Received: 17 November 2020 | Revised: 28 August 2021 | Accepted: 15 September 2021
DOI: 10.1111/exsy.12839

ORIGINAL ARTICLE

Expert Systems WILEY

An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition

Chakradhar Pabba | **Praveen Kumar**

Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology, Nagpur, India

Correspondence
Chakradhar Pabba, Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology, Nagpur, Maharashtra, 440010, India.
Email: pck2507@gmail.com

Abstract

Students' disengagement problem has become critical in the modern scenario due to various distractions and lack of student-teacher interactions. This problem is exacerbated with large offline classrooms, where it becomes challenging for teachers to monitor students' engagement and maintain the right-level of interactions. Traditional ways of monitoring students' engagement rely on self-reporting or using physical devices, which have limitations for offline classroom use. Student's academic affective states (e.g., moods and emotions) analysis has potential for creating intelligent classrooms, which can autonomously monitor and analyse students' engagement and behaviours in real-time. In recent literature, a few computer vision based methods have been proposed, but they either work only in the e-learning domain or have limitations in real-time processing and scalability for large offline classes. This paper presents a real-time system for student group engagement monitoring by analysing their facial expressions and recognizing academic affective states: 'boredom,' 'confuse,' 'focus,' 'frustrated,' 'yawning,' and 'sleepy,' which are pertinent in the learning environment. The methodology includes certain pre-processing steps like face detection, a convolutional neural network (CNN) based facial expression recognition model, and post-processing steps like frame-wise group engagement estimation. For training the CNN model, we created a dataset of the aforementioned facial expressions from classroom lecture videos and added related samples from three publicly available datasets, BAUM-1, DAiSEE, and YawDD, to generalize the model predictions. The trained model has achieved train and test accuracy of 78.70% and 76.90%, respectively. The proposed methodology gave promising results when compared with self-reported engagement levels by students.

KEYWORDS
affective computing, computer vision, convolutional neural network, deep learning, offline classroom analytics, student engagement

1 | INTRODUCTION

The recent technological advances of artificial intelligence (AI) enable the development of intelligent systems in many fields, like health care (Davenport & Kalakota, 2019), traffic management (Abduljabbar et al., 2019) and education (Chen, L., et al. 2020). The implementation of such systems in large offline classroom management assists teachers to reduce the burden of laborious activities such as attendance tracking, collecting

 © 2021 John Wiley & Sons Ltd.

lecture feedback, students' engagement or attention monitoring, thereby strengthening the best teaching-learning outcomes (Kim et al., 2018). Hence, in recent years it has become active and challenging research field.

Students' engagement in classroom learning is essential as it improves the overall class learning quality and academic progress (De Villiers & Werner, 2016). Currently, students' disengagement problem is rising every day due to various reasons, like short attention span, lack of student-teacher interactions, inadequate teaching methodology (Bradbury, 2016; Lamba et al., 2014). This problem is exacerbated with large offline classrooms (student count is 60 or more). Well-experienced teachers can monitor students' engagement by observing their behaviour and interactions in SMALL classes. Still, even such experienced teachers face difficulties as the class strength increases and become unscalable beyond a threshold number of students due to human/manual limitations (Exeter et al., 2010). Moreover, many universities (significantly higher education institutes) do not have all the faculties who are well-experienced teaching experts. They typically receive no or very little training/allocated time to teach and improve pedagogical skills that can increase students' engagement and participation. There remain several challenges for instructors who want to improve their teaching, including the lack of sufficient feedback opportunity on their pedagogical skills. Currently, the most effective practice for this kind of professional development employs a professional human expert to observe one or more lectures and provide personalized formative feedback to the instructor. Of course, this is expensive, unscalable and more importantly, precludes the instructor's continuous learning feedback loop. Therefore, the proposed automated approach in this work for students' engagement monitoring can support professional development for novice teachers at scale and could also potentially assist experienced teachers in assessing and improving students' engagement and the overall teaching-learning process for large classes.

In the educational research literature, student's engagement is defined with multi-dimensions and components. Fredricks et al. (2004) defined it in three ways: behavioural, emotional, and cognitive engagements. Behavioural engagement describes behavioural actions in the learning process, such as proper body posture and writing notes. Emotional engagement describes positive and negative emotional reactions towards learning, such as attention, boredom, frustrated. Cognitive engagement leads to learning to increase cognitive ability, including problem solving, knowledge and creative thinking. According to (Li, Y., & Lerner, R. M. 2013), behavioural and emotional engagements are bi-directionally related to each other. Besides, behavioural engagement influences cognitive engagement, which is an essential outcome of the learning process. The work (Sathik & Jonathan, 2013) has statistically proven that students' facial expressions are significantly correlated to their behavioural and emotional states, which can help recognize their engagement towards the lecture.

Whitehill et al. (2014) classified student engagement estimation methods into three categories, namely: manual, semi-automatic and automatic methods. Manual methods include paper-based or computer-based self-reports (Haddad 2014), survey-based methods like National Survey of Student Engagement (NSSE) (Kuh, G.D. 2003), Student Engagement Instrument (SEI) (Appleton et al., 2006), observational checklist, and rating scale (Odiri Amatari, 2015; Dzelzkaleja & Kapenieks, 2016) methods. These methods remain laborious, tedious, interruptive, and vulnerable to biases. Semi-automatic methods include knowledge tracing and physiological based methods. In knowledge tracing, teachers estimate student's engagement by assessing their responses to the questions asked while teaching. The work done in (Griol, D., et al. 2017; Mogwe, A. W. 2018) is used to perform this method effortlessly. In physiological based methods, student's engagement is estimated by processing physiological signals like brain signals (Electroencephalogram [EEG]), heart signals (Electrocardiogram [ECG]) using wearable devices like fitness bands, electrodermal activity sensors (Di Lascio et al., 2018). These semi-automatic methods have limitations such as susceptibility to human intervention, expensive. Besides, to measure physiological signals, different wearable electrodermal activity sensors are attached to the user by cables, which causes difficulty to wear for a longer time (Dirican & Göktürk, 2011). Automatic methods include vision based methods that measure a student's engagement by analysing non-verbal cues like facial expressions and head gaze from the video captured by high-definition (HD) surveillance cameras. Such an automated method is non-intrusive, effective, easy, unbiased, and inexpensive way to measure student engagement in any learning environment, such as online or offline classroom learning.

AI technological advances like affective computing, computer vision and deep learning are used to develop the automatic engagement monitoring systems (AEMS). AEMS monitors and reports the students' engagement automatically by analysing the non-verbal cues without human intervention. In recent literature, many works have been proposed to develop AEMS in the education domain using student's behavioural and emotional cues. Out of them, most of the works addresses the e-learning environment with a single student in a single video frame (Bosch et al., 2016; Krithika, L., & GG 2016; Ruipérez-Valiente et al., 2018; Sharma et al., 2019; Whitehill et al., 2014; Zhang, H., et al. 2019; Mukhopadhyay et al., 2020; Bhardwaj, P., et al. 2021). A few works are supported in the offline classroom environment with multiple students in a single video frame (Zaletelj & Košir, 2017; Klein & Celik, 2017; Thomas & Jayagopi, 2017; Soloviev, 2018; Ashwin & Guddeti, 2019; Zheng, R., et al. 2020; Luo, Z., et al. 2020; Vanneste et al., 2021; Peng, S., et al. 2021). Most of these works have scalability issues for large offline classroom environments and the inability to estimate student group engagement in real-time. Moreover, these works have used basic emotions like happy, angry, fear, sad, and surprise (Ekman, 1992) as emotional cues for engagement estimation, which is not suitable for academic environment. Early research has proven that there are some distinctions between academic and basic emotions (Pekrun, 2000).

This paper addresses this research gap by exploring an automatic real-time system to monitor student group engagement by analysing their facial expressions and recognizing academic affective states. Academic affective states are part of specific affective states that cannot be represented by basic emotions (Wei, Q., et al. 2017). For this study six meaningful academic affective states are used, namely: 'boredom,' 'confuse,' 'focus,' 'frustrated,' 'yawning,' and 'sleepy,' pertinent in the learning environment (D'Mello, S. 2013; Tonguç & Ozkara, 2020). We created a facial

expressions dataset for these academic affective of students' faces from classroom lecture videos. The dataset has been added with similar facial expression samples from three public datasets: BAUM-1 (Zhalehpour et al., 2016), DAiSEE (Gupta et al., 2016; Kamath et al., 2016), and YawDD (Abtahi et al., 2014, March) to increase the dataset samples and generalize the model predictions to some extent. This work also describes how the proposed system addresses the ethical and students' data privacy issues which are to be considered in this type of work.Towards this aim, the following research questions are explored and the attempted solution is presented in this paper: Can we detect each student's face in every image frame of a large offline classroom video? Can we recognize a student's academic affective states through facial expressions? Can we compute a single student group engagement score for each video frame? Can we estimate overall students' engagement in real-time with adequate computing resources? Can we validate the correlation between students' engagement which the AEMS model and self-reports estimate? Many researchers have already addressed the face detection in video frame issue (Kumar, A., et al. 2019; Viola, P., & Jones, M. 2001; Zhang, K., et al. 2016). Hence, this paper focuses on addressing the remaining issues.

The following are key highlights of this paper:

- Develop a vision-based automated system for student group engagement level monitoring in a large offline classroom environment by analysing students' academic affective states through facial expression.
- Created a facial expressions dataset for academic affective states from the classroom lecture video frames of more than 40 students to train the facial expression recognition system.
- Develop a convolution neural network based facial expression recognition model with train and test accuracy of 78.70% and 76.90% respectively, which works on reasonably low-resolution face images of size 48 $\times$ 48 pixels.
- Compared the proposed methodology output values with students' engagement levels self-reported ground truth values yielded a correlation of 73.68%.

The remaining parts of the paper are organized as follows: Section 2 discusses the related work. Section 3 gives a brief background of related AI technologies. Section 4 describes the proposed methodology. Section 5 presents the experimentation methodology, results, discussions, and limitations. Section 6 concludes with summary remarks and future research directions.

2 | RELATED WORK

This section provides a broad overview of recent works that have been used to address the problem of estimating students' engagement in online and offline learning environments. As discussed in introduction section, engagement estimation methods are categorized into three groups, namely: Manual, semi-automatic and automatic methods. Due to the success of various vision-based machine-learning and deep learning solutions, automatic methods are gaining attention in recent years. These automated methods can be particularly useful in large offline classrooms and also in the e-learning/online learning scenarios.

