark
4. $Detect_{head-movement} \leftarrow$ landmark
5. Evaluation of $\delta_{emotions} \leftarrow$ Algorithm 2
6. Evaluation of $\theta_{count} \leftarrow$ EQ. 5, EQ. 6
7. Evaluation of $\zeta_{head-movement} \leftarrow$ Algorithm 1

If$\{\delta_{happy} \parallel \delta_{surprised} \parallel \delta_{neutral}\}$ AND $\{(\theta_{count} \leq$ to 15 $) \parallel (\theta_{Lcount} \leq$ to 2$)\}$ AND $\{\zeta_{straigh}\}$
EI→ **Engaged**
elseif $\{\delta_{sad} \parallel \delta_{angry}\}$ AND $\{\theta_{count} \geq 15\}$ AND $\{\zeta_{left} \parallel \zeta_{right}\}$
EI→ **Disengaged**
else
EI→ **Disengaged**

Algorithm 3 Engagement evaluation.

5 Experimental results and discussion

The experimental analysis of the proposed system is presented in this section. The system was tested with 50 undergraduate and graduate students from different fields between 19-25 years old. Students watched the subject, and non-subject related video lectures with 2 minutes and 20 seconds long. The video lectures include verbal and diagrammatic information. Figure 12 represents the visual view of the working system in a real-time e-Learning environment and shows information related to different modalities that contribute to deciding the engagement state at any time instance. The output information of three modalities are fed as input to Algorithm 3, and the EI value is calculated. The value of EI helps to decide the student's engagement state in an e-Learning environment. This will help teachers make the e-Learning content adaptable to the individual student's needs when their engagement level is low.

Figure 13(a) shows that when an eye blink occurs, the EAR value goes close to '0' for closing an eye. The blink period is also shown in Fig. 13(a) that can be used to detect various abnormal patterns. Figure 13(b) shows frequent eye-blinks and long eye-closure, which are computed using (5) and (6) respectively. If the blink count of students studying in an e-Learning environment exceeds 15 minutes, then the student is disengaged. Additionally, if a long eye-closure count is greater than 3 minutes, the state is also disengaged otherwise engaged.

Using Algorithm 1, the Euler angle is calculated to estimate the position of the head at any given time. The results retrieved from Algorithm 1 is displayed in Table 2.

It is observed that if the student is looking right or left, then the engagement level is low and if student looks straight towards the laptop screen then the engagement level is high. The graphs corresponding to different head position is shown in Fig. 14.

For recognising facial emotion, the weights of each emotion is calculated using Algorithm 2. The results retrieved from Algorithm 2 is displayed in Table 3 and representation of results in bar-graph and pie-chart are shown in Fig. 15.

