Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Comprehensive Survey of Students' Engagement Detection for Online Classroom Management

Sanjiu Raja Brahma
Central Institute of Technology Kokrajhar
Ranjan Maity
r.maity@cit.ac.in

Central Institute of Technology Kokrajhar

Research Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7331623/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Comprehensive Survey of Students' Engagement Detection for Online Classroom Management

Abstract

Online classes have become a mainstream source of learning over the past few years, especially during the COVID-19 pandemic and its post-period. The convenience and ease of use of online learning have attracted many learners to e-learning, alongside traditional learning modes. In the online learning approach, continuously tracking each learner's affective state or level of engagement often becomes difficult, especially in a large classroom. Consequently, there is a need for an automatic method to detect the learners' affective state. Over the last few years, several works have been targeted at addressing this problem. Among different techniques reported in the literature, machine/deep learning-based approaches used for image analysis are mostly common. This paper reviews the existing computer vision-based engagement detection methods in online learning to understand and summarize the process of identifying the affective state using emotion recognition. To carry out this survey, we used a different indexed dataset of literature published in the last ten years, from 2015 to 2025. Using four search strings and six inclusion-exclusion criteria, fifty-one research papers were considered in this study. To categorically analyze these works, nine research questions were formed. It was observed that existing works can identify student emotions with high accuracy, going up to 99% in some works. However, there is a need to focus on the multiple occluded images, images in low light conditions, as they cause loss of facial details and introduce image noise, which can adversely affect face parsing. There is also a need for further research on the association of emotion with the engagement detection of the students.

Keywords: engagement detection, affective state, emotion recognition, online class monitoring, computer vision

1 Introduction

Active online learning is a form of distance education, where the teacher and students are physically separated and often use computers, mobiles, and the internet as the delivery mechanism [1]. It was first introduced in 1982 by the Western Behavioral Sciences Institute in La Jolla, California [2]. Since then, significant development has occurred in the area, particularly the COVID-19 pandemic, which forced educational

institutes to close temporarily has highlighting the importance of online learning platforms when teaching in a physical classroom was not possible. Continuing to this day, the availability of massive open online courses(MOOCs) has attracted learners looking to upgrade their skills to take up online learning. Online learning has the advantage of accessibility, flexibility, and convenience. Learners can participate in asynchronous recorded video lectures or in synchronous interactive live lectures in the form of online learning, whichever meets their convenience [3]. Thus, helping learners from different geographical locations to connect and collaborate with their peers[4]. However, there are many limitations. Often, it becomes difficult for a teacher to continuously monitor the affective state of engagement of each student as the number of students increases.

A learner's engagement is one of the key factors for effective learning. Studies have shown that students who are engaged in the learning process perform better than their peers [5, 6]. Engagement in learning is defined as being mentally and physically involved or engaged in an effective learning process. Lase et al. [7] describe student engagement as the degree of attention, curiosity, interest, optimism, and passion students show when they are learning or being taught.

Many different factors can contribute to a learner's engagement and disengagement. A student can get disengaged due to distractions and boredom [8]. Long, monotonous lectures and teaching styles can also affect the engagement state. Consequently, there is a need for a teacher to identify the level of engagement of each student and adjust the teaching style most suitable. Another reason for disengagement in online learning is the *Computer Vision Syndrome*. Prolonged staring at the computer screen for consecutive classes can put strain on the eyes.

Methods for automated engagement detection by analyzing engagement are essential to maintain a healthy and effective learning environment. Apart from that, it is also required that the engagement detection method generate its result in real-time for identifying disengagement. Consequently, measures to re-engage the learners need special attention.

Several works have been reported for detecting the engagement level of an individual. Dewan et al. [9] reviewed different methods of engagement detection and classified them into three main categories- automatic, semi-automatic, and manual. Each of these categories can again be divided into subcategories based on the type of data used. In the manual category, the learners are directly involved in the engagement detection process. These techniques include observational checklists and self-reporting. Engagement tracing is a method in the semi-automatic category. In this approach, a learner's indirect involvement is necessary for engagement detection. Automatic category of engagement detection uses non-intrusive data from learners through computer vision and physiological sensors such as EEG and heart rate monitors. In computer vision-based systems, webcams or surveillance cameras are used to capture the learners' participation to estimate learner's engagement level. The taxonomy of the learner's engagement detection methods, as given in [9], is shown in Figure 1. This paper primarily focuses on the automatic engagement detection category using computer vision, which analyses visual features - facial expression, emotion, and behavior or action.

Manual, semi-automatic, and log-file based engagement detection methods generally assess learner engagement after the learning process, making them unsuitable

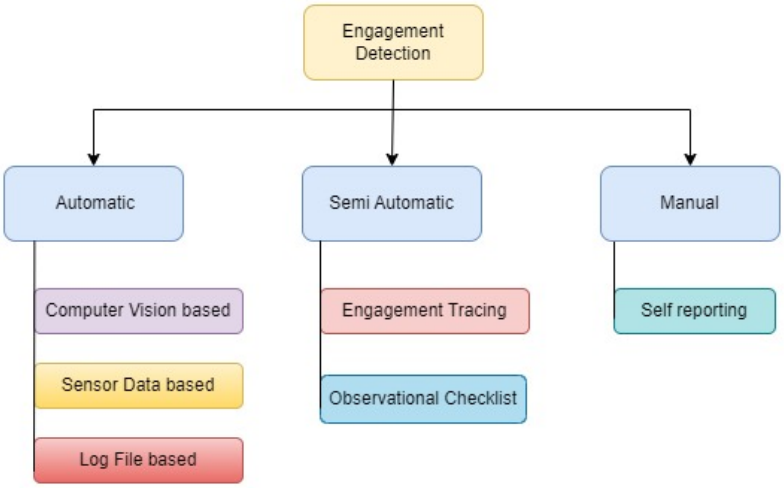


Fig. 1 Taxonomy of learners engagement detection methods as given in [9]

.

for real-time implementation. Automatic methods using psychological or neurological sensors require additional hardware, increasing implementation cost and complexity, while also requiring proper user training for device usage and data interpretation.

Considering widely accessible technological devices, this work focuses exclusively on computer vision-based methods, which only require a webcam-equipped computer or a mobile phone to continuously and non-intrusively capture learners' facial images.

One major challenge for Computer Vision-based methods is protecting the user's privacy. Since this method involves capturing images and videos using a webcam, which is inherently personal and biometric in nature, ethical practices are not optional but fundamental to how the system should be designed, deployed, and evaluated.

There are already survey papers on engagement detection in e-learning environments [9–11]. However, they reviewed engagement detection methods as a whole rather than a specific category of methods. In such reviews, a great deal of introductory knowledge about different ways of engagement detection can be learned. However, a detailed and in-depth review of any particular method is often ignored. Focusing on a specific area within a broader subject helps us understand the methods, the proposed methodology, the progress, and the issues related to that specific area of study. Therefore, it is important to narrow down the area of study for an effective review of the subject. In addition, [12] reviewed the academic emotion classification in their work but did not precisely analyze the relationship between emotion and engagement.

In this paper, we present a review of recently developed computer vision-based Engagement detection techniques using Facial Expression Recognition(FER) for an online classroom setup published between 2015 and to date. Our work, in addition to reviewing existing FER methods, also analyzes and reports teacher-student relationships and their roles in creating a healthy learning environment. We also discuss the ethical practices required to ensure a safe environment for the learners, in the absence

of which, sensitive private data may be misused. The remainder of this paper is organized as follows. Section 2 presents the general framework and taxonomy of computer vision-based methods. Section 3 reviews the literature related to engagement detection by dividing it into four steps and an additional step to discuss the datasets. Section 4 describes the methodology for collecting relevant literature, performing the research work, and forming the research questions. Section 5 presents answers to the research questions formed in the previous section, followed by a brief discussion about identified research gaps and ethical issues in Section 6. Finally, Section 7 concludes the paper.

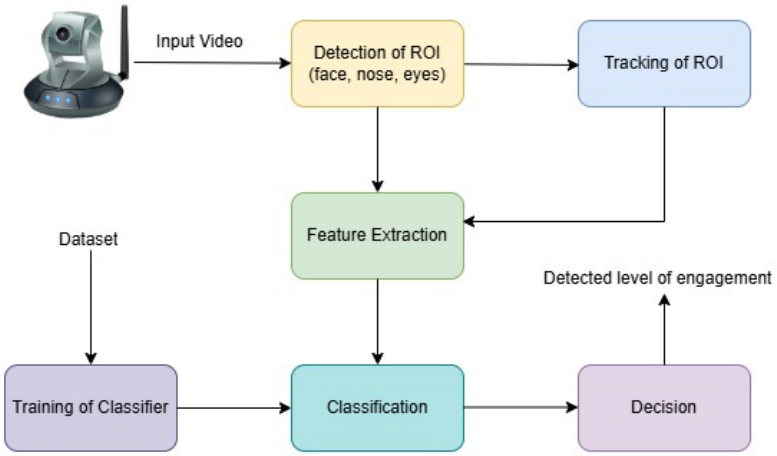


Fig. 2 Framework for computer vision-based method

2 General Framework and taxonomy

Engagement detection techniques are generally a sequential combination of different types of smaller algorithms, each defining a step in the process. Based on the existing literature and our understanding, the general framework for computer vision-based methods can be visualized as shown in Figure 2. The first step is to detect the presence of the object of interest and the regions within it. Optionally, the detected region of interest may be tracked to reduce computation rather than re-detecting it every frame. The next steps that follow are the feature extraction, where facial features are extracted in the form of graphs or activation maps, before the final decision by the classifier. Methods like CNN usually do not require an independent feature extractor, as CNN is already designed to handle both. The classifier is required to be pre-trained for the specific task, which in this case is FER, using an appropriate dataset before actual classification of emotion for the input. Finally, the predicted emotion is mapped to the engagement state.