2.1 | Automatic methods

Whitehill et al. (2014) proposed a machine learning based system for classifying student's engagement in an e-learning environment by analysing student's facial expressions and behaviour patterns. Their experiments concluded that SVM with Gabor features performed the best in classifying student engagement, with an area under the curve (AUC) value of 0.729. The work (Bosch et al., 2016) used computer-vision and machine-learning algorithms to detect student's affect from facial expressions and body movements during interactions with educational games on a computer. They build 14 different machine-learning models like SVM, decision trees for this purpose. Each affective state classification accuracy measured in terms of AUC values: bored (0.61), confused (0.65), delighted (0.87), engaged (0.68), and frustrated (0.63). Krithika, L., & GG (2016) developed a system that recognizes and monitors student's emotions in an e-learning environment and provides real-time feedback on student's concentration levels. They used emotions like excite, boredom, yawning, drowsiness, abnormal head and eyes moving patterns to predict concentration level. The work (Sharma et al., 2019) proposed a real-time student's engagement estimation system for the e-learning environment by analysing the student's basic facial expressions. They trained a CNN-based emotion recognition model with a validation accuracy of 70%. Zhang, H., et al. (2019) proposed an inflated three-dimensional convolutional network (I3D) based binary classification model for students' engagement recognition system for online learning environment on the DAiSEE data set. It achieved an accuracy of 0.98% for binary engagement classification, engaged or non-engaged. Mukhopadhyay et al. (2020) presented a method for assessing students' affective states in online learning by combinations of the basic facial expressions. They proposed and trained a convolutional neural network (CNN) based model with the FER2013 dataset and achieved a classification accuracy of 62%. P Bhardwaj et al. (2021) presented a deep learning based methodology for real-time student engagement classification in an online learning environment by analysing the basic facial expressions. All the aforementioned methods address

the automatic student engagement monitoring in an e-learning environment, with a single student in a single video frame. Therefore, these works are unfeasible to address the automatic student group engagement estimation problem in a large offline classroom environment, with multiple students in a single video frame.

Zaletelj and Košir (2017) attempted to automatically estimate students' attention in an offline classroom environment using non-verbal cues. They developed a model with extracted 2D and 3D features from the Kinect One camera (Zhang, Z. 2012) using machine learning algorithms like decision trees and k-nearest neighbours. Their system achieved 0.753% test accuracy, which was assessed by comparing predicted attention to ground-truth attention given by human annotations. Due to technological limitations with the Kinect cameras, the analysis was limited to only six students rather than the whole classroom. Klein and Celik (2017) developed wits intelligent teaching system (WITS), a CNN-based approach to assist teachers with real-time feedback of students' engagement using positive and negative behavioural cues in the large offline classroom environment. They trained a model based on Alexnet architecture (Krizhevsky et al., 2012) with the created students' classroom behaviour dataset and achieved a validation accuracy of 89.60%. The study did not use emotional cues to estimate students' engagement, besides it is also involved computational overhead. The work (Thomas & Jayagopi, 2017) used computer vision and machine learning algorithms to classify student's engagement with non-verbal facial cues. They created a dataset comprising relevant features that comes to a 27-dimensional feature vector, using an open-source real-time facial analysis toolbox called OpenFace (Baltrušaitis et al., 2016). They have trained models utilizing machine learning algorithms like SVM, logistic regression with the dataset and achieved a classification accuracy of 0.89% and 0.76%, respectively. The study was not experimented in the large offline classroom environment. Besides, it was not tested for estimating overall class student group engagement in real-time. Soloviev (2018) presented a system for continuously analysing visual data streams from classroom cameras by classifying students' basic facial expressions as positive or negative emotions. They trained a model with a two-class boosted decision tree (Adaboost) method, which resulted in a classification accuracy of 84.80%. The study did not consider the student's academic emotions to classify their engagement levels.

The work (Ashwin & Guddeti, 2019) developed a CNN based system that analyzes non-verbal cues and classifies students' engagement into one of its four levels: "Not engaged at all," "Nominally engaged," "Engaged in the task," and "Very engaged." Their system was trained and tested on faces, hand gestures, and body postures in a large offline classroom environment with 71% of engagement classification accuracy. The method requires 2153 ms (2.153 s) to process a single image frame, indicating a computational overhead. As a result, it cannot be used in real-time implementations. The work (Zheng, R., et al. 2020) designed an intelligent student's behaviour monitoring framework that detects students' behaviours like hand raising, standing, and sleeping, in a classroom environment. They trained the model with an improved Faster R-CNN object detection algorithm to recognize the aforementioned behaviours with a mean Average Precision (mAP) value of 57.6%. Since this model works only to detect students' behaviours, it cannot be used to predict overall student group engagement using academic affective cues. The study Luo, Z., et al. 2020 proposed a 3D model that includes hierarchical and conditional random forest algorithms and an interactive platform that uses head pose, facial expressions, and smartphones to estimate student interest in classroom settings. The model was achieved a classification accuracy of 87.5%. Peng, S., et al., (2021) presented a multi-modality fusion of face cues, heart rate, and auditory features to monitor students' mental states. A set of machine learning algorithms SVM, random forest, and multilayer perceptron have been trained using various multimodal fusion techniques. The above two studies required variety of physical devices to measure student's multi-modality data, which involves high cost to use in large offline classroom environments. The study (Vanneste et al., 2021) suggested a technique for assessing students' engagement in classroom settings by recognizing their behaviours such as raising hands and taking notes. A deep learning model was trained to recognize such behaviours, and it achieved recall and accuracy scores of 63% and 45%, respectively. The study did not experiment with a large classroom setting for real-time engagement estimation. Furthermore, it did not considered students' academic affective states in its approach. None of the above mentioned works have attempted the development of AEMS for real-time students' engagement monitoring by analysing the students' academic affective states through facial expressions in a large offline classroom environment. This work is an attempt to address these research gaps. Some of the aforementioned works that use classroom settings are summarized in Table 1.

3 | PRELIMINARIES

This section briefly presents the significance of AEMS systems and background of different AI-related technologies used in this work: affective computing, computer vision, and CNN.

3.1 | Significance of AEMS

Implementation of AEMS in the education domain could have a vast range of applications. (1) In distance-learning settings, human teachers can receive real-time feedback on their students' engagement levels (low, medium, high) (Whitehill et al., 2014). (2) The student's reactions to instructional videos could automatically recognize and modify the video content, causing the audience to lose interest (Whitehill et al., 2014). (3) Educational analysts could obtain significant data volumes to mine the factors and variables that influence students' engagement. In comparison to self-

TABLE 1 Summary of works related to offline classroom students' engagement monitoring

Works	Methodology	Outcome	Limitation
S Peng et al., (2021)	Trained a set of machine learning algorithms SVM, random forest, and multilayer perceptron using features from facial, heart rate, and auditory modalities.	They monitor student's mental states (boredom, frustrated, concentration, and confusion) during classroom discussions.	The study needs physical devices to measure students' multi-modality data like Apple watch and Air pods, which involve a high cost to use in large offline classroom settings.
Vanneste et al., (2021)	Trained a deep learning based action classifier with the dataset of 8 students' classroom actions like hand-raising, hand on face.	They assess students' engagement in the offline classroom by recognizing their classroom actions through body pose.	The study used students' behavioural cues but not academic emotional cues in estimating their engagement.
Luo, Z., et al., (2020)	They trained a 3D model that includes hierarchical and conditional random forest algorithms and an interactive platform.	The 3D model uses head pose, smiling facial expression, and smartphones to estimate student interest in classroom settings.	They study needs physical devices to get interactive data, which involves cost, and it uses only smiling facial emotion in the student attention estimation.
Zheng, R., et al. (2020)	Trained an improved Faster R-CNN model on the created classroom students' behavioural dataset of hand raising, standing and sleeping behaviours.	They identify students' engagement by analysing the detected behaviours by the trained model.	This study works only for students' behavioural analysis, not for academic emotions based students' engagement estimation.
Ashwin & Guddeti (2019)	Trained an anchor based detection framework, a CNN-based architecture using self-created classroom student behavioural dataset.	Classifies engagement into one of the four engagement levels, that is, no, low, medium and high engagement.	The work does not emphasized on facial emotions analysis and proposed system takes 2153 ms (2.15 s) to process single image frame.
Soloviev (2018)	Trained a two-class boosted decision tree model using facial features like head pose, facial expressions.	Classify either engaged or not engaged student.	The authors do not considered learning-centred facial emotions and the proposed system has not tested in a large offline classroom settings.
Zaletelj and Košir (2017)	Trained decision tree and K-nearest neighbours algorithm using extracted 2D and 3D features.	Classify attention levels into high, medium, and low.	The Kinect sensor has a short range of imaging system. Hence, it does not work for large classrooms.
Klein and Celik (2017)	Performed transfer learning for AlexNet using self-developed dataset of common classroom actions and postures.	Categories attention levels either interested or not interested.	The work has not considered the students' academic emotions in their engagement estimation.
Thomas and Jayagopi (2017)	Trained an SVM model on a dataset of student eye gaze, head pose and facial expressions.	Classifies engagement into engagement levels of engaged or not engaged.	The work has experimented in a small classroom with fewer students. Also not supports for real-time engagement monitoring.

report and questionnaire results, these data will have a high temporal resolution (Whitehill et al., 2014). (4) The analysis of students' engagement can be used as immediate feedback to adjust the instructional strategy to enhance the student learning process (Ashwin & Guddeti, 2019). (5) The day to day feedback on instructional strategy is beneficial to novice teachers to improve their teaching experience quickly (Ashwin & Guddeti, 2019). (6) There are many learning environments on campus in the present era of smart campuses and smart universities, including classrooms, web seminars, and e-learning environments. Manual student monitoring is difficult, and can be solved using AEMS (Al-Nawashi et al., 2017; Ashwin & Guddeti, 2019).

In addition to being used for the education domain, AEMS can also be used in many other domains like entertainment (Wang, S., & Ji,Q. 2015), healthcare (Singh & Goyal, 2021), shopping (Yolcu et al., 2020) and more. Since AEMS can be used in various fields, each field needs to redesign a different set of contextual features according to the engagement dimensions to obtain better predictions. Working with people's visual data is sensitive. The development and use of such autonomous systems based on emotional-AI and affective computing technology introduce a new set of ethical issues that require responsible behaviour, such as system design, ethical data usage, transparency, and privacy (Gretchen Greene 2020; Robin Murdoch 2020).

3.2 | Affective computing

Affective computing (AC) is an area of research for developing systems and devices that are capable of perceiving, recognizing and processing human affects. It is an interdisciplinary domain that includes computing, psychology, and cognitive science. With the aid of AC, we can transform the computing machines into emotionally intelligent machines that can interpret human emotions and respond accordingly. AC has a vast range of applications in many domains like education, healthcare, smart homes, entertainment, etc. According to AC researchers, human communication not only depends on verbal communication like vocal and text but also depends on non-verbal communications like facial expressions, eye gaze, head gaze, hand gestures, and body postures (Poria et al., 2017).

The work (Sathik & Jonathan, 2013) proved that the facial expressions are more often used non-verbal communication channels by the students attending lectures in the classroom. These facial expressions are fewer occluded non-verbal parameters despite the seating arrangement and size of classrooms. Also, processing this parameter is computationally less intensive when compared to the processing of other non-verbal parameters like body posture estimation. Hence, in this paper, facial expressions have been considered as a meaningful feature for estimating student group engagement.

3.3 | Computer vision

Computer Vision (CV) facilitates computers to see and understand visual image content, like photos and videos. Generally, based on computing complexity, CV methods are categorized into three levels, that is, low, medium, and high. Low-level methods include techniques for low-level vision processing like quality enhancement, image transformations, and edge detection. Mid-level methods include mid-level vision processing techniques like object segmentation, motion estimation, and image foreground or background subtraction. High-level methods include image classification, object detection, activity recognition, and pose estimation techniques for high-level vision processing. CV has a vast application domain like the education sector, human-machine interaction, machine inception, medical imaging, surveillance, and health care (Wiley & Lucas, 2018). This work applies several CV techniques like face detection and face alignment to build a framework for AEMS development.