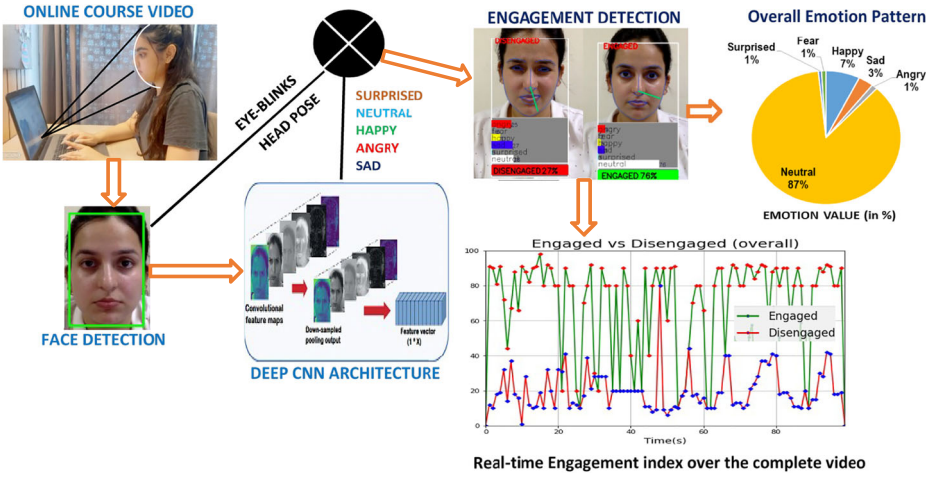

Fig. 12 Visualisation of real-time engagement detection system

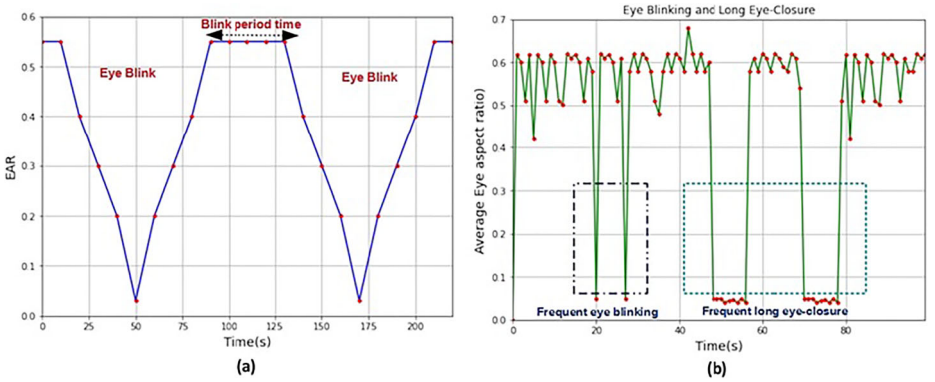

Fig. 13 Example of detected eye movement where (a) Representation eye-blink based on EAR value over time frame (b) Representation of frequent blinks and long eye-closure based on the average EAR

Table 2 Engagement level for a corresponding head position using the Algorithm 3

S.No.	Head Movement	Engaged	Disengaged
1.	Looking Right (angle >threshold)	LOW	HIGH
2.	Looking Left (angle >threshold)	LOW	HIGH
3	Straight Face (angle =0)	HIGH	LOW
4.	Looking Right (angle <threshold)	LOW	MEDIUM
5.	Looking Left (angle <threshold)	LOW	MEDIUM

Note: the angle is eular angle and threshold is -15 to 30

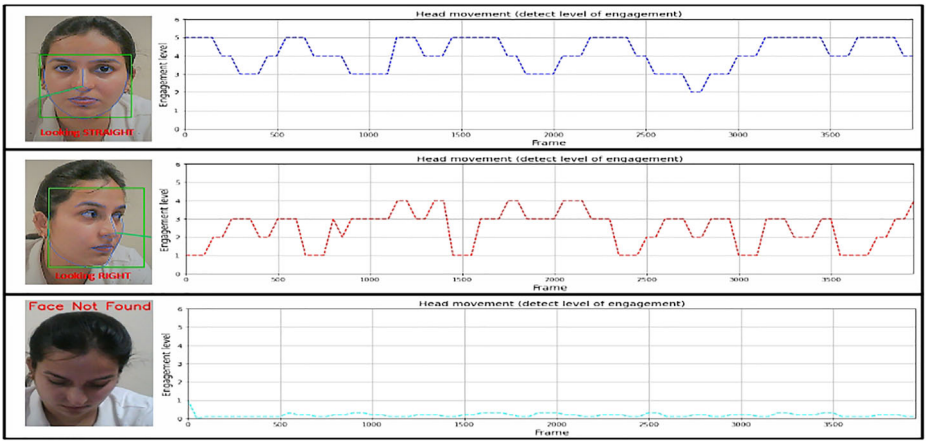

Fig. 14 Graphs show how the head position of student is inferred while studying in a e-Learning environment, within the lapse of time

Table 3 The calculated weight of every emotion using the Algorithm 2

Emotion	Calculated weight
Angry	5.5337
Fear	3.0590
Happy	10.3072
Sad	0.7991
Surprise	9
Neutral	71.301

It has been observed that students who display happy, neutral and surprised emotions while studying are more engaged in studies than those who display sad, angry and fear emotional states. The proposed emotion recognition algorithm is trained and tested with sample data from FER-2013 + CK(+) + own created datasets. And its effectiveness is demonstrated by conducting extensive experiments using deep CNN-based models such as VGG19 and ResNet-50.The hyper-parameters used for VGG19 and ResNet-50 used in the experiments are represented in Table 4. The experimental results are summarised in Table 5.

-Performance metrics The proposed system has been evaluated on four metrics and these metrics are evaluated in terms of false-negative (FN), false-positive (FP), true-negative (TN) and true-positive (TP). The equations for computing accuracy, precision, recall and F1-Score metrics.

Figure 16(a) and (b) display the confusion matrix using the VGG19 and ResNet-50 respectively for the testing data sample collected from FER-2013, CK+ and own created datasets. The accuracy and loss of the proposed model with VGG19 and ResNet-50 are shown in Fig. 17(a), (b), (c) and (d) respectively. From the confusion matrix perspective, the best overall results is shown by VGG19 (Fig. 16(a)).

$\delta_{Emotion}$, θ_{Ecount}, θ_{Lcount}, ζ are fed into the proposed Algorithm 3 and EI is calculated to detect the engagement of the online student. The multimodal system successfully detected whether the student is Engaged or Not Engaged and the EI graph over the complete video lecture are also shown in Fig. 12. The performance of the proposed engagement detection system is then calculated using a confusion matrix (see Table 6). From Table 5 PROP+VGG19 model has highest accuracy rate. FER2013+CK(+)+Own dataset has considered for the evaluation for engaged and dis-engaged. Based on the confusion matrix, the value obtained for accuracy is 92.58%.