In the previous section, the classification of different engagement detection approaches based on the type of input data and the involvement of the individual

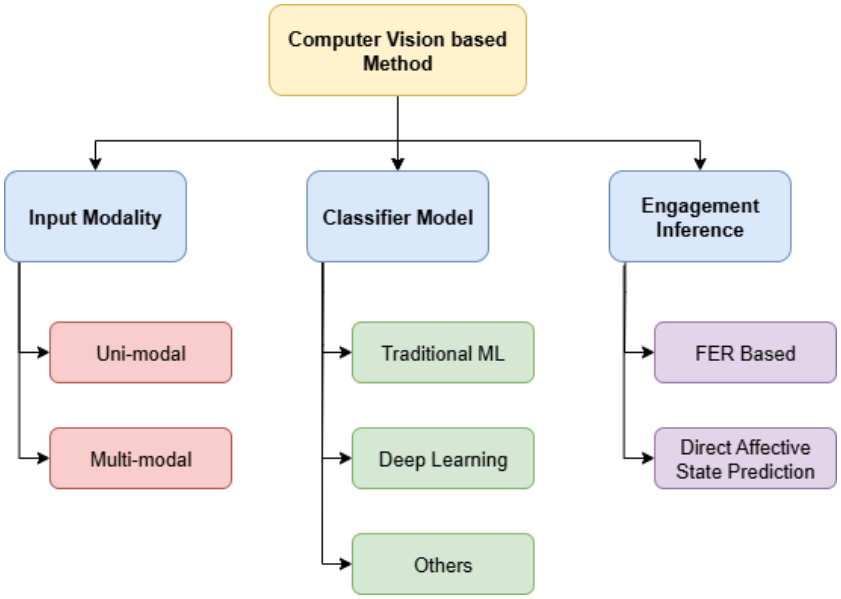


Fig. 3 Taxonomy for computer vision-based methods.

providing the data was presented. Continuing the classification, the taxonomy of computer vision-based methods can be extended as given in Figure 3. Modality refers to the different data used for detecting the learner's engagement. It can be categorized into two types: unimodal and multimodal. The uni-modal approach is based solely on the FER for engagement detection, while multi-modal implements FER as the primary method, and additional data like body posture, mouse movement, keyboard strokes, and physiological sensor data is used to finetune the mapping of engagement state. Based on the type of classifier model used, it can again be categorized into three types. Methods that implement traditional machine learning are relatively faster than deep learning-based methods, but deep learning models achieve better accuracy for challenging and complicated tasks [13, 14]. Traditional ML algorithms, unlike some deep learning algorithms (CNN), are not designed to directly extract features from an image, so image processing techniques are implemented for feature extraction. Another difference between the two is that traditional ML can be trained on a small dataset while deep learning requires thousands or even millions of data samples. Again, based on the engagement inference, there are two categories of methods: FER-based and Direct Affective-State prediction. In FER-based methods, first the learner's facial emotion is recognized, then engagement is predicted based on the dominant emotion or set of emotions. In the direct method, the FER classifier model is trained directly to predict the engagement or affective state without classifying the emotion.

3 Literature Review

Over the past few years, many research studies have been performed to monitor the learners' engagement during online classes using computer vision and machine learning

or deep learning. The process of engagement detection can be divided into four major steps. The algorithms used in these steps by the existing methods have been discussed in this section, along with a brief description of the datasets used for training the algorithms at the beginning of the section.

3.1 Dataset

Data is an important component of any machine learning model. It has been widely acknowledged that a large, well-annotated, and relevant dataset is required for training and evaluating the detection or recognition algorithm. As such, many datasets related to facial expression and emotion recognition have been developed in the past few years. These datasets exist in both single-individual settings for individual emotion recognition and in a group setting for multiple individual emotion recognition in a single frame. But, in the context of online learning, the facial image obtained is usually in a single individual setting, and so only those datasets have been used in the collected literature and hence reviewed in our work. There also exist datasets for engagement prediction similar to FER datasets. The datasets found to have been used for facial expression recognition and engagement detection in related literature are listed in Table 1.

Table 1 Datasets used in related literature.

Dataset	**Total Samples**	**Total Subjects**	**Condition**	**Expression distribution**	**Used in**
DAiSEE[15]	9068 video snippets	122 individuals	Spontaneous	4 affective states and 4 intensities	[16–22]
AffectNET[23]	450000 images	Internet images	Wild	8 basic emotions	[24–26]
FER2013[27]	35887 images	N/A	Posed & Spontaneous	7 basic emotions	[22, 28–37]
RAF-DB[38]	29672 images	Internet images	Posed & Spontaneous	7 basic and 12 compound emotions	[39]
JAFFE[40]	213 images	10 individuals	Posed	7 basic emotions	[41, 42]
CK+[43]	593 image sequence	123 individuals	Posed & Spontaneous	7 basic emotions	[37, 42, 44]
Jonathan Oheix[45]	35.9k images	N/A	Posed	7 basic emotions	[46]
BAUM-2[47]	1047 video clips	286 individuals	Posed	7 basic emotions	[48]
EngageWild[49]	195 video clips	78 individuals	Spontaneous	4 affective states	[25]
VERESEE[50]	3525 video snippets	88 individuals	Spontaneous	4 affective states	[50]
NVIE[51]	236 images	215 individuals	Posed & Spontaneous	6 basic emotions	[52]
UPNA[53]	120 videos	10 individuals	Posed	n/a	[54]
Independent or Own	N/A	N/A	N/A	6 basic emotions(mostly)	[24, 37, 39, 55–60]

Among all the literature, the dataset that has been used the most is the DAiSEE or Dataset for Affective States in E-Environments introduced by Gupta et al.[15]. This dataset consists of 9068 video snippets captured from 112 individuals, which were captured using a webcam mounted on a computer while watching a test video. This dataset recognizes four different affective states of the users- boredom, confusion, engagement, and frustration. The data is crowd-annotated into four levels - very low, low, high, and very high for each of the affective states.

Mollahosseini et al. [23] introduced the AffectNet dataset, which consists of over 1 million facial images collected from the internet using three major search engines and 1250 emotion-related works in six different languages. Almost half of the images were manually annotated by twelve human experts against the presence of seven different facial expressions and the intensity of valence and arousal.

Another widely used dataset introduced by Goodfellow et al. [27] is the FER2013 dataset. This dataset consists of approximately thirty thousand grayscale facial images of size 48x48 pixels. Each image has been annotated as classified into one of the seven basic emotions based on the facial expression. The dataset has been divided into training and testing sets with 28,709 samples in the training set and two sets of 3,589 samples in the testing sets.

The Real-world Affective Faces Database (RAF-DB), as introduced by Li et al. [38] which consists of about 30K facial images sourced from the internet. The data has been labeled by 40 annotators with seven basic or twelve classes of two-label compound emotions. Lyons et al. [40] introduced the Japanese Female Facial Expression or JAFFE dataset, which consists of 213 facial images from 10 individuals showing seven basic facial expressions. The Extended Cohn-Kanade or CK+ data set was introduced by Lucey et al. [43]. It consists of 593 image sequences collected from 123 individual subjects. Each video sequence shows a transition of the facial expression from a neutral expression to one of the seven basic expressions. The videos are recorded at 30 frames per second (FPS).

Another widely used dataset for FER is the dataset developed by Jonathan Oheix from Kaggle [45]. It consists of about 35.9K images of size 48x48 pixels and is annotated into one of the seven basic emotions. The dataset has been organized into a training and validation dataset. For each image, the face region occupies about the same amount of space. Erdem et al. [47] created the audio-visual dataset named BAUM-2. It is an emotion detection dataset and consists of 1047 Video clips collected from 286 subjects. The data is annotated as one of the seven basic emotions. Kaur et al. [49] created a dataset for engagement prediction named EngageWild. It consists of 195 video clips, each 5 minutes long, totalling 16.25 hours, collected from 78 individuals, consisting of 25 females and 53 males. The data was collected while the subject watched an educational video about 5 minutes long. The data was annotated by 5 annotators without audio in 4 levels of intensity(0-3) of engagement, where intensity 0 meant disengaged, intensity 1 meant barely engaged, intensity 2 meant engaged, and intensity 3 meant highly engaged.

Selim et al. [50] introduced the VRESEE dataset. This dataset consists of 3525 video recordings of 88 students, with 53 males and 35 females. The videos are 10

seconds long and recorded during e-learning sessions. The subjects themselves annotated their engagement across four levels. Wang et al. [51] introduced the USTC-NVIE dataset. This dataset consists of 236 natural as well as infrared face images collected from 215 subjects, 157 males and 58 females. The annotation was done by five students. UPNA is a head pose estimation dataset introduced by Ariz et al. [53]. This dataset consists of 120 videos collected from 10 subjects, 6 female and 4 male. Translation and rotation about three axes are identified in this dataset.

Apart from datasets for emotion detection, some algorithms are pre-trained on the face detection dataset. The most commonly used dataset in pre-training is the VGGFace2[61] dataset, which consists of 3.31 million images of 9131 subjects. The images were collected using a Google search, and the images consist of various poses, ages, ethnicities, illuminations, and professions.

3.2 Detection of Face and ROI

The most common approach for ROI detection found in the literature [16, 18, 19, 24, 28, 30, 34, 41, 42, 46, 48, 57, 58, 62] is the use of the Viola-Jones algorithm[63]. It is the oldest and the most commonly used face detection algorithm, which uses *Haar-like* features to locate facial features. It can be used for both object detection and feature extraction. The algorithm was designed to detect front-facing faces. It was observed in [64] that the algorithm performance degraded for the side-facing faces. But implementation of this algorithm can still be seen today due to its efficiency on CPUs, low latency, and ease of implementation. Another commonly used algorithm is the Multi-Task Cascaded Convolutional Neural Networks or MTCNN[65]. It uses three neural networks in cascade. The three networks are the *Proposal network*, the *Refine network*, and the *Output network*. [17, 25, 31, 35] implemented this method. [24, 37, 60] implemented versions of You Only Look Once or *Yolo* [66] to detect face. The advantage of Yolov5 is that it can detect small objects and has a fast inference speed.

Das et al.[67] implemented *OpenFace* while Ma et al.[20] implemented *OpenFace2.0* [68] for face detection, tracking, and feature extraction. It is an extension of the OpenFace toolkit and can detect faces and head poses. It can also extract facial landmarks and track the target. Okur et al. [56] used a 3D camera(Intel® RealSense™ F200 Camera) and the raw image data was fed into the camera SDK(Intel® RealSense™ SDK) for getting face location, head position, and pose. A total of 78 facial landmark localizations, 22 facial expressions, and seven basic facial emotions were used in their study. A region-based convolutional neural network, known as Faster R-CNN[69] for face detection, was reported in [39]. In [33, 59] OpenCV was used for both face detection and tracking. [54, 70] implemented Histogram Oriented Graph(HOG) for face detection. The use of Dlib with 68 landmark prediction was observed in [26].

3.3 Tracking of ROI

Performing face detection for each frame of a video is a computationally expensive process for a real-time system. In [29, 55], the face detection process occurs after every few seconds. There are works [28, 62] which use Viola-Jones on each frame

for tracking. However, such methods suffer a drawback as the frames in between the time intervals are just ignored, even though they might contain information about changes in learners' affective state. Other methods [16, 19, 24] use dedicated face tracking algorithms like Deep Simple Online and Realtime Tracking(Deep SORT)[71] and Incremental Learning Visual Tracking(IVT)[72] for tracking the detected face frame to frame. [24] uses *Deep SORT*. The Deep SORT algorithm uses *Kalman* filter for object tracking. It can track multiple objects in a video stream. It has the advantage of handling situations in which the face may temporarily disappear or become obscured but it requires a GPU for real-time performance. IVT is a fast and lightweight particle-based filter that can handle the problem of changing target appearance even on a CPU. [16, 19] has implemented IVT for face tracking in their proposed methods.