3.4 | Deep learning

Deep Learning (DL) is a specialization of Machine Learning (ML) techniques that deals with algorithms inspired by the brain structure and function called artificial neural network (McCulloch & Pitts 1943). ML based systems are capable of learning and improving knowledge from experience without explicit programming. One of the fundamental facts that influence the ML model's performance is the quality of discriminate features between the given classes. However, manual extraction of such features is a tricky and sometimes complex task. This limitation can be addressed by self-feature extraction models, which come under the DL models' umbrella. The DL architecture uses a hierarchical network with multiple neural network layers. DL frameworks influenced many fields of research like human action recognition (Khaire et al., 2018), stock trading (Chakole & Kurhekar, 2020), bioinformatics (Dasari & Bhukya, 2020), computational biology (Amilpur & Bhukya, 2020) and more (LeCun et al., 2015). The CNN is one such DL framework.

3.4.1 | Convolutional neural networks

The CNN is a type of feed-forward multi-layer neural network that can extract features and optimize model weights using the back-propagation algorithm (Rumelhart et al., 1986). In recent years, CNN based architectures have become one of the most popular DL networks in a broad range of visual computing tasks, like facial emotion recognition (FER) (Pons & Masip, 2017), object detection (Krizhevsky et al., 2012). The CNN has three essential characteristics: region perception, weight distribution, and down-sampling. These characteristics resolve low-level feature extraction overhead and make the model obtain high-level features that are invariant to translation, rotation, and distortion. Figure 1 shows a sample CNN model comprising three blocks. First is the input block, which consists of an input layer. The second is a feature extraction block, which consists of successive pairs of convolution and pooling layers. The third is a classification block that consists of multiple densely connected layers followed by an output layer.

CNN's input layer reads a 2-dimensional (2D) image pixel data and forwards it to the next feature extraction block. The feature extraction block comprises two layers, a convolution layer (Conv2D) and a pooling layer. The Conv2D layer automatically extracts image feature maps. The pooling layer is responsible for the down-sampling of the feature maps. The Conv2D layer consists of multiple independent kernels by randomly initialized learnable weights and biases. The weights are real values that represent the strength of the connection between neurons. The biases are the extra parameters in the network that are used to change the output in addition to the weighted sum of the neuron's inputs. During the

convolution process, each of these kernels slides over the entire image in both horizontal and vertical direction as per the stride count and performs a dot product operation known as a convolution operation in the receptive fields to generate respective feature maps. The receptive field is the region where the image data overlaps with the kernel. The convolution operation can be expressed using Equation 1.

$$F_m(m,n) = Conv2D(I,K)_{mn} = \sum_{i=1}^{d}\sum_{j=1}^{d} K_{(i,j)} \times I_{(i+m,\ j+n)} + b, \tag{1}$$

where $F_m(m,n)$ is the convolution value in the resultant feature map at location (m, n), I is an input image, K is the kernel of size d, and b is the bias term. The feature map dimension depends on the parameters like image padding, kernel size, and stride count. The padding (p) refers to the number of pixels added around to an image when a CNN's kernel is processing it. The kernel size (k) refers to the height and width of a 2D kernel. The stride (s) denotes the number of positions to be moved by the kernel over the image. The size of an output feature map (O) is determined by the Equation 2, where n is the size of an image. Stack of such independent feature maps becomes an input to the next layer, the pooling layer.

$$O = (\lfloor (n-k+2p)/s \rfloor + 1). \tag{2}$$

The pooling layer works on each feature map independently to create a set of down-sampled feature maps by extracting dominating scores. These pooled feature maps make the model invariant to rotation, translation, and distortion. There are different approaches to perform pooling operations like max-pooling, min-pooling, avg-pooling, and it is chosen depending upon the application. These max, min, avg pooling operations return the maximum, minimum, and average values respectively from the image region overlapped with the pooling filter called pool. The pair of convolution and pooling operations considered as one cycle in the feature extraction block. Many such cycles are usually performed to get the most abstract and independent feature maps by the CNN model. As the feature extraction block's final output, all these 2D abstracted feature maps are converted into a single dimension feature vector by a flatten layer. The flatten layer output becomes an input to the next layer called the dense layer in the classification block. The Figure 2 shows schematic illustration of convolution and max-polling operations.

The dense layer comprises neurons, weights, biases, and activation functions. This layer is responsible for the interpretation of the extracted features. An activation function (AF) works as a switch that determines whether or not a neuron should be activated by calculating a weighted sum of input values and weights. The AF aims to introduce non-linearity into a neuron's output. Sigmoid, ReLU, Tanh, and Softmax are some of the popular activation function. These densely connected layers update their randomly initialized weights and biases using the back-propagation

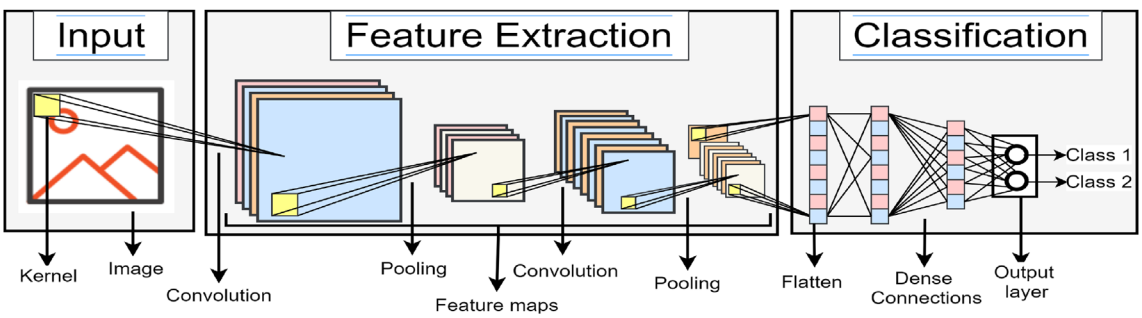

FIGURE 1 A sample CNN's block diagram

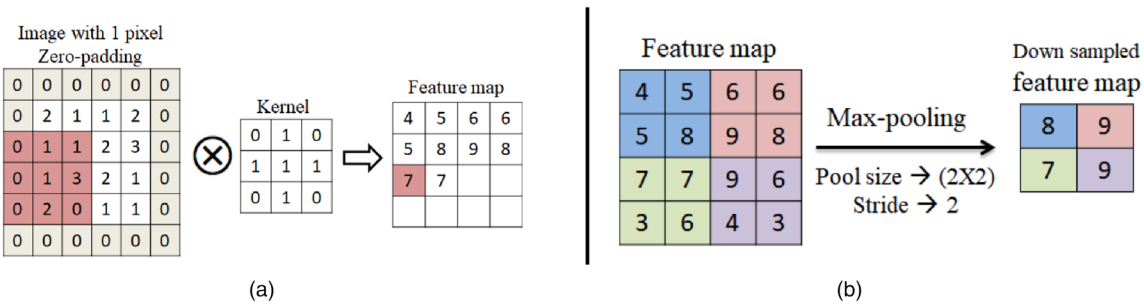

FIGURE 2 Schematic illustration of (a) convolution operation, (b) max-pooling operation

algorithm. Finally, the densely connected layer is linked to the output layer. The output layer contains an AF either sigmoid or softmax to produce the final results. The sigmoid is used for binary class classification tasks, whereas softmax is used for multi-class classification problems. These functions normalize the output values in the range of 0–1, so that they can be represented as class probabilities. The class which gets the highest probability will be assigned as a result of the given input. In this way, CNN classifies the given image into one of classes. In this paper, CNN architecture is proposed and trained to build a FER model.

Overfitting, exploding gradient, vanishing gradients, and dataset's class imbalance are some of the significant challenges when training a CNN-based model (Joshi et al., 2019). These issues can degrade the performance of the model. These challenges can be addressed by fine-tuning the model's hyper-parameters and using regularization methods. The hyper-parameters are the parameters whose values are used to control the learning process of the model, such as batch size, kernel size, loss function, and optimizer. Regularization is a set of methods that can help to avoid the model's overfitting problem. These methods include L1 and L2 regularization, dropout, data augmentation, early stopping (Chatfield et al., 2014). These issues were handled while training the FER model in order to achieve relatively good classification accuracy.

4 | PROPOSED METHODOLOGY

This section describes the proposed methodology to process the classroom video for analysing students' engagement levels (low, medium, and high) and presenting to the teacher in real-time. The framework of the methodology comprises two modules, offline and online modules, as shown in Figure 3. The offline module is a CNN-based trained FER model, and the online module runs in real-time to estimate students' engagement using the CNN model trained in offline module.

4.1 | Privacy protection

As mentioned in the introduction section, the development and use of such automated video-based emotional-AI and affective computing technology introduce a new set of ethical and privacy issues. Some of these issues are system design, transparency, data usage, and data privacy. The system design stipulates that the objectives and purpose should be explained to related stakeholders and be mutually agreed upon. The transparency implies that participants' visual data should be collected only after their consent is obtained by notifying them of the type of data being collected and how it will be used for the research. The consent should be obtained by participants if they are adults and by their parents / guardians if the participants are minors. The data usage describes that the study should ensure that the collected visual data is used only for ethical purposes and should not be misused. The data privacy refers to the measures that should be taken to preserve the data privacy of participants. To ensure this the system should not store, share, sell, or disclose the participant's identity. The measures that have been considered in the current work to address these issues are mentioned below.

There are general privacy and ethical concerns when capturing video data. The proposed work addresses these issues by taking the following measures. The proposed system does not store the complete classroom video. Instead, every incoming video frame will persist in a temporary cache until it is processed, after which it is automatically deleted. Further, to provide more privacy preservation, the system does not individually

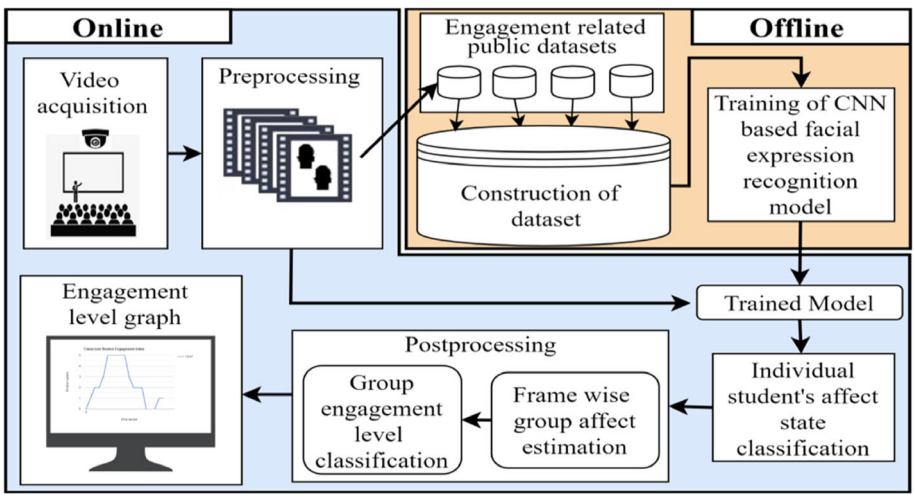

FIGURE 3 The proposed framework for students' engagement estimation

recognize, process, or disclose any student's visual data. Instead, it would return only a higher level value that is "class/group aggregated engagement level index." Only for the purpose of dataset creation, visual data of some students were collected after explaining the goal and intent to all participants and obtaining their consent. We temporarily saved those few classroom video data used for dataset creation to manually label the ground truth of video frames. This dataset will be used purely for academic research and development purpose.