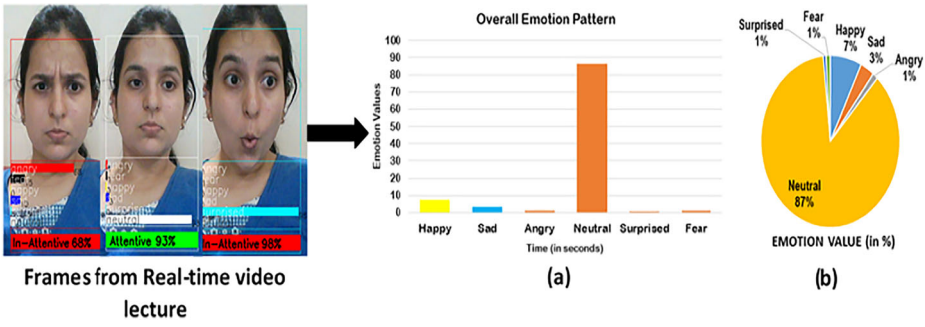

Fig. 15 Facial emotion recognition results from Algorithm 2 for student studying in e-Learning environment

Table 4 Hyper-parameters used in the VGG19, and ResNet50 model

Parameters	VGG19	ResNet-50
Input	(228,228,3)	(228,228,3)
Weight	Initialized to WIDER datset	Initialized to WIDER datset
Optimizer	SGD	Adamax
Loss function	Binary cross entropy	Binary cross entropy
Classifier	Softmax	Softmax
Epocs	20	20
Batch size	128	128
Drop-out Rate	Nil	Nil
Regularization	BatchNormalization	L2 Regularization

Table 5 Performance analysis of ResNet-50 and VGG19 on FER-2013, CK+, and Own Created datasets

S.No	Dataset	MODEL	ACC	P	R	F1
1.	**FER-2013**	PROP+ResNet-50	72.02	71.25	70.11	71.20
		PROP+VGG19	73.40	73.65	71.76	72.69
2.	**CK+**	PROP+ResNet-50	86.23	83.89	85.11	87.38
		PROP+VGG19	89.56	88.92	88.62	88.77
3.	**FER2013+CK(+)**	PROP+ResNet-50	88.89	88.19	88.53	90.14
	+Own Created	PROP+VGG19	90.83	90.40	90.61	92.32

Note:ACC –Accuracy, P - Precision, R-Recall and F1–F1-Score

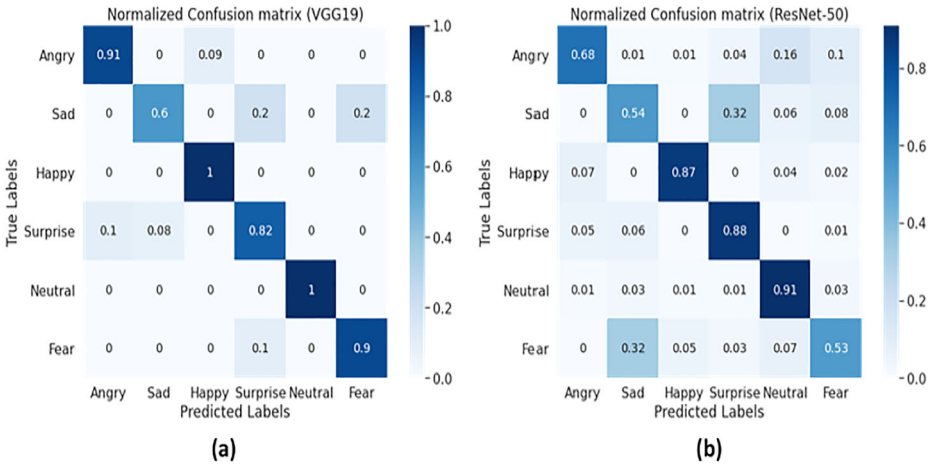

Fig. 16 Normalized confusion matrix for (a) PROPOSED+VGG19 (b) PROPOSED+ResNet-50 models which are trained and tested on (FER2013+(CK+)+Own Created datasets