3.4 Feature Extraction

In many works [16, 18, 24, 28, 46] Viola-Jones algorithm was used for feature extraction too. Liao et al. [17] used *Sequeeze-and-Excitation Network* or SENet(SE-ResNet-50)[73], which is pre-trained on VGGFace2[61] to convert facial images into facial spatial features. Dewan et al, [19] used *Local Directional Pattern(LDP)*[74] to extract person-independent edge features for the different facial expressions and *Kernel Principal Component Analysis(KPCA)* to capture the nonlinear correlations among the extracted features. An adaptive weight *Local Gray Code Patterns(LGCP)* and *Fast Sparse Representation* as a feature extraction method is reported in [55]. In [20], a Convolutional 3D(C3D) for body-pose and OpenFace2.0[68] for head-pose, gaze and action unit.Similarly, in [21], C3D was adopted for feature extraction. Savchenko et al. [25] proposed the use of *EfficientNet* to extract the emotional features. Standard Emotion Face image collected for each learner as features for emotion recognition was adopted in [57]. Gupta et al. [39] developed a modified face-point extractor(MFACEXTOR), which can extract 470 key points from a facial image.

3.5 Emotion Recognition and Engagement Detection

Human emotion can be classified in different ways. Those include 4, 6, 7, 8, and 4-level compound emotions, as reported in the literature. In Table 2, we reported these different levels along with their associated states. Except for the different levels of human emotions, there exist different levels of engagement, particularly in a classroom. From the literature, we identified three different levels of engagement as shown in Table 3.

Facial expression recognition or facial emotion recognition, abbreviated as **FER**, is one of the major steps in learners' engagement detection. Apart from emotion detection, FER algorithms are also used for behavior detection, like pose and action detection. These methods require pre-training using a relevant dataset before being implemented in the wild. It has been observed that both traditional machine learning, deep learning, and other techniques were used in FER. In the following, we precisely reported them.

Table 2 Different levels of emotions.

No of Levels	States
4 Level	Happy, Neutral, Confused and Sleepy/Bored
6 Level	Happy, Surprise, Anger, Sad, Fear and Neutral
7 Level	Happy, Surprise, Anger, Sad, Fear, Disgust and Neutral
8 Level	Happy, Surprise, Anger, Sad, Fear, Disgust, Contempt and Neutral
4 Level Compound	Confusion, Satisfaction, Disappointment and Frustrated

Table 3 Different levels of affective state.

No of Levels	States
2 Levels	Engaged, Not engaged/ Disengaged
3 Levels	Highly Engaged, Normally Engaged, Not engaged/ Disengaged
4 Levels	Highly Engaged, Engaged, Barely Engaged, Not engaged/ Disengaged

3.5.1 Traditional Machine Learning based Approaches

Traditional machine learning approaches have been widely used for engagement and emotion recognition, particularly through comparative evaluations of classical classifiers. Ayvaz et al. [58] and Gopinathan.N et al. [70] compared algorithms such as KNN, Random Forest, CART, and SVM. In [58], models trained on a self-curated dataset classified six emotions, later grouped into positive (happiness, surprise) and negative (sadness, fear, disgust, anger) categories for motivation assessment, where SVM achieved the highest accuracy of 98.24% and Random Forest the lowest at 92.97%. Similarly, [70] used the WACV dataset to predict two- and three-level engagement, with Random Forest performing best for binary classification (90%) and SVM achieving the highest accuracy (57%) for three-level classification. These results highlight the effectiveness of classical classifiers, particularly SVM and Random Forest.

Some works extend beyond standalone classification by incorporating contextual and visual cues. Okur et al. [56] proposed a two-phase framework where learner engagement is first inferred from active URLs (on-platform vs. off-task), followed by appearance-based analysis using head pose, facial expressions, and basic emotions. A Random Forest classifier is then used for final classification, achieving 77% accuracy, demonstrating the benefit of combining contextual and visual information.

Another direction involves multimodal frameworks integrating deep learning with traditional machine learning. Altuwairqi et al. [29] combined a mini-Xception-based CNN [75] for emotion detection with behavioral log data (mouse and keyboard activity). The fused features were classified using Naive Bayes into three engagement levels, achieving 95.23% accuracy and an MSE of 0.04. Similar hybrid approaches were also explored in [22, 67], highlighting the effectiveness of combining deep feature extraction with lightweight classifiers.

Additionally, several studies employ traditional machine learning directly, with SVM used in [54, 76] and Random Forest in [77], reinforcing their relevance in scenarios with limited data or computational resources.

3.5.2 Deep Learning based Approaches

Deep learning-based approaches are the most widely used for emotion recognition and engagement detection, with CNNs dominating due to their superior performance on complex tasks despite higher computational cost. Variants of CNN, along with some RNN and ANN-based methods, have been explored.

Custom CNN architectures have been proposed by Murshed et al. [16] and Dash et al. [18], combining features from models such as *All-CNN, NiN-CNN, VD-CNN*, and *CPConv*. The former achieved 92.33% accuracy for three-level engagement classification, while the latter employed a two-stage pipeline (on/off-task detection followed by engagement classification) with an average accuracy of 91.1%.

Multi-stage CNN frameworks were introduced by Bhardwaj et al. [30] and Sharma et al. [28], where one CNN detects focus via eye gaze and a second performs emotion recognition. Bhardwaj et al. used a basic CNN for seven emotions, while Sharma et al. adopted mini-Xception [75]. Engagement was computed using a Mean Engagement Score (MES), achieving accuracies of 93.6% and 66%, respectively. Similarly, Mukhopadhyay et al. [32] used a CNN over temporally sampled frames to classify combined emotional states into four categories, achieving 65% accuracy. Hasnine et al. [33] combined mini-Xception [75] with eye-region detection [78] and a Concentration Index (CI) for engagement estimation, achieving 68% emotion recognition accuracy.

ResNet-based approaches have also been widely explored. Abedi et al. [21] and Liao et al. [17] used ResNet for feature extraction, followed by Temporal Convolutional Networks (TCN) and LSTM with global attention (GALN), respectively, achieving 63.9% and 58.84% accuracy for multi-level engagement classification. In contrast, Zeng et al. [31] and Gupta et al. [39] used ResNet-50 directly for emotion recognition. Zeng et al. also incorporated FaceNet [79] for identity tracking, while Gupta et al. compared ResNet-50 with Inception-V3 [80] and VGG-19 [81], achieving 71.2% and 73.4% accuracy, respectively. Similarly, Karan et al. [59] fine-tuned VGG-16 [81] for four-class emotion detection and derived engagement via an index, achieving 77.5%.

EfficientNet-based methods were explored by Savchenko et al. [25], who used EfficientNet-B0 and B2 for emotion recognition and statistical (STAT) feature aggregation for group-level engagement estimation, achieving 63% (8 classes) and 66% (7 classes). In contrast, Ahmed et al. [62] focused on eye gaze using MobileNet [82], classifying engagement based on gaze direction and achieving 99.37% accuracy.

Object detection-based approaches using YOLO have also been investigated. Trabelsi et al. [24] compared multiple YOLOv5 [66] variants, combining behavioral and emotional cues and addressing occlusion using MaskTheFace [83], achieving 76% accuracy. Similarly, [37] applied YOLOv5 on FER2013 and CK+ datasets, achieving 72% and 98.4% accuracy respectively. Ferreira et al. [60] used YOLOv8 for emotion detection and ResNet50 with SVM for engagement classification, reporting 94.5% precision and 92.3% recall.

Beyond CNNs, alternative deep learning models have been explored. Yang et al. [41] and Sidhu et al. [46] used ANN-based methods for seven-class emotion recognition, achieving 87.25% and 67% accuracy, respectively. Dewan et al. [19] proposed a Deep Belief Network (DBN) [84] for engagement detection on the DAiSEE dataset, achieving 90.89% (two-level) and 87.25% (three-level) accuracy. RNN-based approaches were

explored by Jagadeesh et al. [48], who proposed a heuristically optimized HM-RNN using C-DHOA for emotion recognition (92.3% accuracy), and Ma et al. [20], who introduced a Neural Turing Machine (NTM)-based multimodal framework incorporating gaze, facial, and body features for four-level engagement classification (61.3% accuracy). Additional CNN-based methods were also reported in [44, 50, 52, 85–89].

3.5.3 Other Approaches

Few pieces of literature were found that do not use any form of machine learning in their method for engagement detection, but implement other conventional approaches. Zhang et al.[55] implemented a multimodal approach. Their method used facial data and mouse movement data for engagement detection. The model used adaptive weighted Local Gray Code Patterns(LGCP)[90] for FER trained on a self-built dataset. Then the extracted features are classified using a fast sparse representation into two engagement states. The proposed method used two self-made datasets with the same data, but the first dataset referred to mouse movement during annotation. An accuracy value of 94.60% was observed on the first dataset that referred to mouse movement data, while 91.51% accuracy was observed on the second dataset that did not refer to mouse movement data.

Panichkriangkrai et al.[57] proposed a method to recognize the emotion level of a learner. They used Local Binary Patterns Histogram(LBPH)[91] trained on the collected standard emotion faces for each learner to classify into one of the four emotions. Each of the emotion levels was assigned a state, where Sleepy/Bored was given 1, Confused as 2, Neutral as 3, and Happy as 4. Then the interest level of each learner is assessed. Learners in states 3 and 4 were considered as positive interest levels, while learners in states 1 and 2 were considered inattentive or inactive. Learners in state 2 are unable to understand the lessons that can lead to state 1. Their method achieved an emotion classification accuracy of 88.7%. Implementation of LBPH was also observed in methods proposed by [92, 93].

Table 4 summarizes the literature reviewed in this section. It highlights the key methods used for classification, face detection, dataset use for training the classifier and the result reported in the respective existing work.