4.2 | Offline module

The offline module executes once to develop a CNN-based FER model, which accepts a face image as input and predicts an appropriate label of affective states as output. A dataset is also constructed to train the CNN architecture as part of the offline module.

4.2.1 | Dataset construction

This section discusses the construction of a dataset for training the FER model, suitable for the learning environment. A variety of research studies have been suggested to tackle the issue of FER (Mollahosseini et al., 2016). Most of these studies focused on basic emotion elicitation (Ekman, 1992), which is not suitable for learning environments (Pekrun, 2000). As a result, these studies are not directly related to students' academic affective states analysis.

4.3 | Academic affective states definition

In developing the FER model for the learning environment, it is essential to identify meaningful academic affective states. Academic affective states are part of specific affective states that cannot be represented by basic affective states (Wei, Q., et al. 2017). Some related works in the literature have reported the relevant students' facial expressions, which help in determining their academic affective states. The work (Ashwin & Guddeti, 2020) reported that engaged, boredom, sleepy, frustrated, confused, and neutral are the dominant affective states that students experience in their classroom learning. The study (Bian, C., et al. 2018) described enjoyment, confusion, fatigue, distraction, and neutral as students' common feelings in online learning. For the current research, the academic affective state needs to have visible facial expressions. As a result of combining the above research with actual needs, six academically meaningful affective states were used, namely: focus, boredom, frustrated, confusion, yawning, and sleepy.

This study uses the definition for these affective states as specified in (Bian, C., et al. 2018; S. D'Mello, Picard et al. 2007; D'Mello, S. K., et al. 2010; Ramón et al., 2017). Focus is a state of attention that comes from participation in an activity. It is coded when the student is highly engaged while listening to the classroom lecture. Boredom is a state of lack of interesting in doing an activity. It is coded when the student is looking away from the lecture or feels not interested in the learning context. Frustrated is a state of dissatisfaction or anger towards an activity. It is coded when the student is dissatisfied or angry while listening to a class lecture. Confusion is a state of emotional disequilibrium. It is coded when the student is confused and unsure how to proceed and have difficulty in understanding the lecture. Yawning is a state of being lethargic or not showing any interest in an activity. It is coded when a student widely opens his or her mouth while listening to a lecture. Sleepy is a state of notably no interest or distraction and a mental state of sleep with an activity. It is coded when a student closes his or her eyes while listening to a classroom lecture.

4.4 | Data collection and participant

Students' spontaneous facial expressions were used to create our "Classroom Spontaneous Facial Expression Dataset" (CSFED) for the aforementioned affective states. These facial expressions were extracted by students' faces from four 30-minute video clips recorded while listening to a lecture course in a classroom environment. These video clips are recorded at 30 frames per second with a resolution of 1920 $\times$ 1080. In total 50 Indian undergraduates, postgraduates and doctoral students from the Computer Science and Engineering department of the Visvesvaraya National Institute of Technology, Nagpur participated in the video recordings. Participants were between 19 and 35 years old, with an average age of 21.02. Among them, 31 are males and 19 are females. Prior to the video recordings, all participants were informed of the purpose and objectives of the study and signed written consent form were obtained acknowledging that their facial photographs could be used for experiments and publication purposes.

4.5 | Annotation of facial expressions

To train a CNN-based supervised model, ground truth labelling for the data samples is necessary. Five well-experienced teachers with a minimum of 8 years of teaching and student counselling experience were invited to help classify the ground truth affective state labels for the extracted students' facial expressions. We believe that it is not very difficult for experienced teachers to use students' facial expressions as the primary clue in estimating their involvement in classroom learning. At least three experienced teachers annotated each facial image with a ground truth affective state label and the label with majority count was assigned as the ground truth label for that facial image. As an additional verification step, the annotated facial images were confidentially shared with psychologist researchers as well to validate the consistency of the ground truth labels. In this process, a total of 1193 students' facial expression images were annotated and validated with the respective academic affective state labels. Figure 4 shows facial expression images of sample classroom students that have been annotated with relevant affective states.

In order to increase the number of data samples and generalize the model predictions, the CSFED has been added with similar facial expression samples from multiple sources, including images from the Internet, and from three publicly available datasets: "BAUM-1," "DAiSEE," and "YawDD." When adding new data samples to the CSFED, the original CSFED characteristics such as age and ethnicity were followed. That means the new data samples' age was chosen to be in the same range and ethnicity was chosen to match the Indian ethnicity as well as the white skin tone. A total of 588 data samples for the six affective states were downloaded from the Internet using a web image search engine like Google images. The "Bahcesehir University Multi modal Face Database (BAUM-1)" is a public dataset of audio-visual facial video clips of acted and spontaneous facial expressions. This dataset comprises 1502 video clips from 31 subjects with various facial expressions, like happiness, anger, sadness, boredom, confusion, concentration, thinking, and bothered. The "Affective States in E-Environments (DAiSEE)" is a public dataset of student's facial expressions in the e-learning environment. It comprises 9068 video clips captured from 112 subjects with affective states of boredom, confusion, engagement, and frustrated. The "Yawning Detection Dataset (YawDD)" is a dataset used for developing drowsiness detection systems. It comprises 351 video samples from 90 subjects, which includes yawning and sleepy affective states. The summary of all datasets used in the construction of the enlarged CSFED is listed in Table 2.

FIGURE 4 Sample facial expressions concerning each affective state, focus (first row), confuse (second row), frustrated (third row), boredom (fourth row), yawning (fifth row) and sleepy (six row)

TABLE 2 Datasets used in the construction of the enlarged CSFED dataset

Dataset		Description	Emotions	Environment
BAUM-1		A dataset of audio-visual facial video clips.	Happiness, sadness, anger, boredom, confused. and so forth.	Indoor
DAiSEE		A dataset of student engagement videos.	Engaged, confused, boredom, and frustrated.	E-learning
YawDD		A dataset of driver drowsiness videos.	Yawning and drowsiness.	In wild
CSFED	Student's video Internet images	A dataset consists of six most frequent academic affective states' facial expressions.	Focus, confusion, frustrated, boredom, yawning and sleepy.	In classroom In wild

TABLE 3 Enlarged CSFED dataset construction details

				CSFED		
	BAUM-1	DAiSEE	Yaawd	Student's images	Internet images	Enlarged CSFED
Focus	250	100	77	166	114	**707**
Confusion	500	84	30	101	89	**804**
Frustrated	234	106	11	121	52	**524**
Boredom	209	277	18	314	91	**909**
Yawning	6	8	217	128	173	**532**
Sleepy	174	68	112	363	69	**786**
Total	**1369**	**637**	**465**	**1193**	**588**	**4262**

Note: The bold numbers represent the total number of samples for each class label across all datasets (rows), the total number of samples from each dataset (columns), and the total number of samples from the CSFED dataset (4262).

The proposed methodology relies on static facial images for estimating student engagement, whereas most of the dataset listed in Table 2 contains video files. So, static facial images are extracted by applying pre-processing techniques, explained in Section 4.2.2 on datasets videos. Finally, a balanced hybrid dataset has been built for all six academic affective facial expressions in a total of 4262 colour images of size 48x48 Table 3 provides details of the contributions of various sources in the enlarged CSFED dataset. This CSFED dataset is utilized for training the proposed CNN architecture to develop a FER model.

4.5.1 | Proposed CNN model architecture

FER is one of the tasks that come under image classification, which classifies human affective states through facial features. FER has become one of the most interesting problems in the research community of AC and CV due to its wide range of applications in many domains. In recent years, many works have been performed to identify human affective states from video sources or picture frames (Canedo & Neves, 2019). Some traditional ML techniques, like SVM, K-nearest neighbours, decision trees, etc., are also used to build FER models. However, these techniques are too limited by the traditional handcrafted feature extraction issues. In contrast, the DL-based models are more robust than conventional ML techniques because of their self-feature extraction ability.

The VGG16, ResNet50, and InceptionV3 models are some of the start-of-the-art (SoTA) models for CNN-based image classification tasks utilized for FER. Typically, the SoTA models comprise several deep neural network layers and a fixed input image shape. For example, VGG16 has 16 layers with an input image shape of ($224 \times 224 \times 3$). Therefore, training the SoTA models demands a large dataset comprising good-resolution (at least $224 \times 224 \times 3$) images to achieve good classification accuracy. The CSFED dataset, on the other hand, has size and image resolution constraints as it was created by extracting students' faces from a large real offline classroom's video frames using a typical 2MP wall-mounted surveillance camera. Hence, that prevents it from being used to train the SoTA models for the current work. That is why we chose to construct a simpler architecture by adapting the VGGNet (Simonyan, K et al., 2014) architecture and by reducing its number of layers. This section discusses the proposed CNN architecture for the FER model, as shown in Figure 5.

As shown in Figure 5, the proposed CNN architecture comprises many convolution-2D, max-pooling-2D layers, one flatten layer followed by two densely connected hidden layers, and one softmax layer. The functionality of each of these layers has been already discussed in Section 3.4.1. In the above blueprint, the self-feature extraction part is divided into three blocks. Each block consists of two consecutive convolution layers followed by one max-pooling layer. The first block begins by reading an image of size $48 \times 48 \times 3$ as input and continues into its first convolution

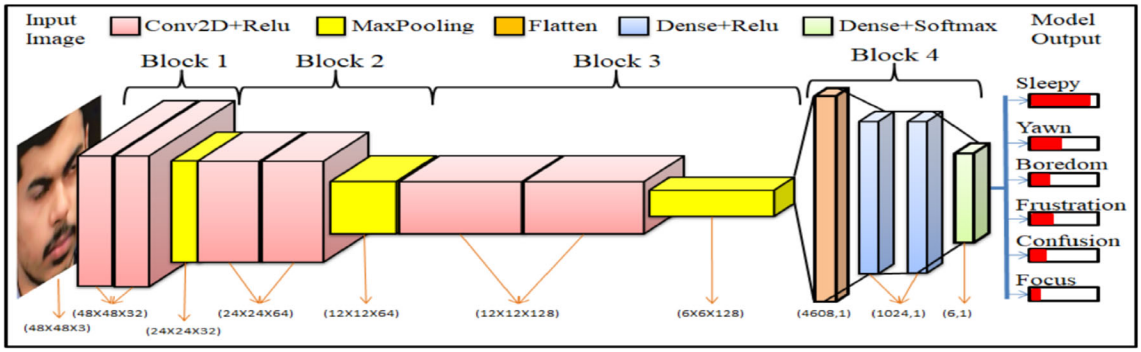

FIGURE 5 The proposed CNN architecture for facial expression recognition model

layer. The first convolution layer comprises 32 kernels of size 3 × 3 each. The kernel weights are initialized by a normal statistical distribution, and kernel regularization is set to L2_regularization. These kernels are convolved on the input image independently for generating 32 non-linear feature maps. A non-linear feature map is generated by applying a non-linear activation function to the calculated dot product from each receptive field in the convolution process. Here the ReLU is used as the non-linear activation function.