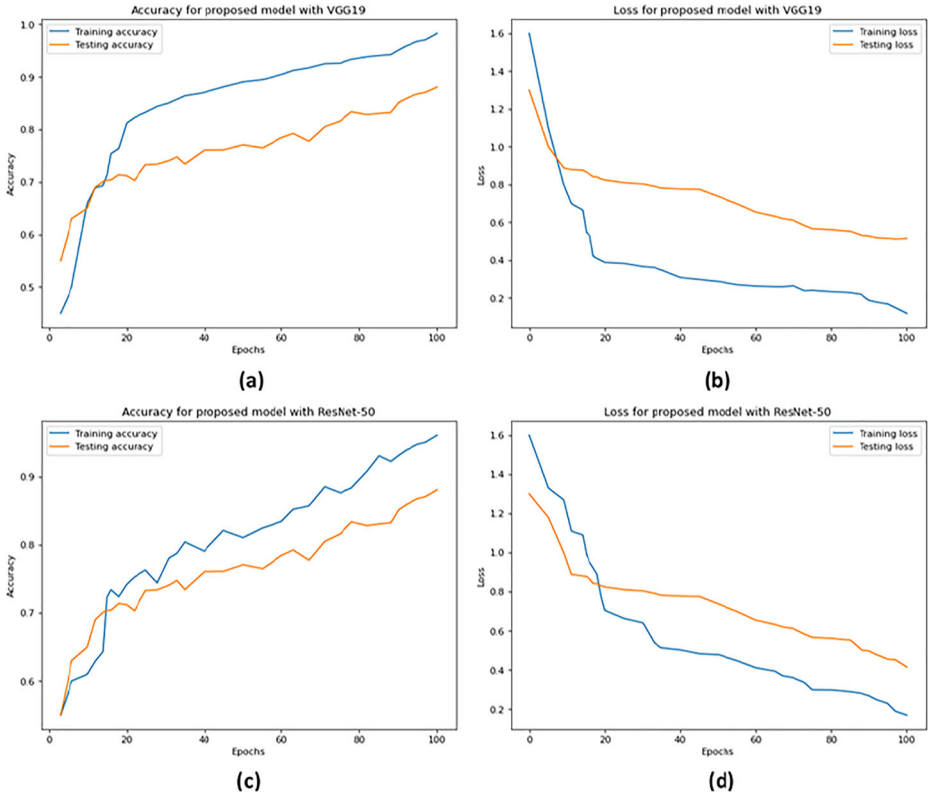

Fig. 17 Analysis of (a) training versus testing accuracy and (b) training versus testing loss for proposed system with VGG19, (c) training versus testing accuracy and (d) training versus testing loss for proposed system with ResNet-50 at different epochs on FER2013+CK(+)+Own dataset

The engagement level obtained is represented in graphical form in Fig. 18. Figure 18(a) shows visualization of engaged state and Fig. 18(b) shows visualization of disengaged state. Figure 18(c)(d) shows overall engagement level for the complete learning session.

6 Discussion

E-Learning has provided the facility of attending lectures from home and watching previously recorded lectures for as much time as students want. But, most of the students leave the

Table 6 Confusion matrix to visualize the performance of algorithm using PROP+VGG19

		Predicted class	
		Engaged-positive	Not-engaged-negative
Actual	Engaged-positive	TP=192	FN=30
	Not-engaged-negative	FP=7	TN=270

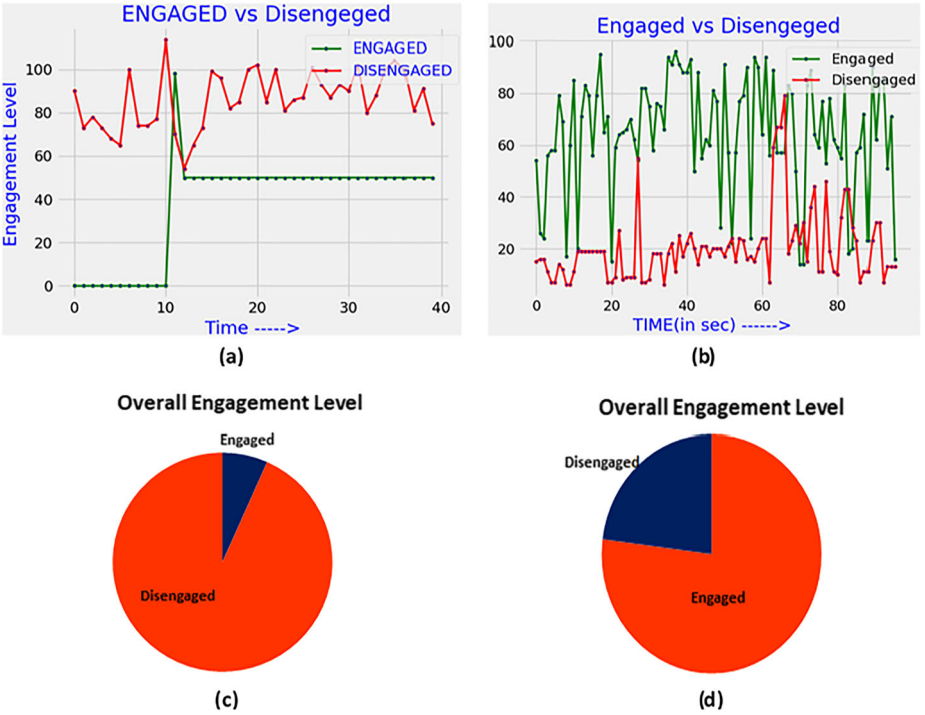

Fig. 18 visualisation of (a) Engaged state (b) Disengaged state over time and Visualisation of overall engagement level for (c) Engaged state (d) Disengaged state