Table 4: Literature review

#	Method	Face detector	Dataset	Accuracy %
[58]	ML: KNN, Random Forest, CART, SVM	Viola-Jones	own	98.24
[56]	ML: Random Forest	Intel RealSense SDK	own	77
[76]	ML: SVM, MLR, Boost	Viola-Jones	own	72.9
[54]	ML: SVM	HOG	UPNA	72.4
[77]	ML: Random Forest	n/a	own	86
[70]	ML: Random Forest, SVM	HOG+SVM	WACV	2 Level- 90 3 Level- 57

Table 4 continued from previous page

#	Method	Face detector	Dataset	Accuracy %
[29]	CNN+ ML: MiniXception + Naive Bayes	Cosine similarity	FER2013	95.23
[22]	CNN+ML: CNN + KNN	n/a	FER2013, DAiSEE	Emotion- 60.57 Engagement- 68.57
[67]	CNN,ML: EfficientNet, XGBoost	OpenFace, MediaPipe	WACV	ML- 82.90 CNN- 47.2
[16]	CNN	Viola-Jones	DAiSEE	92.33
[18]	CNN: Cm-CNN	Viola-Jones	DAiSEE	Behaviour- 92.5 Emotion- 83.4
[30]	CNN	Viola-Jones	FER2013	93.6
[28]	CNN: CNN and miniXception	Viola-Jones	FER2013	66
[32]	CNN	n/a	FER2013	65.00
[33]	CNN: Residual module + Depth wise Convolution	OpenCV	FER2013	68
[21]	CNN: TCN	n/a	DAiSEE	63.9
[17]	CNN: SENet + GALN	MTCNN	DAiSEE	58.84
[31]	CNN: ResNet-50	MTCNN	FER2013	71.2
[39]	CNN: Inception V3, VGG-19, ResNet-50	Faster RCNN	own	73.34
[59]	CNN: VGG-16	n/a	own	77.5
[25]	CNN: EfficientNet-B0 and B2	MTCNN	AffectNet, EngageWild	7 Class- 66 8 Class-63
[62]	CNN: MobileNet	Viola-Jones	Eye Dataset-Kaggle	99.37
[24]	CNN: YOLOv5	YOLOv5	Own + AffectNet	76
[50]	CNN:EffecientNet-B7 + TCN/ LSTM/Bi-LSTM	n/a	DAiSEE + VRESEE	94.47
[44]	CNN: VGG-19, ResNet-50	Faster R-CNN	FER-2013, CK+, own	92.58
[52]	CNN: VGG16, Inception-ReNet-V2	n/a	ImageNet, NVIE	76.08
[86]	CNN	n/a	own	90.47
[87]	CNN + LSTM	n/a	EmotiW 2018 Challenge	n/a
[88]	CNN: MobileNet V2	n/a	ImageNet	74.55
[89]	CNN: Vgg-16 + ResNet-101	n/a	wacv2016	62.03
[34]	CNN: mobileNet-V2	Viola-Jones	FER2013	73.40
[35]	CNN	MTCNN	FER2013	73

Table 4 continued from previous page

#	Method	Face detector	Dataset	Accuracy %
[36]	CNN	n/a	FER2013	93
[37]	CNN: Yolo-v5	Yolo-v5	FER2013, CK+, own	FER2013- 72 CK+- 94.8
[26]	CNN: EmoNet + LSTM	Dlib	AffectNet	70
[60]	CNN: Yolo-v8, ResNet50 +SVM	Yolo-v8	own	n/a
[42]	CNN: Vgg-16	Viola-Jones	JAFEE, KDEF, CK+	95
[41]	ANN	Viola-Jones	JAFFE	87.25
[46]	ANN	Viola-Jones	Facial expression recognition- Jonathan Oheix	67
[19]	DBN	Viola-Jones	DAiSEE	2 level- 90.89 3 level-87.25
[48]	RNN: HM-RNN	Viola-Jones	BAUM-2	92.3
[20]	NTM	OpenFace2.0	DAiSEE	61.3
[55]	Other: LGCP + FSR	n/a	own	94.6
[57]	Other: LBPH	Viola-Jones	own	88.7

4 Research Methodology

In this section, we present the techniques implemented to source the literature as well as the research questions formed from analyzing the selected literature to carry out the proposed literature review. For sourcing the literature, firstly, we select the database that contains relevant scholarly papers. There are different databases like ACM Digital Library, ScienceDirect, IEEE Xplore, and Springer. Another source to obtain scholarly papers is through the Google Scholar search engine, which searches through all the different databases to generate a cumulative result of related literature.

Searching the database with a single keyword returns thousands of results. So based on the area of our research, we decided to form the following search string as given in table 5 to filter the results published between 2015 and 2025. The result obtained while searching using the search string in Google Scholar is as given in Table 6. It is to be taken care that the search string does not overfit, as it may result in the exclusion of relevant papers. Moreover, to ensure that the filtered results have only the relevant and useful papers, we set an inclusion and exclusion criterion for the papers obtained using the search string as given in table 7.

Implementing the search criteria and inclusion-exclusion criteria, we obtained the literature for review from different databases. The steps are as shown in Figure 4 in the form of a flow diagram. In the identification section, the number of results obtained by using the first search string S_1 before implementing any form of filtering criteria is shown. Initially, 2480 records were found. Then, 190 records that were published

Table 5 Search string formed for filtering literature.

Sl.No	Search String
S_1	"engagement detection"
S_2	"engagement detection" & "student or learner"
S_3	"engagement detection" & "student or learner" & "online or e-learning"
S_4	"engagement detection" & "student or learner" & "online or e-learning" & "computer vision" & "emotion"

Table 6 Google search findings.

Search String	No of articles
S_1	2290
S_2	1770
S_3	1640
S_4	791

Table 7 Inclusion and Exclusion criteria

No.	Inclusion criteria	Exclusion criteria
IE_1	Papers published between 2013 and 2025	Papers published before 2013.
IE_2	Engagement detection in e-learning	Engagement detection not related to student learning.
IE_3	Engagement detection using computer vision as a primary method.	Engagement detection using the non-automatic method as a primary method.
IE_4	Papers published in the English language	Papers published in languages other than English.
IE_5	Papers related to engineering and computer science	Papers related to other disciplines like medical, social, life science, and psychology.
IE_6	Realtime monitoring.	Not real-time monitoring.

before 2013 were removed. Next in the screening section, from the remaining 2290 records, 1499 records were excluded based on the search strings formed, resulting in 791 records being selected for further screening. Again, 740 records were excluded from the remaining 791 records based on the inclusion-exclusion criteria. Finally, in the included section, we obtained 51 literature for our review.

Based on a preliminary analysis of the literature for learners' engagement detection in online or e-learning, we constructed the following research questions as given in table 8. These research questions were constructed to understand the purpose and process of engagement detection techniques in detail. The list of journals and conferences from where relevant literature was acquired is given in Table 9 and Table 10, respectively.

5 Result

In this section, we present the answers to our research questions (R_1, $R_2 \ldots R_8$) as shown in table 8. These answers were produced by analyzing the articles selected for review. Our findings are as follows.

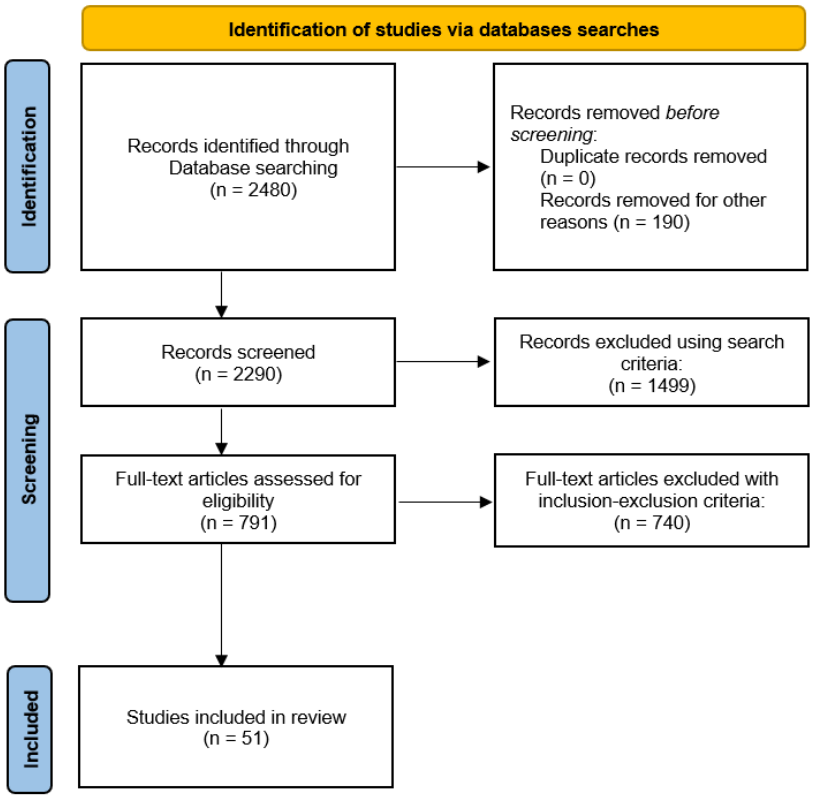


Fig. 4 PRISMA 2020 flow diagram for identifying, screening, and including literature for review.

Table 8 Research questions developed by us.

No	Research Questions
R_1	What are the different methods of engagement detection?
R_2	Is engagement a learner's problem or a teacher's problem?
R_3	How do humans communicate Emotions?
R_4	How emotions affect engagement?
R_5	Is occlusion a hindrance in engagement detection?
R_6	What are the different types of FER methods?
R_7	What other feature can be considered other than facial information?
R_8	How to consider a useful dataset?
R_9	How different teaching styles and learning environments affect engagement?

R_1: What are different methods of engagement detection?

As outlined in [9], engagement detection methods can be broadly categorized into manual, semi-automatic, and automatic approaches based on the level of learner involvement. While this taxonomy provides a useful structural overview, it does not fully capture the practical limitations and deployment challenges of each category, particularly in scalable online learning environments.

Table 9 List of Journals.

No	Journal: year	Total
1	IEEE Transactions on Affective Computing, 2014, 2016, 2022	3
2	IEEE Access, 2019, 2022, 2024	3
3	Multimedia Tools and Applications, Springer, 2022, 2023	2
4	Multimedia Systems, Springers, 2022, 2023	2
5	International Journal of Information and Education Technology, IJIET, 2021	1
6	Computers & Electrical Engineering, Elsevier, 2021	1
7	Applied Intelligence, Springer, 2021	1
8	Big Data and Cognitive Computing, MDPI, 2023	1
9	Science, Engineering and Health Studies, Silpakorn University, 2021	1
10	IEEE Transactions on Visualization and Computer Graphics, 2021	1
11	Signal, Image and Video Processing, Springer, 2021	1
12	Journal of Educational Computing Research, Sage Journal, 2019	1
13	Information Technologies and Learning Tools, Mugla Sitki Kocman University, 2017	1
14	Smart Learning Environments, Springers, 2019	1
15	Proceedings of the IEEE, 2023	1
16	Electronics, MDPI, 2022	1
17	Journal of King Saud University – Computer and Information Sciences, Elsevier, 2018	1
18	arXiv.org, 2024	1
19	Scientific Programming, Wiley, 2024	1
20	Traitement du Signal, International information and Engineering technology Association, 2025	1
21	Measurement Science and Technology, IOP, 2025	1
22	Neural Computing and Applications, Springer, 2025	1
23	Computers, MDPI, 2025	1
24	Social Network Analysis and Mining, Springer, 2025	1
25	Symmetry, MDPI, 2025	1
	Total	**31**

Manual methods, including self-reporting and observational checklists, rely heavily on explicit human input, making them inherently subjective and unsuitable for continuous monitoring. Although they can provide valuable insights into learners' perceived engagement, their lack of objectivity, susceptibility to bias, and inability to operate in real-time significantly limit their applicability in modern e-learning systems. Moreover, observational methods often measure compliance rather than true cognitive engagement, raising concerns about their validity.