Further, these 32 non-linear feature maps are forwarded as input to the second convolution layer. This layer repeats the process similar to the first convolution layer and produces double convolved 32 non-linear feature maps as output. The purpose of double convolution operations is to improve the non-linearity of feature maps. It helps the model to find robust inter-class discriminative feature maps. Next, the max-pooling layer, works on each feature map independently to create a new collection of 32 down-sampled feature maps. The rate of down sampling depends on two parameters, pool kernel size, and stride. Pool kernel is a two-dimensional square matrix that slides on the input image with a given number of pixel shifts called stride value. In each shift, the maximum value is extracted from the respective receptive field. Since pool kernel size and stride are set to 2 × 2 and 2 respectively, the pooled feature maps are down-sampled to half of its original size. After the completion of block1, the model gets 32 feature maps of size 24 × 24 each.

These pooled feature maps are passed as input to the second block. The working process of all the three blocks in the classification section is the same, except the number of kernels used in each block. The second block produces 64 feature maps of size 12 × 12 each. Continuing, the third block accepts these 64 feature maps as input and produces 128 feature maps of size 6 × 6 each. Finally, these two dimensional 128 feature maps are converted into a single column vector of size 4608x1 by the flattening layer. The flattening layer acts as the input layer of CNN's classification block. The classification block contains a fully connected neural network with two hidden layers of size 1024 each and an output layer. The output layer has six neurons (for six emotion classes) as output and uses softmax function to obtain probability score. The internal architecture of the proposed model is shown in Figure 6. As the figure shows, the model is added with batch normalization and dropout layers which are used for faster computation and to avoid overfitting problems respectively.

4.6 | Online module

The next module in the proposed framework is the online module. This online module runs in real-time to estimate students' engagement level for lecture video. It comprises five stages, which are: video acquisition stage, pre-processing stage, students' affective classification stage, post-processing stage, and visualization stage. The overall workflow of the online module is shown in Figure 7. The workflow begins by taking the sequence of the video frame and setting a frame counter to zero. As each video frame is processed into the pre-processing stage, the frame counter is incremented by one. The pre-processing stage returns aligned frontal faces, where affective states from these face images are recognized by the FER model, which was trained in the offline module. Once the frame counter value becomes equal to a predefined threshold, a real-time engagement graph is plotted for the processed video segment by applying post-processing steps (from step-1 to step-4) explained in Section 4.2.4 on recognized affective state labels. After the completion of the lecture, the maximum accumulated group engagement level label is returned as the overall students' engagement feedback for the complete lecture.

4.6.1 | Video acquisition

A classroom lecture video is recorded in the video acquisition stage with a frame rate of 30 frames per second through a 2-megapixel (MP), HD camera. The camera is mounted at the middle position above the blackboard to capture classroom video.

CNN block	CNN Layer	Number of kernels	Input size	Output size
Block 1	Conv2D Layer 1+Input	32	(48,48,3)	(48,48,32)
	BatchNormalization	--	(48,48,32)	(48,48,32)
	Conv2D Layer 2	32	(48,48,32)	(48,48,32)
	BatchNormalization	--	(48,48,32)	(48,48,32)
	MaxPooling2D Layer1	--	(48,48,32)	(24,24,32)
	Dropout Layer 1	--	(24,24,32)	(24,24,32)
Block 2	Conv2D Layer 3	64	(24,24,32)	(24,24,64)
	BatchNormalization	--	(24,24,64)	(24,24,64)
	Conv2D Layer 4	64	(24,24,64)	(24,24,64)
	BatchNormalization	--	(24,24,64)	(24,24,64)
	MaxPooling2D Layer2	--	(24,24,64)	(12,12,64)
	Dropout Layer 2	--	(12,12,64)	(12,12,64)
Block 3	Conv2D Layer 5	128	(12,12,64)	(12,12,128)
	BatchNormalization	--	(12,12,128)	(12,12,128)
	Conv2D Layer 6	128	(12,12,128)	(12,12,128)
	BatchNormalization	--	(12,12,128)	(12,12,128)
	MaxPooling2D Layer3	--	(12,12,128)	(6,6,128)
	Dropout Layer 3	--	(6,6,128)	(6,6,128)
Block 4	Flatten Layer	--	(6,6,128)	(4608)
	Dense Layer 1	--	(4608)	(1024)
	BatchNormalization	--	(1024)	(1024)
	Dropout Layer 4	--	(1024)	(1024)
	Dense Layer 2	--	(1024)	(1024)
	BatchNormalization	--	(1024)	(1024)
	Dropout Layer 5	--	(1024)	(1024)
	Dense Layer 3+Softmax	--	(1024)	(6)

FIGURE 6 Description of the proposed CNN architecture

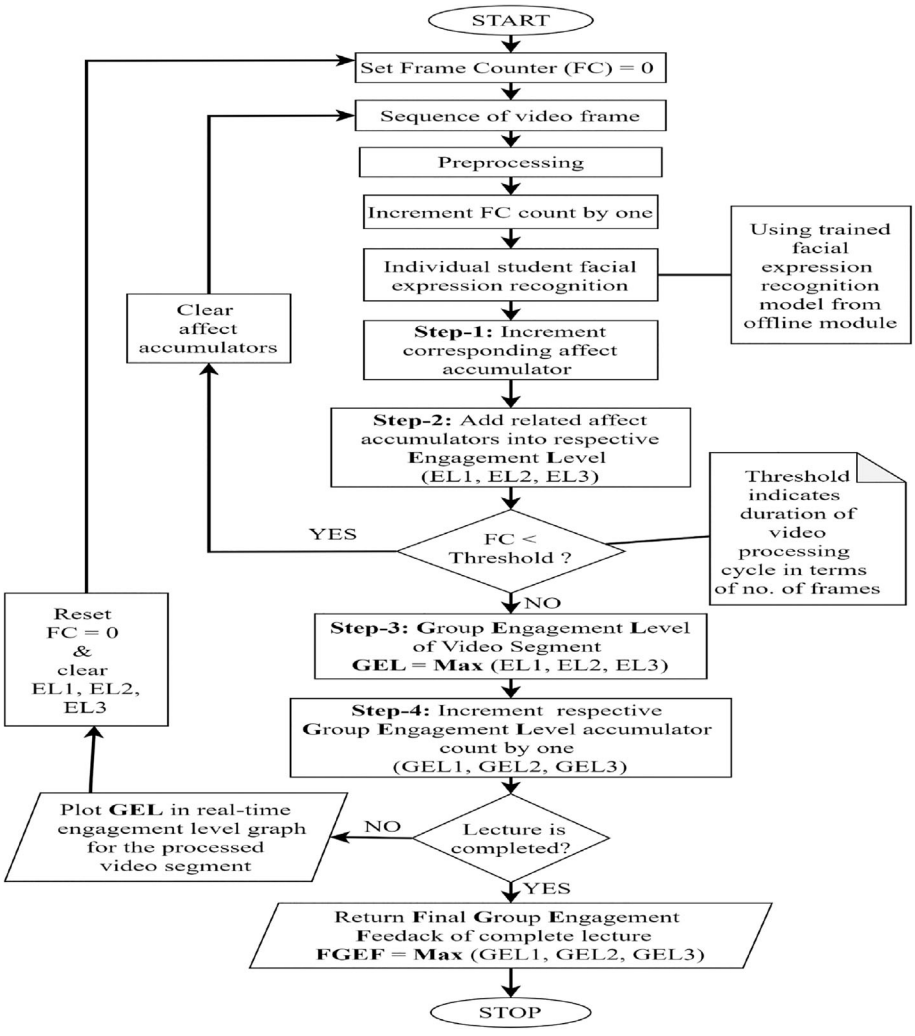

FIGURE 7 Proposed methodology flowchart

4.6.2 | Pre-processing

In this stage, the input video undergoes a set of pre-processing steps; frame sampling, face detection and extraction, head pose estimation, face alignment and resizing, which are described as follows:

4.7 | Frame sampling

The frame sampling step samples a few video frames per second in order to estimate student group engagement level. According to a study by (Whitehill et al., 2014), the result obtained by processing four video frames per second with a time interval of 0.25 s is almost equal to the result which is obtained by processing 30 video frames per second. Therefore, in this frame sampling step, only four video frames are processed per second with a time interval of 0.25 s, thereby reducing computational overhead.

4.8 | Face image extraction

As the input frames are extracted from a classroom video, it is evident that several students will be present in each frame. Hence, it is required to detect and extract student's faces. In the face image extraction step, a pre-trained face detection model is used to extract a maximum number of faces from each video frame. We used the multi-task cascade convolution neural network (MTCNN) as the pre-trained face detection model (Zhang, K., et al. 2016). The MTCNN model achieves state-of-the-art results in detecting even smaller face patches with negligible false-positive results, (where non-face patch is detected as face patch). The face detection step returns a list of face patch coordinates and a list of facial landmarks coordinates. Each face patch coordinate includes four values, which are x and y coordinate values of top-right corner, width and height of the detected face. Using these four values a face image patch is extracted from the video frame. Facial landmarks coordinate list includes coordinate values of two landmarks each at the middle of every eye, one at the nose tip, and another two at mouth corners. Figure 8 shows a sample output of the MTCNN model along with face patch coordinates (15, 12, 18, and 20) and with five facial landmarks.

4.9 | Head pose detection

The head pose detection step removes all non-frontal faces from the detected faces, including left-skewed, right-skewed, upward, and downward. Because the FER model cannot assign appropriate affective state labels to such non-frontal faces, which reduces the proposed method's efficiency. Head pose estimation involves computing the 3D orientation of the head concerning the camera pose using digital images. For this, we implemented a method presented in the work of (Mallick, 2016). In this work, three pose-related degrees, that is, yaw, pitch, and roll, are calculated by associating six 2D facial landmarks (5 landmarks as shown in Figure 8 and the sixth landmark on chin) with respective 3D location landmarks (calculated as per world coordinates). The sixth facial landmark coordinate is explicitly computed using coordinates of the nose tip and mouth corner facial landmarks generated by the MTCNN model. The rotation of an object in vertical movement is called pitch. The rotation of an object in horizontal movement is called yaw. The rotation of an object in a circular (clockwise or anticlockwise) movement is called roll. Left and

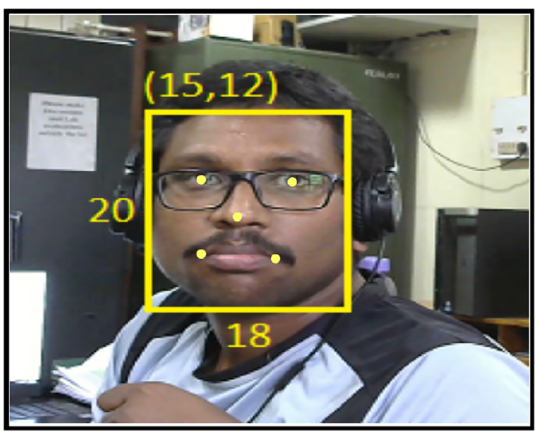

FIGURE 8 MTCNN face detection model output

right-skewed faces were eliminated by a threshold of the degree of yaw movement. Similarly, up and down faces were eliminated by a threshold of the degree of pitch movement. Figure 9 depicts a sample output of the head pose detection.