online courses without completing them. This problem is rising for many reasons, such as zero physical interaction with the teacher, no student-to-student interaction, limited online lectures, etc. Hence, an engagement detection system must provide feedback on the learning content so that the e-learning environment can be as interactive as a physical classroom. Therefore, this paper has provided a framework in which different modalities are introduced and combined to predict student engagement in the e-Learning environment. The proposed framework in Fig. 2 can be described as the acquisition, processing, and detection phases. The system automatically collects key image frames from the real-time streaming after every 20 seconds in the acquisition phase. These collected images are buffered and embedded to detect face key points to provide information on eye-blinking (equation 2 to equation 6 for eye-blinking count and long-eye closure) and head movement estimation (Algorithm 1 is proposed) based on the landmark approach. And also, recognition of facial emotion based on deep learning approaches, such as VGG-19 and ResNet-50, as given in Algorithm 2. All three modalities give different engagement levels. Based on the output information from these modalities, we have proposed Algorithm 3, i.e., engagement evaluation, to combine all three detection results to predict the student engagement state. Hence, the proposed system is implemented on various datasets such as FER-2013, CK+ and own datasets for facial emotion recognition, which provides efficient results.

To compare our experimental results with the results of existing works, no standardised system has utilised the combined data of face, eye, and head. A comparison shown in Fig. 7 is based on individual modality accuracy.

Table 7 Comparison of the existing systems with a proposed system for real-time engagement detection

S.NO.	Work	Dataset	Modality	Accuracy (in %)
1.	Thomas et al. 2017 [69]	Own collected	Eye gaze, Head pose	85
2.	Liu et al. 2018 [47]	Own collected	Head pose	70
3.	Nezami et al. 2019 [55]	FER-2013	Facial emotion	72.38
4.	Hasnine et al. 2021 [28]	CK+	Eye gaze, Emotion detection	73.40
5.	Proposed system	FER-2013, CK+, Own collected	Facial emotion, Eye-blinking, Head-movement	**92.58**

The proposed system outperformed other existing engagement detection systems in terms of performance accuracy, as given in Table 7. Our proposed system provided better results than the other approaches used by [28, 47, 55, 69]. Our system used the information from three modalities: facial emotion, eye-blinking, and head movement. Authors in [69] used eye-gaze and head pose as two different modalities with an accuracy value of 85% for a 10-sec video. Authors in [47] used head pose, and an accuracy value of 70% was achieved on their own collected dataset. In [55], authors used emotions for automatic engaged detection with an accuracy value of 70% on the FER-2013 dataset. Authors in [28] used eye-gaze and emotions to visualise emotion detection on the CK+ dataset with an accuracy of 73.4%. Conclusively, we achieved an accuracy rate of 92.58% for engagement state prediction.

6.1 Limitation

As the proposed system is considered for real-time setting to provide the feedback to the teacher after the combined analysis of the students' facial, eye and head-movement features. There are many limitations which can be improved in future research. From the model point of view, there are a lot of convolution layers and filters from which frame by frame images are needed for evaluation. Both VGG19 and ResNet-50 models need GPU for computation. As it considers real-time videos, the buffer space also needs to be high. With respect to system configuration, the system should be well equipped with cameras, GPU, and processors. Another limitation is the lack of pre-trained models. There is a need for a pre-trained model to reduce the burden of evaluation and achieve high accuracy. Therefore, this is also a major aspect which needs to tackle in future research. In addition to this, there is a lack of a standardised data set which specifically provides engagement and disengagement. Hence, in this paper, we proposed a real-time system with better accuracy.

7 Conclusion

The increasing demand for online education during the COVID-19 pandemic required a system capable of determining student engagement for providing regular feedback to the students, one of the biggest challenges for researchers and policy-makers. Because a reliable system using visual data can intelligently adapt according to the student's mental state. The

main aim of this paper is to support the online education system by detecting the engagement level of students. The multimodal approach-based system is proposed that detects the engagement state as either engaged' or disengaged' in the real-time scenario. The multimodal approach contains information from three modalities: facial expression, eye, and head. The deep learning-based models have been applied for recognising facial emotion from facial expression analysis and landmarks detection to monitor eye-blinking and head movement. And the EI (engagement index) is calculated by combining all three modalities of data to predict the real-time engagement state of the student. Finally, the proposed model is used by teachers to know how actively the students are engaged during e-classes by analysing their EI. The experimental results showed that VGG19 had outperformed ResNet-50 with an accuracy of 90.83% for classifying facial emotion. The results also showed that the disengagement is highly related to the frequency of eye blinks, frequent long-eye closure and constantly moving head rather than looking straight at the video lecture streaming screen in a specific period. The proposed engagement detection system is evaluated on 50 online students, and an accuracy of 92.58% is recorded (see Table 7). The proposed system is compared with state-of-the-art methods, and our results have outperformed other existing approaches. In the future, the multimodal approach can be combined with a sensor-based approach such as data from EEG, ECG, etc. The proposed system can be evaluated for special needs students. The performance can be measured using other deep learning approaches also. The present study can be implemented for fields other than the education sector.