Semi-automatic methods, such as engagement tracing based on response time and answer accuracy, reduce direct user involvement but still rely on indirect representations of engagement. While effective in controlled environments like intelligent tutoring systems, these methods assume a strong correlation between performance and engagement, which may not always hold. E.g., a learner may be engaged but struggling, or disengaged yet answering correctly. Additionally, their reliance on task-based interaction limits their applicability in passive learning scenarios and constrains real-time adaptability.

Automatic methods address many of these limitations by enabling continuous, real-time, and unobtrusive monitoring using sensor data. These include log file analysis, physiological sensing, and computer vision-based approaches. However, log-based methods may fail to capture internal cognitive states, while physiological methods

Table 10 List of Conferences.

No	Conference: year	Total
1	International Conference on Image Processing, IEEE, 2019	2
2	Technology and Innovation in Learning, Teaching and Education, Springers, 2022	1
3	International Conference on Sustainable Technologies for Industry 4.0 (STI), IEEE,2019	1
4	IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI),2018	1
5	International Conference on Innovative Computing and Communications, Springers, 2021	1
6	18th Conference on Robots and Vision (CRV), IEEE, 2021	1
7	Artificial Intelligence in Education, Springer, 2017	1
8	Procedia Computer Science, 6th International Conference on Smart Computing and Communications, ICSCC, Elsevier, 2017	1
9	Proceedings of the 5th International Conference on Intelligent Information Technology, ACM, 2020	1
10	Procedia Computer Science, 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems,Elsevier,2021	1
11	Innovations in Computational Intelligence and Computer Vision, Springers, 2022	1
12	IEEE Delhi Section Conference (DELCON),2022	1
13	International Conference on Contemporary Information Technology and Mathematics, IEEE,2022	1
14	IEEE Global Engineering Education Conference, 2023	1
15	International Conference on Advanced Intelligent Systems for Sustainable Development, Springers, 2023	1
16	Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress, IEEE, 2019	1
17	International Conference on Development and Learning, IEEE, 2022	1
18	International Conference on Emerging Smart Computing and Informatics, IEEE, 2023	1
19	9th International Conference for Convergence of Technology (I2CT), IEEE, 2024	1
	Total	**20**

(e.g., EEG, heart rate) offer higher fidelity but require specialized hardware and intrusive setups, limiting scalability. In contrast, computer vision-based methods provide a practical trade-off between accuracy and deployability, leveraging widely available imaging devices while capturing rich behavioral cues such as facial expressions, gaze, and posture. Nevertheless, they are still challenged by occlusion, illumination variability, and privacy concerns.

Overall, while the three-category taxonomy is conceptually sound, it overlooks important distinctions in scalability, intrusiveness, and reliability. Among these, automatic methods, particularly computer vision-based approaches emerge as the most suitable for real-world online learning systems, motivating their focus in this work.

R_2: Is engagement a teacher's problem or learner's problem?

Engagement in learning can be seen as a shared responsibility of both teacher and learner, where each side contributes to their own extent [94, 95]. At the same time, some studies believe there exists asymmetry in the roles with the teacher being the key driver of engagement [96, 97]. The primary role of the teacher is to create a suitable learning environment that caters to the collective requirements of the learners.

While for the learner, it is necessary to bring themself to participate in that learning process. The teacher's teaching style, follow-up, approach, quality of teaching, and classroom ambiance strongly influence the engagement of the learner [98]. For effective learning, a teacher must keep track of their needs and adapt accordingly to maintain a good level of engagement [24]. A teacher should also understand that students who struggle to keep up with the pace of their peers in academics or lack of feedback and interaction from the teacher can develop a sense of isolation, which can contribute to disengagement from learning [3, 99, 100]. It is a challenge for a teacher to keep track of every student's engagement level in a crowded classroom [24]. From the learners' side, their emotional state highly influences the engagement level [6]. Negative emotions like anger, disgust, and sadness can distract the learners from learning activities. Lack of intrinsic motivation to learn can also lead to disengagement. Apart from that, another reason for disengagement is computer vision syndrome. Prolonged staring at the digital screen for consecutive classes can put a strain on the learners' eyes, which causes fatigue and headaches, which then results in disengagement.

The actions of either the teacher or the learner, or both, can result in the engagement or disengagement of a learner from the learning activity. It can be concluded that it is a shared challenge where the teacher drives to sets the stage for engagement, but learners must mediate and take responsibility for their participation.

R_3: How humans communicate Emotions?

Humans express their emotions in verbal and non-verbal forms[101]. Verbal communication of emotion involves speech. A person's laughter represents positive emotions like happiness and excitement, while the extent of sobbing infers the level of sadness. Apart from that, the tone of voice, like pitch, speed, volume, and rhythm of speech, also conveys the person's emotional state. For example, a raised voice can indicate anger. Also, language and word choice convey a lot about the person's emotions. For example, using negative or positive words can reflect one's mood. It has been found that verbal communication expresses one-third of emotion, while nonverbal communication expresses the rest, as reported in [102, 103]. Non-verbal communication includes facial expression or emotion, body posture or action, and behavior. For facial emotion, one or more facial muscle position and their movements are considered. Facial regions like the eyes and mouth are the most expressive regions of human emotions. In fact, emotion detection using computer vision uses facial Action Units(AUs) generated by observing the facial muscle movement.

R_4: How emotions affect engagement?

Emotions have a substantial influence on a learner's cognitive ability; they can either improve or degrade our affective state of engagement [58]. Researchers have found that learners who experience positive emotions during the learning activity show faster learning and increased intellectual ability. Emotions have been found to directly impact attention and cognitive processing in learning environments. Positive emotions can help students focus on the task at hand and engage in deeper cognitive processing. Negative emotions make the learner distracted and uninterested in the learning

process. Negative emotions can keep the mind preoccupied and decrease the motivation to learn. Learners experiencing negative emotions may disengage, avoid academic challenges, or perform poorly due to stress [104].

Emotional states can influence how students engage with their peers in collaborative learning environments. Positive emotions, such as empathy or enjoyment, foster positive social interactions and encourage cooperative learning, leading to greater engagement and more productive group work [105]. Negative emotions, such as anger, frustration, or shame, can create conflict, reduce collaboration, and limit engagement in group activities. These insights highlight that the maintenance of positive emotional states, as a shared responsibility of both educators and learners, is critical for optimizing engagement and collaborative learning outcomes.

R_5: Is occlusion a hindrance in engagement detection?

FER techniques require the extraction of features from the whole face region to efficiently decide the current emotion of the subject. Occlusion, whether full or partial, can greatly reduce the accuracy of the FER algorithm as it prevents the extraction of the features from every region. Most of the literature reported their methods to underperform on subjects with occlusions. Many methods have adopted a work for partial occlusion like wearing a face mask, face covered with a hand, etc. [24] proposed to apply a facial mask on the existing dataset using MaskTheFace and train the classifier on the new masked faces and existing annotation, but the performance of the method on wild occlusions is unknown. [17] proposed the use of CNN to extract facial spatial features that are robust to occlusion, but these methods require significantly larger computation time compared to traditional machine learning algorithms. [106] has reported an average drop of approximately 3% and 3.5% in FER accuracy on FERPlus and RAF-DB datasets, respectively, when comparing occluded and non-occluded samples across different methods. So, it can be said that occlusion is still a problem for real-time engagement detection.

R_6: What are the different types of FER methods?

FER methods can be categorized in several ways. Based on the type of input, the FER algorithm can be divided into two groups: static FER(frame-based) and Dynamic FER(video-based)[101, 107]. In static FER, a single image frame is processed for emotion recognition at once, while in dynamic FER, the change in facial expression in a video is tracked over time. Again, based on how the facial information is analyzed, FER methods can be classified into part-based and appearance-based [93, 101].In the part-based approach, only certain regions of the face, like eyes, nose, and mouth, are considered for emotion recognition, while in the appearance-based approach, the global facial region is considered. Again, based on the type of algorithm, a FER method can be a Traditional Image Processing algorithm or a Machine Learning algorithm. Based on modality, a FER method can be Single-Modal or Multi-Modal [29, 55]. A single-modal technique uses only a single type of data, like image or sensor data log-file data, while a multi-modal technique uses a combination of more than one type of data, like facial and mouse movement data, or facial and speech data, or facial, keyboard stroke, and mouse movement data.

R_7: What other feature can be considered other than facial information for a computer vision-based method?

Computer vision-based methods mainly consider facial expression and eye movement for engagement detection. Apart from facial features, body posture, action, and behavior have also been considered in the existing literature as they convey a great deal of information about engagement. Actions and behaviors like raising a hand, writing notes, and proper sitting pose are considered positive factors for engagement, while facing away from the screen, talking, eating, or drinking are considered negative factors contributing to disengagement [24]. Many methods first check for engagement using these factors before analyzing the facial expression or emotion of the learner. Only if the body posture or action and behavior show positive factors for engagement, are the facial expressions or emotions analyzed for final engagement detection. Literature [29, 44, 52, 55, 86, 93] has proposed the use of additional parameters to be used along with computer vision-based methods, such as mouse clicks and movements, and keyboard strokes. These parameters, though they have low accuracy of engagement detection on their own, have been found to increase the accuracy of engagement detection when used together with FER.

R_8:How to consider a useful dataset?

A useful dataset for emotion or engagement detection should have a large sample size, diverse subjects, reliable multi-annotator labeling, balanced classes, and preferably spontaneous expressions. However, existing datasets only partially satisfy these criteria.

Among the reviewed datasets, DAiSEE and AffectNet best meet these requirements. DAiSEE provides spontaneous video data in e-learning contexts with multi-level annotations for affective states, making it highly suitable for engagement detection. AffectNet offers large-scale diversity and valence-arousal annotations, though it may include noisy samples due to internet-based collection. Datasets such as RAF-DB and FER2013 provide sufficient scale for training but lack task-specific engagement labels, and FER2013 additionally suffers from low resolution. In contrast, datasets like JAFFE and CK+ are well-annotated but contain posed expressions and limited subjects, restricting real-world applicability.