4.10 | Face alignment

The frontal face patches are further refined in the face alignment step. In general, it is not guaranteed that all frontal faces are exactly aligned as shown in Figure 9a. Some frontal faces may be tilted towards the right or left. Face alignment is a pre-processing technique to obtain a canonical alignment of the face in digital images based on translation, scale, and rotation transformation. For this, we implemented a method presented in the work of (Rosebrock, 2017). Another implicit advantage of this step is image enhancement; it reduces an image blur by re-positioning the pixels. Figures 10a, b show sample output for face alignment and enhancement, respectively. Finally, these aligned frontal faces are resized to 48 × 48 (width × height) pixels then given as an input to the trained FER model.

4.11 | Data augmentation

Data augmentation is a data pre-processing technique used to generate a new set of images by randomly applying different data augmentation techniques on a given image, e.g., rotation, translation, flip, zoom, and more. Generally, data augmentation is used only in the model training phase to address the over-fitting problem. Below are a few most widely used data augmentation techniques. Table 4 shows the data augmentation techniques used in the training of the proposed CNN-based architecture to build FER model.

- Rotation_range: Randomly rotates an image within a certain degree range.
- Shear_range: Randomly apply shearing transformations on the image.
- Width_shift and height_shift range: Randomly translate pictures in both vertical and horizontal directions.
- Zoom_range: Randomly zoom the picture with a given scale.
- Horizontal_flip: Randomly flip input image horizontally.
- Fill_mode: Fill in newly generated pixels that may appear after a rotation or a width/height change.

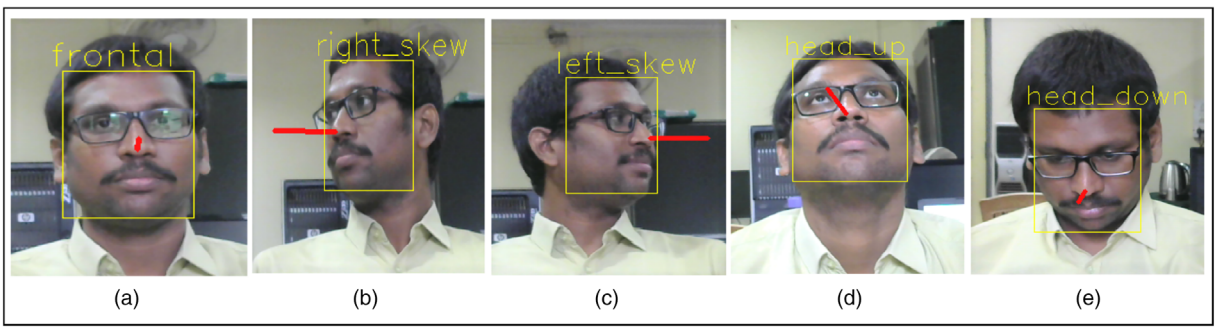

FIGURE 9 Head pose detection output, (a) Frontal face, (b) Right-skewed face, (c) Left-skewed face, (d) Upwards face, (e) Downwards face

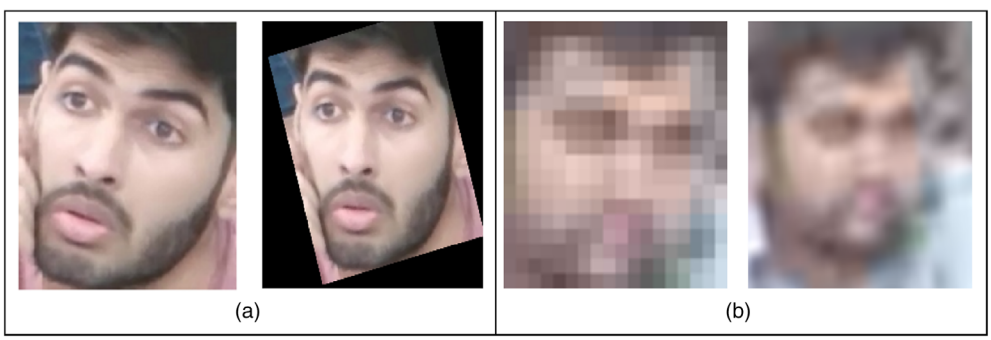

FIGURE 10 Face alignment output (a) Face alignment, (b) Image enhancement

TABLE 4 Types of data augmentation techniques used

Augmentation technique	Augmentation value
rotation_range	20
shear_range	0.2
zoom_range	0.2
width_shift_range	0.2
height_shift_range	0.2
horizontal_flip	True
fill_mode	Nearest

4.11.1 | Student's affective state classification

At this stage, the trained FER model in the offline module is used to recognize the facial expressions of the aligned frontal faces. The predicted expression labels depict the student's affective state or mood while listening to a lecture. These labels are processed into the post-processing stage to estimate overall student group engagement level. From related literature, it has been found that student's engagement is not persistent at the same level throughout the lecture; it transitions into different levels of engagement. Many works (Ashwin & Guddeti, 2019; Whitehill et al., 2014; Zaletelj & Košir, 2017) have used engagement level (EL) taxonomy to classify students' engagement and its transition pattern over time by analysing behavioural cues. This motivated us to use similar EL taxonomy in the proposed work to classify students' academic affective states using facial expressions.

4.12 | Engagement definition and its dimensions

Students' engagement can be measured using classroom recordings, but real engagement cannot be measured without close observation and questioning, which is impractical in large-scale offline classroom settings. Hence, students' affective states (like moods and emotions) analysis is performed using facial expressions to estimate the overall students' engagement. In this work, a *student's engagement* is defined as the degree of a student's emotional engagement while listening to a lecture. In order to classify the level of student's engagement (low [EL1], medium [EL2], and high [EL3]), a local survey was conducted with a group of 100 participants which comprised 15 faculty and 85 students. In the survey form, a sample facial expression image was provided from each academic affective state without specifying its label. Participants were then asked to scale the engagement of each image on a three-level engagement scale. Figure 11 shows the screenshot of the survey form. Figure 12 shows that, out of the 100 votes cast for each expression, 70 were cast for expression 1 with EL3, 65 for expression 2 with EL2, 64 for expression 3 with EL2, 74 for expression 4 with EL1, 62 for expression 5 with EL2, and 80 for expression 6 with EL1. Hence, the six academic affective states' facial expressions labels were categorized into their maximum voted EL group.

- Low engagement (EL1): Boredom, sleepy.
- Medium engagement (EL2): Yawning, frustrated, confused.
- High engagement (EL3): Focused.

4.12.1 | Post-processing

Post-processing is another important step of the proposed methodology. This section discusses the proposed post processing approach to estimate the overall group engagement level in the classroom. It begins by collecting the frame-wise outputs of the trained FER model and gets completed by estimating the group engagement level. Finally, in the visualization stage, the estimated level gets plotted on a real-time engagement graph. As described in Figure 7, the complete post-processing phase is divided into four steps. In step-1, all the predicted affective labels of students faces extracted from a video-frame are accumulated into the respective accumulators (which work as counters for each affective state label). In step-2, these accumulators are merged into the respective EL accumulators, EL1 (low), EL2 (medium), and EL3 (high). This process is repeated until the frame counter becomes equal to the pre-defined threshold. Once the frame counter becomes equal to the pre-defined threshold, the input video stream is divided into a video segment. In step-3, the maximum EL accumulator label is returned as a group engagement level (GEL) of that video segment, like 1 for low, 2 for medium, and 3 for high. In step 4, the GEL label of every video segment is accumulated into the respective GEL accumulators, like GEL1 (low), GEL2 (medium), and GEL3 (high). These accumulators are used to estimate the final group

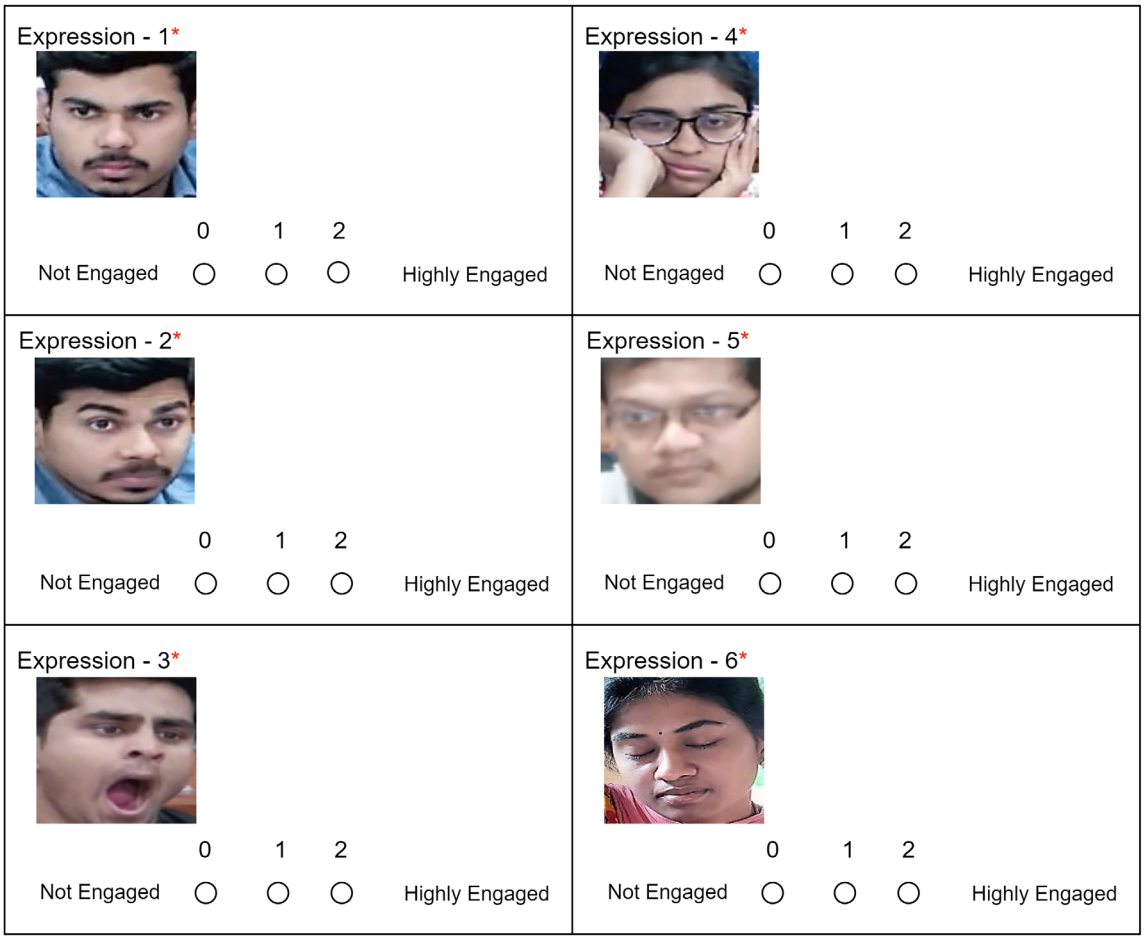

FIGURE 11 Screenshot of the engagement level classification survey form. Facial expressions*: (1) Focused, (2) Confused, (3) Yawning, (4) Boredom, (5) Frustrated, and (6) Sleepy facial expressions

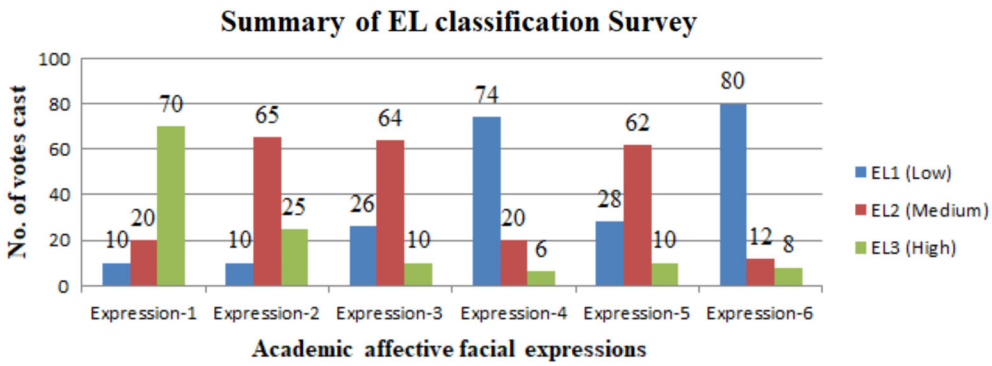

FIGURE 12 Local survey conclusions on engagement level classification, expression- (1) Focused, (2) Confused, (3) Yawning, (4) Boredom, (5) Frustrated, and (6) Sleepy facial expressions

engagement feedback (FGEF) of the complete lecture. After this processing, the processed video segment's GEL is plotted on a real-time graph. Finally when the input video stream is finished, the maximum GEL accumulator label is returned as the FGEF for the complete lecture.