Data Availability The data presented in this study are available on request from the corresponding author. Some data are publicly available and some are not.

Declarations

Conflict of Interests 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. Achour B, Belkadi M, Filali I, Laghrouche M, Lahdir M (2020) Image analysis for individual identification and feeding behaviour monitoring of dairy cows based on convolutional neural networks (cnn). Biosyst Eng 198:31–49. Elsevier
2. Anajemba JH, Iwendi C, Mittal M, Yue T (2020) Improved advance encryption standard with a privacy database structure for iot nodes, pp 201–206. IEEE
3. Anas ER, Henriquez P, Matuszewski BJ (2017) Online eye status detection in the wild with convolutional neural networks. In: International conference on computer vision theory and applications, vol 7, pp 88–95. SciTePress
4. Artífice AFVP, Sarraipa J, Jardim-Goncalves R (2021) Improvement of student attention monitoring supported by precision sensing in learning management systems. IntechOpen
5. Azlan CA, Wong JHD, Tan LK, Huri MSNA, Ung NM, Pallath V, Tan CPL, Yeong CH, Ng KH (2020) Teaching and learning of postgraduate medical physics using internet-based e-learning during the covid-19 pandemic–a case study from malaysia. Physica Med 80:10–16. Elsevier
6. Bargshady G, Zhou X, Deo RC, Soar J, Whittaker F, Wang H (2020) Enhanced deep learning algorithm development to detect pain intensity from facial expression images. Expert Syst Appl 149:113305. Elsevier
7. Bhardwaj P, Gupta P, Panwar H, Siddiqui MK, Morales-Menendez R, Bhaik A (2021) Application of deep learning on student engagement in e-learning environments. Computers & Electrical Engineering 93:107–277. Elsevier
8. Biju S. M, Salau A, Eneh J, Sochima V, Ozue I (2020) A novel pre-class learning content approach for the implementation of flipped classrooms

9. Cai Z, Gu Z, Yu ZL, Liu H, Zhang K (2016) A real-time visual object tracking system based on kalman filter and mb-lbp feature matching. Multimed Tools Appl 75(4):2393–2409
10. Carlotta Olivetti E, Violante MG, Vezzetti E, Marcolin F, Eynard B (2020) Engagement evaluation in a virtual learning environment via facial expression recognition and self-reports: a preliminary approach. Appl Sci 10(1):314. Multidisciplinary Digital Publishing Institute
11. Chang C, Zhang C, Chen L, Liu Y (2018) An ensemble model using face and body tracking for engagement detection. In: Proceedings of the 20th ACM international conference on Multimodal interaction, pp 616–622
12. Chauhan S, Mittal M, Woźniak M, Gupta S, Pérez de Prado R (2021) A technology acceptance model-based analytics for online mobile games using machine learning techniques. Symmetry 13(8):1545
13. Chowdary MK, Nguyen TN, Hemanth DJ (2021) Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications, pp 1–18. Springer
14. Daza R, Morales A, Fierrez J, Tolosana R (2020) Mebal: a multimodal database for eye blink detection and attention level estimation. In: Companion publication of the 2020 international conference on Multimodal interaction, pp 32–36
15. Dewan MAA, Lin F, Wen D, Murshed M, Uddin Z (2018) A deep learning approach to detecting engagement of online learners, pp 1895–1902. IEEE
16. Dewan M, Murshed M, Lin F (2019) Engagement detection in online learning: a review. Smart Learning Environments 6(1):1–20
17. Fabian Benitez-Quiroz C, Srinivasan R, Martinez AM (2016) Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5562–5570
18. Giannopoulos P, Perikos I, Hatzilygeroudis I (2018) Deep learning approaches for facial emotion recognition: a case study on fer-2013. In: Advances in hybridization of intelligent methods, pp 1–16. Springer
19. Goldberg P, Sümer Ö, Stürmer K, Wagner W, Göllner R, Gerjets P, Kasneci E, Trautwein U (2021) Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. Educ Psychol Rev 33(1):27–49. Springer
20. Gopane S, Kotecha R (2022) Enhancing monitoring in online exams using artificial intelligence. In: Proceedings of international conference on data science and applications, pp 183–193. Springer
21. Gotlieb RJ, Yang X-F, Immordino-Yang MH (2021) Measuring learning in the blink of an eye: adolescents' neurophysiological reactions predict long-term memory for stories. In: Frontiers in education, vol 5, pp 285. Frontiers
22. Gupta S (2015) A correction model for real-word errors. Procedia Computer Science 70:99–106
23. Gupta S (2018) Facial emotion recognition in real-time and static images. In: 2018 2nd international conference on inventive systems and control (ICISC), pp 553–560. IEEE
24. Gupta SK, Ashwin T, Guddeti RMR (2019) Students' affective content analysis in smart classroom environment using deep learning techniques. Multimedia Tools and Applications 78(18):25321–25348. Springer
25. Gupta S, Gouttam D (2017) Towards changing the paradigm of software development in software industries: An emergence of agile software development pp 18–21. IEEE
26. Gupta S, Kumar P (2021) Attention recognition system in online learning platform using eeg signals, pp 139–152
27. Gupta S, Kumar P, Tekchandani RK (2022) Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimed Tools Appl, pp 1–30
28. Hasnine MN, Bui HT, Tran TTT, Nguyen HT, Akçapınar G, Ueda H (2021) Students' emotion extraction and visualization for engagement detection in online learning. Procedia Computer Science 192:3423–3431. Elsevier
29. Herbig N, Düwel T, Helali M, Eckhart L, Schuck P, Choudhury S, Krüger A (2020) Investigating multimodal measures for cognitive load detection in e-learning. In: Proceedings of the 28th ACM conference on user modeling, adaptation and personalization, pp 88–97
30. Hung K-C, Lin S-F (2022) An adaptive dynamic multi-template correlation filter for robust object tracking. Appl Sci 12(20):10221
31. Iwendi C, Khan S, Anajemba JH, Mittal M, Alenezi M, Alazab M (2020) The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems. Sensors 20(9):2559
32. Javed AR, Sarwar MU, Khan S, Iwendi C, Mittal M, Kumar N (2020) Analyzing the effectiveness and contribution of each axis of tri-axial accelerometer sensor for accurate activity recognition. Sensors 20(8):2216