Engagement-specific datasets such as EngageWild and VRESEE provide relevant annotations but are comparatively smaller. Other datasets such as BAUM-2, USTC-NVIE, and UPNA are designed for specialized tasks (e.g., multimodal emotion, infrared imaging, and head pose) and are less directly applicable to engagement detection. Overall, no dataset satisfies all desirable properties, highlighting a trade-off between scale, realism, and task relevance, and indicating the need for more comprehensive datasets tailored to engagement detection.

R_9:How different teaching styles and learning environments affect engagement?

Teaching styles and learning environments play significant roles in shaping student engagement in online learning. Traditional teaching styles, where the teacher controls the pace, content, and direction of the lesson, can hinder student engagement in online settings as students may feel passive and detached. To counter this, teachers

should incorporate interactive elements like discussions and quizzes that require active participation. Active learning methods that are student-centered, such as project-based learning, case studies, and collaborative group activities, are more effective in engaging online learners. When students take responsibility for their learning, they are more likely to participate and stay engaged [3, 99]. Teachers' feedback and interaction are critical for engagement. Regular feedback and communication (via email, discussion boards, video messages, etc.) help maintain student motivation. Research shows that timely, constructive feedback positively impacts student engagement and achievement. Clear guidelines and expectations from the teacher also impact student engagement. If students know what is expected, they are more likely to stay motivated and engaged.

Different learning environments, like synchronous and asynchronous modes, facilitate the learners to learn with convenience. Synchronous mode facilitates real-time interaction and feedback from the teachers, thus promoting personal participation, while asynchronous mode allows students to access materials at their convenience, which facilitates cognitive participation as students can process the information at their own pace[3].

6 Discussion

This literature study reviews computer vision-based learners' engagement detection methods developed for e-learning environments during the past decade. We have collected research papers from multiple sources, as mentioned in Section III. Our study analyzed the existing methods for engagement detection by dividing the process into four major steps:

1. Detection of ROI.
2. Tracking of detected ROI.
3. Feature extraction.
4. Facial emotion recognition and engagement detection.

The literature was selected for review using the proposed inclusion and exclusion criteria as given in Table 7. This review is limited only to papers where computer vision-based methods are the primary methods, as they are the only methods that can be implemented in real-time using only already available hardware, like computers and web cameras. Some of the literature consists of analyzing additional parameters(multimodal approach), like mouse and keyboard data, context identification, along with emotion recognition.

Our work attempts to understand the techniques for engagement detection implemented for e-learning setups. To analyze the existing works in the field, we formed nine research questions (R_1, R_2,.....R_9) as given in Table 8. It has been observed that many of the recently developed methods of engagement detection are based on Machine Learning algorithms, with CNN models being used the most, while some methods also implement vanilla neural networks and traditional machine learning algorithms. Some literature also adopted the use of traditional image-processing algorithms for emotion recognition. While a clear superior category of algorithms cannot be ascertained, as

each category produced an accuracy score spread over a range, it was observed that methods [29, 55]that implement a multi-modal approach generated better results than the single-modal approach using the same algorithm.

It has been observed that there is no precise rule set to decide which emotion represents which particular level of engagement, but rather the engagement level is decided based on the inferred perception of the author about that particular emotion. This is both a challenge and an advantage. The challenge is that it becomes complicated to perform comparative analysis among different methods, and the advantage is in the cultural context. Different cultures have slight variations in their ways of expressing certain emotions [108, 109]. For example, some East Asian cultures, like Japanese, tend to value emotional restraint and harmony within the group, thus avoiding public display of strong emotions, while public emotional displays are often seen as a healthy means of communication in Western cultures. But since the basic emotions are universal and common to all cultures, and so just fine-tuning the method to suit emotional intensities of different cultures, the methods could be universally applicable. Then again, some cultures refrain from displaying emotions like anger and sadness in public and disguise them with other emotional expressions. In such cases, analysis of micro-expressions that cannot be hidden using the Facial Action Coding System may be more suitable for predicting the actual emotion. Although there exist many studies on cultural differences, FER methods that consider variation in emotion expression are still underexplored.

Existing studies often rely on heuristic associations, which may vary across datasets and contexts. However, structured mappings can still be defined by considering emotional dimensions such as valence and arousal. In such formulations, high-arousal emotions (e.g., happiness, surprise, or even confusion) are generally indicative of higher engagement, whereas low-arousal states (e.g., boredom or neutrality) are more closely associated with disengagement. Additionally, certain emotions, such as confusion or frustration, remain inherently ambiguous and should not be forced into binary categories. To address this, we recommend that a 3-level engagement state be adopted in the form of engaged, nominally engaged, and disengaged. The nominally engaged shall represent the state of confusion and frustration, as they may reflect ongoing cognitive processing but also indicate a risk of disengagement. This intermediate state highlights scenarios where timely instructional intervention may be necessary to prevent a decline in engagement.

This work also analyzed the datasets used in the existing methods for emotion and affective state classification. Each dataset used in the existing methods has a fair number of samples, either as an image or a video, collected from multiple human subjects. It has been observed that the conditions at which the samples for facial expressions were collected can be categorized into posed, spontaneous, and wild. The posed samples are the samples collected in a controlled environment with an acted expression for each said expression or emotion. The spontaneous samples were collected as naturally occurring facial expressions and emotions collected while the subjects were in a controlled environment, such as watching educational videos or through webcam during online classes. The third category of samples, i.e., the wild samples, is the samples collected from either uncontrolled or unknown environments. These

samples are mostly collected through internet searches using specific keywords. The spontaneous and wild samples have an advantage over posed samples as they cover wider variations in each class of facial expression or emotion, which helps in better classifications of emotions at different intensities.

This study has identified several research gaps that can be worked on to improve the effectiveness of the engagement detection algorithm. Though much literature presented high accuracy scores for their methods, almost every literature review mentioned the problem of occlusion. It has been reported that occlusion, both partial and complete, greatly affects the face detection and emotion recognition accuracy of the existing methods. Some literature reported failure to detect the face even for partial occlusion, while many observed a decrease in classification accuracy. Trabelsi et al.[24] implement the MaskTheFace algorithm to introduce face mask-based occlusion into the dataset to train the classifier and achieved an accuracy of 76%. It is unknown if the result will be affected by encountering a wild occlusion. Among existing algorithms for face detection used in the literature reviewed here, it has been observed that Viola-Jones is the most commonly used algorithm, even though it works only with full-view front-facing faces, which is a major drawback for processing images with occlusion and side-facing faces for face detection and feature extraction.

Another research gap identified from the analysis of the literature is the comparative analysis of the algorithms. Since the methods are needed to produce the result in real time, the process is time-sensitive. It is required that the algorithms take the least amount of time for processing, which can be accepted as real-time, while maintaining a decent score of accuracy. Much of the literature claims their implemented algorithms are better among the selected few algorithms without providing a means for reproducing the said comparison result. Also being a time-sensitive process, most of the literature does not include any information on the average throughput time, i.e., the total time taken from step 1 to producing the result after completion of step 4, of their proposed methods, while some literature mentions their throughput time in hundreds of seconds, which contradicts the concept of real-time. This is a major gap in research, as actual analysis about being real-time applicable cannot be ascertained. It was observed from the analysis of the existing works that the classification accuracy is spread over a wide range, i.e., 47%-99% for different methods. The reason for such a large range of results can depend on many factors, like task complexity, training data, feature engineering, and hyperparameter tuning. It can be seen in [67] that traditional ML achieved 82.9% accuracy, but the CNN-based method only scored 47.2%. In their proposed method, the CNN pretrained on ImageNet for object classification was not finetuned for task-specific AU-based affective state classification except for the Fully Connected layers. Thus, while Traditional ML methods like XGBoost performed well on extracted multimodal data, CNN based like EfficientNet were unable to perform due to untuned hyperparameters. Also, it can be seen in [62] that their method generated a very high accuracy of 99.37%, but only eye gaze direction was considered for engagement detection. The drawback of such a method is that it is not possible to detect if the learner is actually engaged or just compliant, as being compliant does not necessarily mean the learner is interested or engaged in the learning activity.

A clear distinction must be made between real-time engagement detection systems and post-session analysis systems, as their design objectives differ significantly. Real-time systems are intended to provide immediate feedback or alerts to instructors during a live session, and therefore prioritize low latency and computational efficiency, often operating under strict time constraints with potentially reduced model complexity. In contrast, post-session systems analyze recorded data to generate summary reports of learner engagement, where accuracy and robustness are more critical than processing speed. Such systems can leverage more complex models and temporal analysis since they are not constrained by real-time requirements. Recognizing this distinction is essential when evaluating existing methods, as performance metrics such as accuracy and inference time should be interpreted relative to the intended application scenario.

Ultimately, the question of the effect of long-term monitoring of student engagement, such as how to adapt to changes in students' emotional states and how to balance privacy and monitoring needs arises. There is a difference in how children and adults express their emotions. While the former is more direct in displaying their emotions, the later shows more maturity and complexity in displaying their actual emotions. As mentioned earlier, the universalisms of the basic emotions reduce the complexity of predicting the emotion. Also, rather than only relying on the visual cues of each individual, the final judgment of engagement should be based on a multimodal approach, which might include log file analysis to refer to the performance log of the individual. By continuously tracking emotional engagement (such as frustration, boredom, or interest), educators can tailor their approach, content, and delivery style to suit the needs of the learners. Schools or educational institutions can gather large sets of data that show patterns in student engagement, which can then inform broader decisions in curriculum development, teaching methods, and even institutional policies. The drawbacks of long-term monitoring are that the teachers might become overly dependent on technology and FER data, potentially reducing the focus on human interactions or nuanced understanding of students' needs that might not be captured through facial recognition alone. The constant monitoring of emotional responses could lead to anxiety or stress among students. Knowing they are being constantly evaluated could affect their behavior, potentially leading to unnatural reactions.

Monitoring requires diverse visual facial data of the learner. Storing such data can pose a threat to the privacy of the individuals; thus, to maintain the ethicality of the monitoring process, it is advisable to process the raw data as it arrives and store only the computed information, like frequency of different emotional states and level of engagement, which could help monitor the progress of the learner, rather than storing the raw facial data. Data collection should be limited to specific times or contexts, such as during certain class activities, assessments, or interactions where engagement is critical. Furthermore, the use of edge AI, where the facial data will be preprocessed in the user's device, can mitigate privacy concerns by minimizing the transmission of sensitive data. Informed consent should be obtained before any monitoring takes place, and the purpose of data collection and how it will be used should be clearly informed to the students(and their parents, if applicable). Also, FER data should be used for constructive insights rather than as the sole basis for judgment. Regular audits of FER

systems must be conducted to ensure they are being used ethically and in compliance with privacy policies.