5 | EXPERIMENTAL RESULTS AND DISCUSSIONS

This section describes experimental details such as experimental setup, CNN model training, evaluation and results.

5.1 | Experimental setup

For the implementation and experimentation, a computing system consisting of 8th Generation Intel Core i5-8300H Processor @2.30GHz, 16 Gigabyte (GB) RAM, and 4 GB NVIDIA GeForce GTX 1050Ti graphics card is used. A 2MP (People Link Elite FHD-1080 20X optical zoom) IP camera installed in the smart classroom of the department was used for recording classroom videos.

5.2 | CNN model training and evaluation

5.2.1 | The model training

The CSFED dataset (Section 4.1.1) was used for training the proposed CNN architecture that classifies the academic affective states of the given face expression image. The dataset was divided into two parts, training data (75% of the total data) and testing data (25% of the total data). The test data has been utilized for fine-tuning the hyper-parameters of the model during training. However, the testing data samples are not utilized for training the model. We conducted a series of experiments with different hyper-parameter values to build an efficient model. These hyper-parameters include network parameters like activation function, kernel size, optimizer, loss function, batch size, and more. The optimal hyper-parameter values that were related after many trials are summarized in Table 5.

The proposed CNN architecture was trained on both non-augmented and augmented training data separately. The corresponding accuracy and loss graphs of these two trained models are shown in Figure 13 and Figure 14, respectively.

Figures 13a, b show that the model trained on non-augmented training data achieved less accuracy and high loss on test data when compared to the training data due to the over-fitting problem. Figures 14a, b show that the model trained on augmented training data achieved almost equal accuracy and loss on both training and test data. The observation clearly tells that the second model predictions are more reliable outputs for new data samples than the first and is therefore utilized in the online module of the proposed framework. The training outcomes of both models are tabulated in Table 6.

5.2.2 | The model evaluation

This section discusses the performance evaluation of the trained model. The trained model was assessed by confusion matrix and evaluation metrics like accuracy, precision, recall, and F1-score. The confusion matrix is a table that allows visualization of the performance of the model. It consists of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) terms, which are required to calculate the evaluation metrics. The percentage of correct predictions made by a model is referred to as accuracy. Precision is the ability of a classifier not to label

TABLE 5 List of hyper-parameter values

Hyper-parameter	Value
Batch size	32
Activation function	Rectified Linear Unit (ReLU)
Optimizer	Adam with a learning rate of 0.001
Loss function	categorical_crossentropy
Range of kernels	32–64 – 128
kernel size	3X3
Kernel weight initialiser	he_normal
kernel_regularizer	L2 with a logarithm scale of 0.01
Padding	Zero padding
Stride	Default
Pool size	(2,2)
Normalization	Batch Normalization
Probability of dropout	0.2
Maximum epochs	500
Early stopping	185

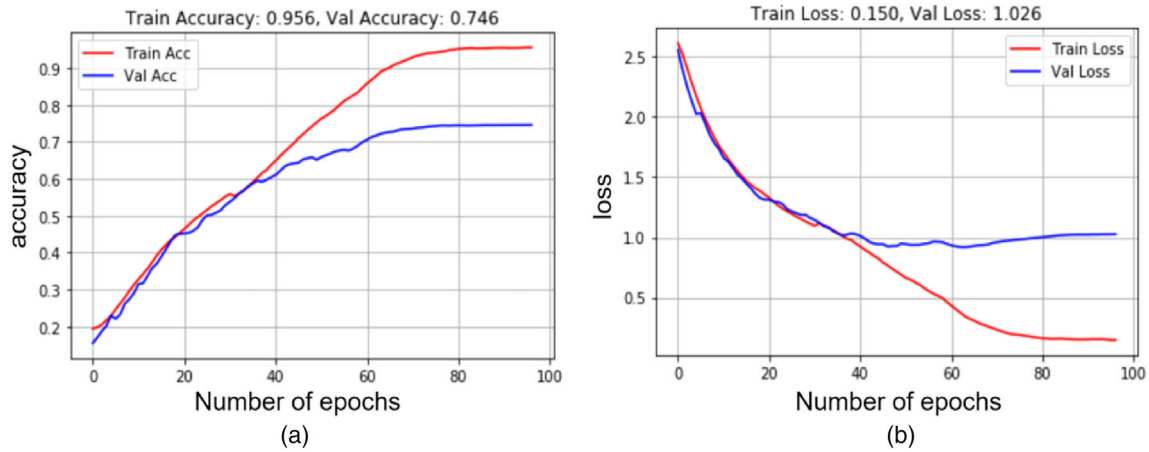

FIGURE 13 Model's learning curves without data augmentation (a) Training and test accuracy graph and (b) Training and test loss graph

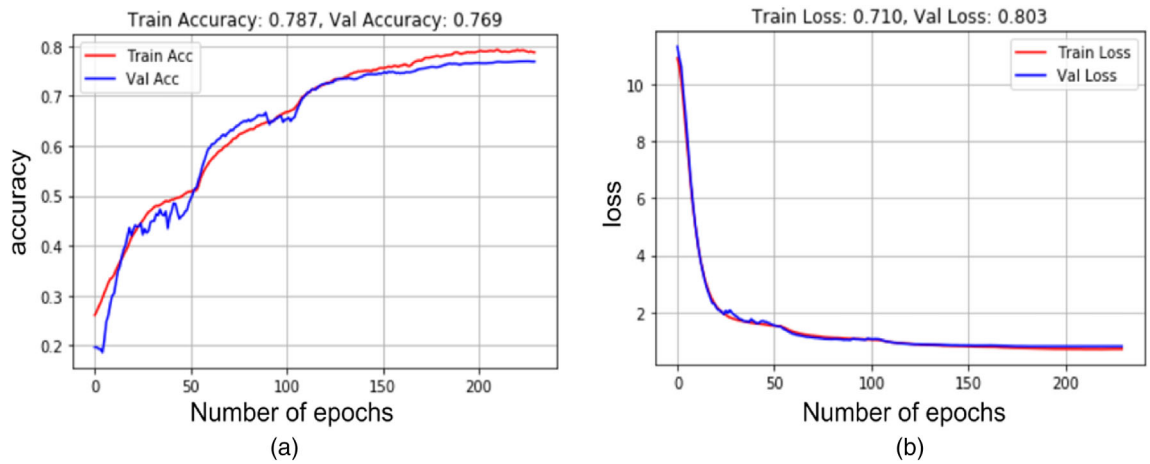

FIGURE 14 Model's learning curves with data augmentation (a) Training and test accuracy graph and (b) Training and test loss graph

TABLE 6 Comparison of trained models evaluation metrics values

Evaluation metrics\trained model	Train accuracy	Test accuracy	Train loss	Test loss
Model without data augmentation	95.60%	74.60%	0.15	1.02
Model with data augmentation	78.70%	76.90%	0.71	0.80

negative instances as positive. The recall is the ability of a classifier to find all positive instances over relevant instances. The F1-score is a weighted harmonic mean of precision and recall. Equations 3, 4, 5, and 6, show the formula to compute the evaluation metrics. The confusion matrix and the evaluation metrics were calculated for the trained model and summarized in Table 7 and Table 8, respectively.

$$Accuracy = ((TP+TN))/((TP+FP+FN+TN)), \tag{3}$$

$$Precision = TP/(TP+FP), \tag{4}$$

$$Recall = TP/(TP+FN), \tag{5}$$

$$F1-Score = (2*(Precision*Recall))/((Precision+Recall)) \tag{6}$$

TABLE 7 Confusion matrix of trained model

	Predicted labels					
Actual labels	**Sleepy**	**Yawning**	**Boredom**	**Frustrated**	**Confuse**	**Focus**
Sleepy	**194**	1	11	0	2	4
Yawning	2	**127**	1	0	2	1
Boredom	6	0	**151**	11	23	26
Frustrated	3	1	21	**87**	29	12
Confuse	1	0	25	17	**142**	23
Focus	1	0	11	3	25	**131**

Note: The bold values represent the trained deep learning model's correct predictions for facial expression recognition in each class.

TABLE 8 Evaluation metrics report of the model

	Precision	**Recall**	**F1-score**	**Support**
Sleepy	0.94	0.92	0.93	212
Yawning	0.98	0.95	0.97	133
Boredom	0.69	0.70	0.69	217
Frustrated	0.74	0.57	0.64	153
Confuse	0.64	0.68	0.66	208
Focus	0.66	0.77	0.71	171
Accuracy	–	–	0.76	1094
macro avg	0.77	0.76	0.77	1094
weighted avg	0.77	0.76	0.76	1094

From Table 7, it is observed that the trained FER model is giving fewer misclassification results for sleepy and yawning affective states compared with other affective classes. This is because these affective states had higher discriminative facial features than other affective states. Whereas there is higher misclassification among other classes due to their facial feature similarities. Since we later group these affectives labels into three engagement levels, the impact of these misclassifications on the overall estimation of GEL is significantly reduced.

From Table 7, it is observed that the model achieved 98%, 94%, 74%, 69%, 66%, and 64% of average precision for unseen instances of yawning, sleepy, focus, boredom, confusion, and frustrated, respectively. We also tested our FER model on a few image samples from one of the popular public available datasets, Imagenet (Deng et al., 2009). Figure 15 shows the result on a sample image, and it can be observed that the model has recognized very close approximate affect states with their facial expressions.