33. Jiang Y, Li C (2020) Convolutional neural networks for image-based high-throughput plant phenotyping: a review. Plant Phenomics 2020. Science Partner Journal
34. Kamath S, Singhal P, Jeevan G, Annappa B (2021) Engagement analysis of students in online learning environments. In: International conference on machine learning and big data Analytics, pp 34–47. Springer
35. Kanematsu H, Ogawa N, Shirai T, Kawaguchi M, Kobayashi T, Barry DM (2016) Blinking eyes behaviors and face temperatures of students in youtube lessons–for the future e-learning class. Procedia Computer Science 96:1619–1626. Elsevier
36. Krithika L, GG LP (2016) Student emotion recognition system (sers) for e-learning improvement based on learner concentration metric. Procedia Computer Science 85:767–776. Elsevier
37. Lai Z, Chen R, Jia J, Qian Y (2020) Real-time micro-expression recognition based on resnet and atrous convolutions. Journal of Ambient Intelligence and Humanized Computing, pp 1–12. Springer
38. Li S, Deng W, Du J (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2852–2861
39. Li Y-Y, Hung Y-P (2019) Feature fusion of face and body for engagement intensity detection. In: 2019 IEEE international conference on image processing (ICIP), pp 3312–3316. IEEE
40. Li S, Lajoie SP, Zheng J, Wu H, Cheng H (2021) Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving. Computers & Education 163:104114. Elsevier
41. Li J, Ngai G, Leong HV, Chan SC (2016) Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics. ACM SIGAPP Applied Computing Review 16(3):37–49. ACM New York, NY, USA
42. Li Y, Zeng J, Shan S, Chen X (2018) Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Trans Image Process 28(5):2439–2450. IEEE
43. Liao X, Li K, Yin J (2017) Separable data hiding in encrypted image based on compressive sensing and discrete fourier transform. Multimed Tools Appl 76(20):2073920753
44. Liao X, Qin Z, Ding L (2017) Data embedding in digital images using critical functions. Signal Process Image Commun 58:146–156
45. Liao X, Shu C (2015) Reversible data hiding in encrypted images based on absolute mean difference of multiple neighboring pixels. J Vis Commun Image Represent 28:21–27
46. Liu S, Tao X, Gui Q (2019) Research on emotional state in online learning by eye tracking technology. In: Proceedings of the 2019 4th international conference on intelligent information processing, pp 471–477
47. Liu YJ, Zhang M, Rao C (2018) Student engagement study based on multi-cue detection and recognition in an intelligent learning environment. Multimedia Tools and Applications 77(21):28749–28775. Springer
48. Majstorović I, Ahac M., Madejski J. (2022) Influence of the analytical segment length on the tram track quality assessment. Appl Sci 12(19):10036
49. Mittal M, Iwendi C, Khan S, Rehman Javed A (2021) Analysis of security and energy efficiency for shortest route discovery in low-energy adaptive clustering hierarchy protocol using levenberg-marquardt neural network and gated recurrent unit for intrusion detection system. Transactions on Emerging Telecommunications Technologies 32(6):3997
50. Mittal M, Kobielnik M, Gupta S, Cheng X, Wozniak M (2022) An efficient quality of services based wireless sensor network for anomaly detection using soft computing approaches. Journal of Cloud Computing 11(1):1–21
51. Mittal M, Kumar K (2014) Network lifetime enhancement of homogeneous sensor network using art1 neural network, pp 472–47. IEEE
52. Mittal M, Saraswat LK, Iwendi C, Anajemba JH (2019) A neuro-fuzzy approach for intrusion detection in energy efficient sensor routing, pp 1–5. IEEE
53. Mittal M, Srinivasan S, Rani M, Vyas O (2017) Type-2 fuzzy ontology-based multi-agents system for wireless sensor network, pp 2864–2869. IEEE
54. Mittal M, de Prado RP, Kawai Y, Nakajima S, Muñoz-Expósito JE (2021) Machine learning techniques for energy efficiency and anomaly detection in hybrid wireless sensor networks. Energies 14(11):3125
55. Mohamad Nezami O, Dras M, Hamey L, Richards D, Wan S, Paris C (2019) Automatic recognition of student engagement using deep learning and facial expression. In: Joint european conference on machine learning and knowledge discovery in databases, pp 273–289. Springer
56. Murshed M, Dewan MAA, Lin F, Wen D (2019) Engagement detection in e-learning environments using convolutional neural networks, pp 80–86. IEEE
57. Parthiban L, Samy SS (2021) Emotion detection in iot-based e-learning using convolution neural network. Fuzzy Intelligent Systems: Methodologies, Techniques, and Applications, pp 27–44 . Wiley Online Library