7 Conclusion

This paper presented a review of computer vision-based learner engagement detection methods for e-learning environments over the past decade. By organizing the analysis into four stages—ROI detection, tracking, feature extraction, and engagement inference—we provided a structured understanding of existing approaches. The review indicates that deep learning, particularly CNN-based models, is most commonly used, although traditional machine learning remains effective in certain scenarios, especially with well-engineered or multimodal features.

A key finding is the lack of standardized emotion-to-engagement mapping, with most methods relying on heuristic or dataset-specific interpretations, limiting comparability across studies. Additionally, multimodal approaches generally achieve better performance, while challenges such as occlusion, dataset bias, and real-time constraints remain insufficiently addressed. The wide variation in reported accuracy (approximately 47%–99%) further reflects inconsistencies in datasets, evaluation protocols, and task complexity.

Overall, future research should focus on developing occlusion-tolerant methods, standardized evaluation frameworks, and structured approaches for interpreting emotional states, particularly for ambiguous cases such as confusion and frustration. Addressing these challenges is essential for building reliable and scalable engagement detection systems. Future research should focus on developing occlusion-aware and region-attentive deep learning frameworks capable of maintaining robust performance under partial facial visibility conditions.Additionally, the creation of large-scale annotated occlusion datasets may further improve the reliability and real-world applicability of occlusion-robust systems.

This review has certain limitations. The study selection was restricted to computer vision-based methods in e-learning contexts, which may exclude relevant work from related domains. Some papers were also excluded due to lack of full-text access or insufficient publication quality, potentially introducing selection bias. Furthermore, the analysis is based on reported results that are not always directly comparable due to variations in datasets and evaluation protocols. The absence of standardized benchmarks in this field further limits definitive comparisons between methods.

Declaration

Availability of data and material:

As this work is a review study, no new experimental datasets or materials were generated. The analysis is based entirely on publicly available literature cited in the references.

Funding:

There is no funding available for this work.

Conflicts of interests:

The authors have no conflicts of interest to declare.

Acknowledgments:

The authors would like to thank their institution for providing the necessary resources to carry out this work.

References

[1] Kentnor, H.E.: Distance education and the evolution of online learning in the united states. Curriculum and teaching dialogue **17**(1), 21–34 (2015) https://doi.org/10.1108/978-1-68123-229-420251014

[2] Feenberg, A.: Building a global network: The wbsi experience. Global networks: Computers and international communication, 185 (1993)

[3] Hrastinski, S.: Asynchronous and synchronous e-learning. Educause quarterly **31**(4), 51–55 (2008)

[4] Eden, C.A., Chisom, O.N., Adeniyi, I.S.: Online learning and community engagement: Strategies for promoting inclusivity and collaboration in education. World Journal of Advanced Research and Reviews **21**(3), 232–239 (2024)

[5] Schnitzler, K., Holzberger, D., Seidel, T.: All better than being disengaged: Student engagement patterns and their relations to academic self-concept and achievement. European Journal of Psychology of Education **36**(3), 627–652 (2021)

[6] Carini, R.M., Kuh, G.D., Klein, S.P.: Student engagement and student learning: Testing the linkages. Research in higher education **47**, 1–32 (2006)

[7] Lase, D., Zega, T.G.C.: How can teachers engage students in online learning? a conceptual framework. Technium Soc. Sci. J. **20**, 143–155 (2021)

[8] Yacek, D.W., Gary, K.: The uses and abuses of boredom in the classroom. British Educational Research Journal **49**(1), 126–141 (2023)

[9] Dewan, M., Murshed, M., Lin, F.: Engagement detection in online learning: a review. Smart Learning Environments **6**(1), 1–20 (2019)

[10] Ibrahim, Q.R., Kurnaz, S.: A systematic literature survey of emotion detection of students in e-learning. In: 2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM), pp. 335–340 (2022). IEEE

[11] Booth, B.M., Bosch, N., D'Mello, S.K.: Engagement detection and its applications in learning: A tutorial and selective review. Proceedings of the IEEE **111**(10), 1398–1422 (2023)

[12] Lek, J.X.-Y., Teo, J.: Academic emotion classification using fer: A systematic review. Human Behavior and Emerging Technologies **2023**(1), 9790005 (2023)

[13] Chauhan, N.K., Singh, K.: A review on conventional machine learning vs deep learning. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON), pp. 347–352 (2018). IEEE

[14] Degadwala, S., Vyas, D.: Survey on systematic analysis of deep learning models compare to machine learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology **10** (2024) https://doi.org/10.32628/CSEIT24103206

[15] Gupta, A., D'Cunha, A., Awasthi, K., Balasubramanian, V.: Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885 (2016)

[16] Murshed, M., Dewan, M.A.A., Lin, F., Wen, D.: Engagement detection in e-learning environments using convolutional neural networks. In: 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/-CyberSciTech), pp. 80–86 (2019). IEEE

[17] Liao, J., Liang, Y., Pan, J.: Deep facial spatiotemporal network for engagement prediction in online learning. Applied Intelligence **51**(10), 6609–6621 (2021)

[18] Dash, S., Dewan, M.A.A., Murshed, M., Lin, F., Abdullah-Al-Wadud, M., Das, A.: A two-stage algorithm for engagement detection in online learning. In: 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), pp. 1–4 (2019). IEEE

[19] Dewan, M.A.A., Lin, F., Wen, D., Murshed, M., Uddin, Z.: A deep learning approach to detecting engagement of online learners. In: 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 1895–1902 (2018). IEEE

[20] Ma, X., Xu, M., Dong, Y., Sun, Z.: Automatic student engagement in online learning environment based on neural turing machine. International Journal of

Information and Education Technology **11**(3), 107–111 (2021)

[21] Abedi, A., Khan, S.S.: Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. In: 2021 18th Conference on Robots and Vision (CRV), pp. 151–157 (2021). IEEE

[22] Malekshahi, S., Kheyridoost, J.M., Fatemi, O.: A general model for detecting learner engagement: implementation and evaluation. arXiv preprint arXiv:2405.04251 (2024)

[23] Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)

[24] Trabelsi, Z., Alnajjar, F., Parambil, M.M.A., Gochoo, M., Ali, L.: Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. Big Data and Cognitive Computing **7**(1), 48 (2023)

[25] Savchenko, A.V., Savchenko, L.V., Makarov, I.: Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Transactions on Affective Computing **13**(4), 2132–2143 (2022)

[26] El Maazouzi, Q., Retbi, A.: Multimodal detection of emotional and cognitive states in e-learning through deep fusion of visual and textual data with nlp. Computers **14**(8), 314 (2025)

[27] Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., *et al.*: Challenges in representation learning: A report on three machine learning contests. In: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20, pp. 117–124 (2013). Springer

[28] Sharma, P., Joshi, S., Gautam, S., Maharjan, S., Khanal, S.R., Reis, M.C., Barroso, J., Jesus Filipe, V.M.: Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In: International Conference on Technology and Innovation in Learning, Teaching and Education, pp. 52–68 (2022). Springer

[29] Altuwairqi, K., Jarraya, S.K., Allinjawi, A., Hammami, M.: Student behavior analysis to measure engagement levels in online learning environments. Signal, image and video processing **15**(7), 1387–1395 (2021)

[30] Bhardwaj, P., Gupta, P., Panwar, H., Siddiqui, M.K., Morales-Menendez, R., Bhaik, A.: Application of deep learning on student engagement in e-learning environments. Computers & Electrical Engineering **93**, 107277 (2021)

[31] Zeng, H., Shu, X., Wang, Y., Wang, Y., Zhang, L., Pong, T.-C., Qu, H.: Emotioncues: Emotion-oriented visual summarization of classroom videos. IEEE transactions on visualization and computer graphics **27**(7), 3168–3181 (2020)

[32] Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P.K.D., Dasgupta, N., Choudhury, P.: Facial emotion detection to assess learner's state of mind in an online learning system. In: Proceedings of the 2020 5th International Conference on Intelligent Information Technology, pp. 107–115 (2020)

[33] Hasnine, M.N., Bui, H.T., Tran, T.T.T., Nguyen, H.T., Akçapınar, G., Ueda, H.: Students' emotion extraction and visualization for engagement detection in online learning. Procedia Computer Science **192**, 3423–3431 (2021)

[34] Sukumaran, A., Manoharan, A.: Multimodal engagement recognition from image traits using deep learning techniques. IEEE Access **12**, 25228–25244 (2024)

[35] Somu, R., Ashok Kumar, P.: Analysis of learner's emotional engagement in online learning using machine learning adam robust optimization algorithm. Scientific Programming **2024**(1), 8886197 (2024)

[36] Alruwais, N., Zakariah, M.: Detecting student engagement with convolution neural network and facial expression recognition. Traitement du Signal **42**(2) (2025)

[37] Wang, S., Chen, M., Ratnavelu, K., Shibghatullah, A.S.B., Keoy, K.H.: Online classroom student engagement analysis based on facial expression recognition using enhanced yolov5 for mitigating cyberbullying. Measurement Science and Technology **36**(1), 015419 (2024)

[38] Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing **28**(1), 356–370 (2019)

[39] Gupta, S., Kumar, P., Tekchandani, R.K.: Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications **82**(8), 11365–11394 (2023)

[40] Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998). IEEE

[41] Yang, D., Alsadoon, A., Prasad, P.C., Singh, A.K., Elchouemi, A.: An emotion recognition model based on facial recognition in virtual learning environment. Procedia Computer Science **125**, 2–10 (2018)

[42] Yao, Q., Wang, M., Li, Y.: Visual geometry group-swishnet-based asymmetric facial emotion recognition for multi-face engagement detection in online learning

environments. Symmetry **17**(5), 711 (2025)

[43] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 Ieee Computer Society Conference on Computer Vision and Pattern Recognition-workshops, pp. 94–101 (2010). IEEE

[44] Gupta, S., Kumar, P., Tekchandani, R.: A multimodal facial cues based engagement detection system in e-learning context using deep learning approach. Multimedia Tools and Applications **82**(18), 28589–28615 (2023)

[45] Jonathan, O.: Face expression recognition dataset. https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset

[46] Sidhu, P.K., Kapoor, A., Solanki, Y., Singh, P., Sehgal, D.: Deep learning based emotion detection in an online class. In: 2022 IEEE Delhi Section Conference (DELCON), pp. 1–6 (2022). IEEE