58. Qureshi SA, Hussain L, Chaudhary Q-u-a, Abbas SR, Khan RJ, Ali A, Al-Fuqaha A (2022) Kalman filtering and bipartite matching based super-chained tracker model for online multi object tracking in video sequences. Appl Sci 12(19):9538
59. Ranti C, Jones W, Klin A, Shultz S (2020) Blink rate patterns provide a reliable measure of individual engagement with scene content. Sci Rep 10(1):1–10. Nature Publishing Group
60. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28:91–99
61. Salau AO, Jain S (2019) Feature extraction: a survey of the types, techniques, applications, pp 158–164. IEEE
62. Salzillo G, Natale C, Fioccola GB, Landolfi E (2020) Evaluation of driver drowsiness based on real-time face analysis. In: 2020 IEEE international conference on systems, man, and cybernetics (SMC), pp 328–335. IEEE
63. Sharma A, Gupta S, Kaur S, Kumar P (2019) Smart learning system based on eeg signals, pp 465–476. Springer
64. Sheikh AA, Mir J (2021) Machine learning inspired vision-based drowsiness detection using eye and body motion features. In: 2021 13th international conference on information & communication technology and system (ICTS), pp 146–150. IEEE
65. Siriaraya P, Takumi K, She WJ, Mittal M, Kawai Y, Nakajima S (2022) Investigating the use of spatialized audio augmented reality to enhance the outdoor running experience. Entertainment Computing, pp 100534
66. Srivastava S (2021) Driver's drowsiness identification using eye aspect ratio with adaptive thresholding. In: Recent trends in communication and electronics, pp 151–155. CRC Press
67. Su M-C, Cheng C-T, Chang M-C, Hsieh Y-Z (2021) A video analytic in-class student concentration monitoring system. IEEE Transactions on Consumer Electronics. IEEE
68. Sümer Ö, Goldberg P, D'Mello S, Gerjets P, Trautwein U, Kasneci E (2021) Multimodal engagement analysis from facial videos in the classroom. arXiv:2101.04215
69. Thomas C, Jayagopi DB (2017) Predicting student engagement in classrooms using facial behavioral cues. In: Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education, pp 33–40
70. Yan H, Wang X, Liu Y, Zhang Y, Li H (2021) A new face detection method based on faster rcnn. In: Journal of physics: conference series, vol 1754, pp 012209. IOP Publishing
71. Yang S, Luo P, Loy C-C, Tang X (2016) Wider face: a face detection benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5525–5533
72. Zhang Z, Li Z, Liu H, Cao T, Liu S (2020) Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. J Educ Comput Res 58(1):63–86. SAGE Publications Sage CA: Los Angeles, CA
73. Zou C, Li P, Jin L (2021) Online college english education in wuhan against the covid-19 pandemic: student and teacher readiness, challenges and implications. Plos one 16(10):025–8137. Public Library of Science San Francisco, CA USA