[47] Eroglu Erdem, C., Turan, C., Aydin, Z.: Baum-2: A multilingual audio-visual affective face database. Multimedia tools and applications **74**(18), 7429–7459 (2015)

[48] Jagadeesh, M., Baranidharan, B.: Facial expression recognition of online learners from real-time videos using a novel deep learning model. Multimedia Systems **28**(6), 2285–2305 (2022)

[49] Kaur, A., Mustafa, A., Mehta, L., Dhall, A.: Prediction and localization of student engagement in the wild. In: 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8 (2018). IEEE

[50] Selim, T., Elkabani, I., Abdou, M.A.: Students engagement level detection in online e-learning using hybrid efficientnetb7 together with tcn, lstm, and bi-lstm. IEEE Access **10**, 99573–99583 (2022)

[51] Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Chen, F., Wang, X.: A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Transactions on Multimedia **12**(7), 682–691 (2010)

[52] Yue, J., Tian, F., Chao, K.-M., Shah, N., Li, L., Chen, Y., Zheng, Q.: Recognizing multidimensional engagement of e-learners based on multi-channel data in e-learning environment. IEEE Access **7**, 149554–149567 (2019)

[53] Ariz, M., Bengoechea, J.J., Villanueva, A., Cabeza, R.: A novel 2d/3d database with automatic face annotation for head tracking and pose estimation. Computer Vision and Image Understanding **148**, 201–210 (2016)

[54] Uçar, M.U., Özdemir, E.: Recognizing students and detecting student engagement with real-time image processing. Electronics **11**(9), 1500 (2022)

[55] Zhang, Z., Li, Z., Liu, H., Cao, T., Liu, S.: Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. Journal of Educational Computing Research **58**(1), 63–86 (2020)

[56] Okur, E., Alyuz, N., Aslan, S., Genc, U., Tanriover, C., Arslan Esme, A.: Behavioral engagement detection of students in the wild. In: Artificial Intelligence in Education: 18th International Conference, AIED 2017, Wuhan, China, June 28–July 1, 2017, Proceedings 18, pp. 250–261 (2017). Springer

[57] Panichkriangkrai, C., Silapasuphakornwong, P., Saenphon, T.: Emotion recognition of students during e-learning through online conference meeting. Science, Engineering and Health Studies, 21020010–21020010 (2021)

[58] Ayvaz, U., Gürüler, H., Devrim, M.O.: Use of facial emotion recognition in e-learning systems (2017)

[59] Karan, K.V., Bahel, V., Ranjana, R., Subha, T.: Transfer learning approach for analyzing attentiveness of students in an online classroom environment with emotion detection. In: Roy, S., Sinwar, D., Perumal, T., Slowik, A., Tavares, J.M.R.S. (eds.) Innovations in Computational Intelligence and Computer Vision, pp. 253–261. Springer, Singapore (2022)

[60] Ferreira, F.R.T., Couto, L.M., Melo Baptista Domingues, G., Saporetti, C.M.: Development of a framework using deep learning for the identification and classification of engagement levels in distance learning students. Social Network Analysis and Mining **15**(1), 37 (2025)

[61] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74 (2018). IEEE

[62] Ahmed, Z.A., Jadhav, M.E., Al-madani, A.M., Tawfik, M., Alsubari, S.N., Shareef, A.A.A.: Real-time detection of student engagement: Deep learning-based system. In: International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 1, pp. 313–323 (2022). Springer

[63] Viola, P., Jones, M.J.: Robust real-time face detection. International journal of computer vision **57**, 137–154 (2004)

[64] Videla, S., Kumar, P.: Modified feature extraction using viola jones algorithm. J. Adv. Res. Dyn. Control Syst **10**, 528–538 (2018)

[65] Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters **23**(10), 1499–1503 (2016)

[66] Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Chaurasia, A., Diaconu, L., Ingham, F., Colmagro, A., Ye, H., et al.: ultralytics/yolov5: v4. 0-nn. silu () activations, weights & biases logging, pytorch hub integration. Zenodo (2021)

[67] Das, R., Dev, S.: Optimizing student engagement detection using facial and behavioral features. Neural Computing and Applications, 1–23 (2025)

[68] Baltrušaitis, T., Robinson, P., Morency, L.-P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10 (2016). IEEE

[69] Hai, L., Guo, H.: Face detection with improved face r-cnn training method. In: Proceedings of the 3rd International Conference on Control and Computer Vision, pp. 22–25 (2020)

[70] Sherly, E., *et al.*: Fostering learning with facial insights: Geometrical approach to real-time learner engagement detection. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), pp. 1–6 (2024). IEEE

[71] Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649 (2017). IEEE

[72] Ross, D.A., Lim, J., Lin, R.-S., Yang, M.-H.: Incremental learning for robust visual tracking. International journal of computer vision **77**, 125–141 (2008)

[73] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

[74] Jabid, T., Kabir, M.H., Chae, O.: Robust facial expression recognition based on local directional pattern. ETRI journal **32**(5), 784–794 (2010)

[75] Arriaga, O., Valdenegro-Toro, M., Plöger, P.: Real-time convolutional neural networks for emotion and gender classification. arXiv preprint arXiv:1710.07557 (2017)

[76] Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Transactions on Affective Computing **5**(1), 86–98 (2014)

[77] Villaroya, S.M., Gamboa-Montero, J.J., Bernardino, A., Maroto-Gómez, M.,

Castillo, J.C., Salichs, M.Á.: Real-time engagement detection from facial features. In: 2022 IEEE International Conference on Development and Learning (ICDL), pp. 231–237 (2022). IEEE

[78] Canu, S.: Eye detection-gaze controlled keyboard with python and opencv p. 1. Pysource, Jan **7** (2019)

[79] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)

[80] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)

[81] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[82] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

[83] Anwar, A., Raychowdhury, A.: Masked face recognition for secure authentication. arXiv preprint arXiv:2008.11104 (2020)

[84] Hinton, G.E., Osindero, S., Teh, Y.-W.: A fast learning algorithm for deep belief nets. Neural computation **18**(7), 1527–1554 (2006)

[85] Vishnumolakala, S.K., Vallamkonda, V.S., Sobin, C., Subheesh, N., Ali, J.: In-class student emotion and engagement detection system (iseeds): an ai-based approach for responsive teaching. In: 2023 IEEE Global Engineering Education Conference (EDUCON), pp. 1–5 (2023). IEEE

[86] Arwa, A., Khawlah, A., Salma, K.J., Nihal, A., Samar, A.: Cnn-based face emotion detection and mouse movement analysis to detect student's engagement level. In: International Conference on Advanced Intelligent Systems for Sustainable Development, pp. 604–626 (2022). Springer

[87] Li, Y.-Y., Hung, Y.-P.: Feature fusion of face and body for engagement intensity detection. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3312–3316 (2019). IEEE

[88] Ahmad, N., Khan, Z., Singh, D.: Student engagement prediction in moocs using deep learning. In: 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 1–6 (2023). IEEE

[89] Xie, N., Liu, Z., Li, Z., Pang, W., Lu, B.: Student engagement detection in

online environment using computer vision and multi-dimensional feature fusion. Multimedia Systems **29**(6), 3559–3577 (2023)

[90] Ji, X., Wang, R.: Facial expression recognition using adaptive weighted lgcp and fast sparse representation. Comput. Eng. Appl **53**, 158–162 (2017)

[91] Zhao, X., Wei, C.: A real-time face recognition system based on the improved lbph algorithm. In: 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), pp. 72–76 (2017). IEEE

[92] Altuwairqi, K., Jarraya, S.K., Allinjawi, A., Hammami, M.: A new emotion–based affective model to detect student's engagement. Journal of King Saud University-Computer and Information Sciences **33**(1), 99–109 (2021)

[93] Monkaresi, H., Bosch, N., Calvo, R.A., D'Mello, S.K.: Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Transactions on Affective Computing **8**(1), 15–28 (2016)

[94] Cao, J.T., Burgess, B., Foster, J.K., Yaoyuneyong, G., Wallace, L.K.: Rapid transitioning to remote learning: Shared responsibility and coping strategies. Journal of Marketing Education **44**(2), 235–249 (2022)

[95] Lubicz-Nawrocka, T., Bao, X.: Partnership in the classroom: Engaging students through inclusive student-teacher relationships to advance social justice. Social Sciences **14**(2), 75 (2025)

[96] Li, R.: The role of teacher-student interpersonal relations in flipped learning on student engagement. Frontiers in Psychology **Volume 12 - 2021** (2021) https://doi.org/10.3389/fpsyg.2021.741810

[97] Shao, Y., Feng, Y., Zhao, X., Liu, G., Zhang, L.: Teacher support and secondary school students' learning engagement: A moderated mediation model. Scientific Reports **15**(1), 2974 (2025)

[98] Sun, B., Wu, Y., Zhao, K., He, J., Yu, L., Yan, H., Luo, A.: Student class behavior dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. Neural Computing and Applications **33**, 8335–8354 (2021)

[99] Martin, F., Bolliger, D.U.: Engagement matters: Student perceptions on the importance of engagement strategies in the online learning environment. Online learning **22**(1), 205–222 (2018)

[100] Singh, A., Singh, D., Chhikara, S.: Online learning: Challenges and suggestions to enhance student engagement in higher education institutions. In: Reshaping Entrepreneurial Education Within an Industry 4.0 Context, pp. 59–80. IGI Global, ??? (2024)

[101] Ko, B.C.: A brief review of facial emotion recognition based on visual information. sensors **18**(2), 401 (2018)

[102] Mehrabian, A.: Communication without words. In: Communication Theory, pp. 193–200. Routledge, ??? (2017)

[103] Kaulard, K., Cunningham, D.W., Bülthoff, H.H., Wallraven, C.: The mpi facial expression database—a validated database of emotional and conversational facial expressions. PloS one **7**(3), 32321 (2012)

[104] Pekrun, R.: The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review **18**, 315–341 (2006) https://doi.org/10.1007/s10648-006-9029-9

[105] Furrer, C., Skinner, E.: Sense of relatedness as a factor in children's academic engagement and performance. Journal of educational psychology **95**(1), 148 (2003)

[106] Li, Y., Liu, H., Liang, J., Jiang, D.: Occlusion-robust facial expression recognition based on multi-angle feature extraction. Applied Sciences **15**(9), 5139 (2025)

[107] Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M.: Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Transactions on Affective Computing **10**(2), 223–236 (2017)

[108] Gendron, M., Roberson, D., Vyver, J.M., Barrett, L.F.: Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. Emotion **14**(2), 251 (2014)

[109] Jack, R.E., Blais, C., Scheepers, C., Schyns, P.G., Caldara, R.: Cultural confusions show that facial expressions are not universal. Current biology **19**(18), 1543–1548 (2009)