5.3 | Computational time

This section discusses each block's computational times in the online module of the proposed methodology to process a single face based on system specifications, as mentioned in Section 5.1. A sample video with a frame resolution of 1920 × 1080 was used to estimate computational time. It is noted that the video acquisition takes on an average of 18 milliseconds (ms) to capture a single frame. The time required to complete the pre-processing steps depends on the number of faces in the frame. It is noted for a single face that the face detection, head pose estimation, face alignment, and resize steps takes on an average 19.74, 0.44, 0.34, and 0.023 ms, respectively. Next, the trained model takes on an average of 23.22 ms per face to recognize the affective label for a single face. Therefore, in

total, it takes 43.763 ms to process a single face in the frame. The post-processing takes almost negligible time since it is a max function that returns a maximum value index out of three EL values. The total time required for processing a frame is 236–893 ms, depending on the number of faces (from 5 to 20), which corresponds to a video processing rate of 4.23 and 1.12 frames per second. It can be concluded that the proposed methodology is fast enough to predict the academic affective state of students for real-time engagement monitoring. The proposed methodology can also be used with more number of students in a large offline classroom environment to work in real-time using higher computing hardware resources like graphics processing units (GPU) (Jaiswal & Kumar, 2019). The measured computation time of each online module block is reported in Table 9.

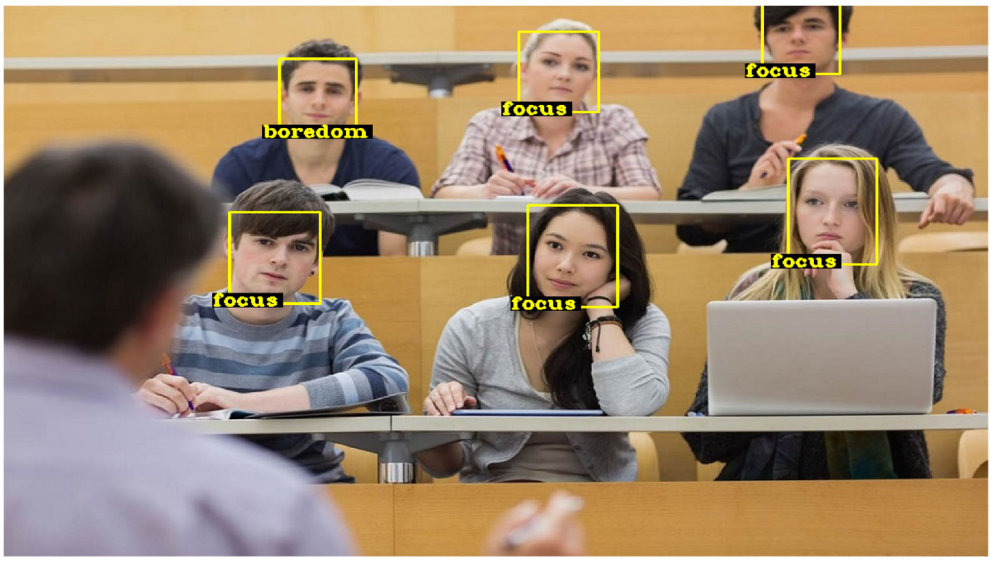

FIGURE 15 The result of the trained FER model on an image of Imagenet dataset

TABLE 9 The computation times of online module blocks averaged over 100 video frames with a single face

Block of online module	Computation time in milliseconds (ms)
Video frame acquisition	18 ms
Face detection	19.74 ms
Head pose estimation	0.44 ms
Face alignment	0.34 ms
Face patch resize	0.023 ms
Affective label recognition	23.22 ms
Total time	43.763 ms

5.4 | Results and discussions

This section shows the result of the overall proposed methodology and validates it by comparing the self-reported ground truth values. Figure 16 shows the screenshot of the pre-processed output of students' video frames in a classroom. As discussed in Section 4.2.2, the pre-processing stage produces only aligned frontal faces from the detected faces.

Figures 17 and 18 show two splits of a single screenshot of the output image frame obtained from the proposed methodology. Here the students' academic affective state labels are recognized by the trained FER (offline module). These model predictions are processed into the post-processing stage as discussed in Section 4.2.4 to compute overall students' engagement level, and that level gets plotted in a real-time engagement graph. The engagement graph is a line graph where time duration (in seconds) is along the X-axis and engagement level (1-low, 2-medium, and 3-high) along the Y-axis.

Two sample video clips of the large classroom were processed to show the real-time engagement level graph generated by the proposed methodology, and a snapshot from those videos is shown in Figure 19a, b, respectively. For the first video, as shown in Figure 19a, the students' engagement is at a low level for the first 10 s, then it increased to medium level and continued up to 130 s; next 10 s it increased to a high level; after that, it drops down to low level. The overall students' engagement for the full video clip is returned as 'medium.' Similarly, students' engagement for the second video is monitored and plotted in the graph as shown in Figure 19b, and it returns the overall students' engagement as 'Low' for the full video clip.

To further validate the proposed methodology's correctness, it was tested on 19 offline classroom lecture video samples of 15 min length. For this purpose, 30 Indian student volunteers (14 female and 16 male) between the ages of 19 and 35 were involved. The 30 volunteers comprised 15 undergraduates, 10 postgraduates, and 5 PhDs from the Computer Science and Engineering department of Visvesvaraya National Institute of Technology (VNIT), INDIA. Informed consent was taken from the volunteers on ethical grounds after outlining the study's goal and purpose. The 30 volunteers were grouped into three groups of 10 people each. The first group comprised 10 undergraduates, the

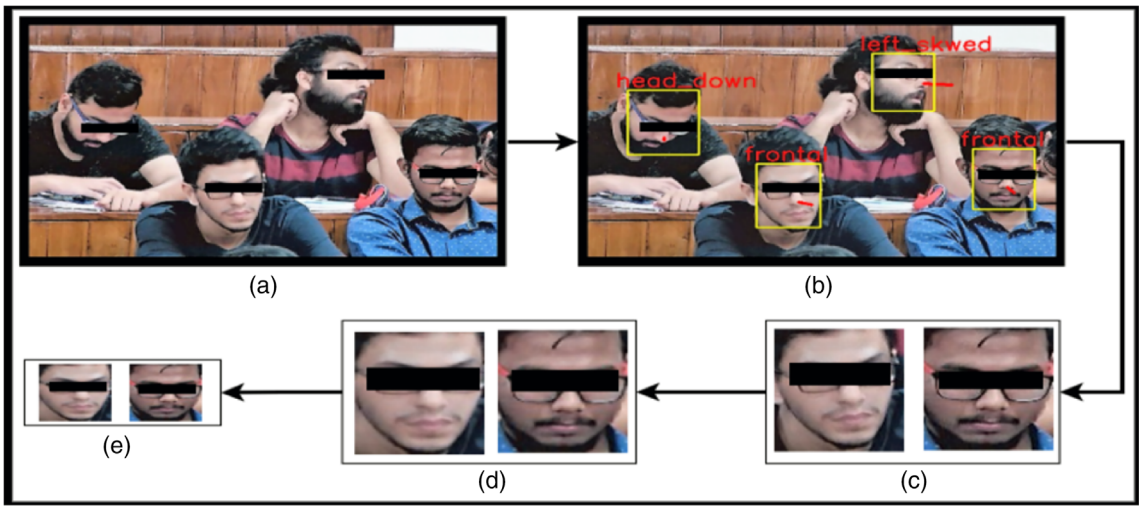

FIGURE 16 Pre-processing techniques used in the proposed methodology. (a) A portion of classroom video, (b) Face detection and head pose estimation (c) Frontal face extraction, (d) Face alignment, (e) Face patch resize to 48×48

FIGURE 17 The left split of the resulting screenshot of the proposed methodology

second group comprised five undergraduates and five postgraduates, and the third group comprised five postgraduates and five doctoral students.

These three groups' visual data have been recorded separately in the five lectures of 60 min each. During these lectures, the students were instructed to note their engagement levels in the given self-report form every 5 min. The self-report form has items for measuring in-class student engagement. To prepare these items, we considered some famous survey-based students' engagement analysis methods like NSSE, SEI. Continuous ground-truth student group engagement levels were calculated for the recorded lectures by analysing and consolidating the submitted self-reports. From these recorded lectures, 19 sample videos were trimmed along with their ground-truth EL reports. Among these 19 videos, six videos were calculated with low EL, seven videos with medium EL, and six videos with high EL as ground truth results. Then, the same videos were also tested with the proposed methodology, and it estimated three videos with low EL, 11 videos with medium EL, and five videos with high EL as predicted results. The ground truth and predicted results were analysed through a confusion matrix, as shown in Table 10.

From Table 10, it is observed that out of a total of 19 videos, the GEL predicted by the system matched with the ground truth values for 14 videos, thereby getting an accuracy of 73.68% using Equation 3. Table 11 compares our proposed methodology with other related works.

FIGURE 18 The right split of the resulting screenshot of the proposed methodology

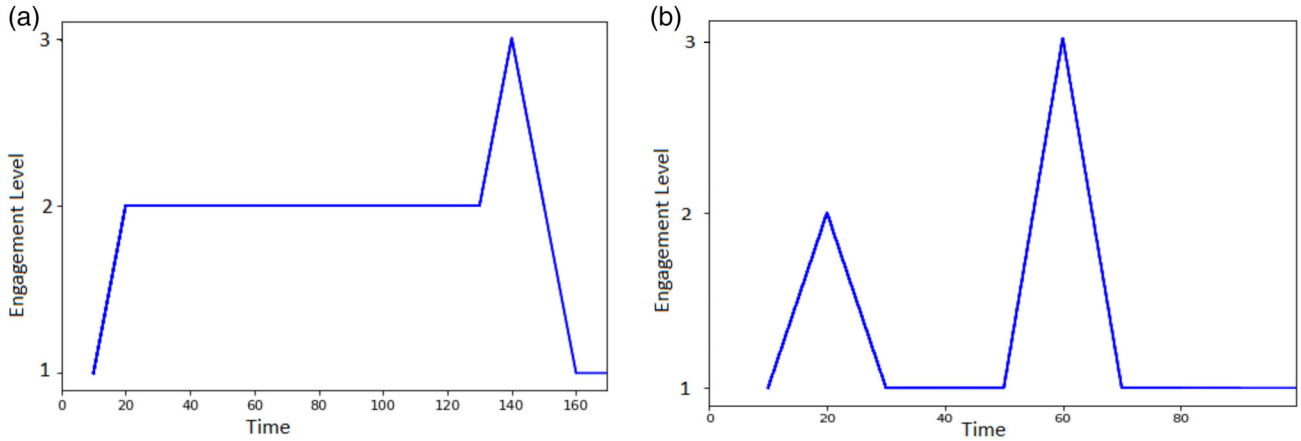

FIGURE 19 Group engagement feedback graphs of sample large classroom video clips through line graphs

5.5 | Limitations

Like most of the related work, the proposed methodology also has some limitations, which are mentioned below. This work is limited to a context where there is no vast degree of heterogeneity in the students' age, culture, and backgrounds in attending the classes. In this case, we assume that the expressions will not vary significantly, and then one model is sufficient to recognize their facial expressions. As a result, the proposed model's operation can work to some extent in the above context, but it could vary when applied on students from different ages, cultures, and backgrounds. However, when the system is deployed in a situation where there is a vast degree of heterogeneity in the student age, culture, and backgrounds, we recommend that multiple FER models be trained considering the different groups of people, and integrate the output of these models to obtain the final result. The current study focuses mainly on facial expressions to estimate students' group engagement. We have not considered the combination of different non-verbal cues like body posture, head movement, and eye gaze. Besides, the proposed study is validated with the students' engagement measure self-reports, but not with the external validations with teachers own judgements, trained observers and students' learning gains.

TABLE 10 Confusion matrix of predicted and actual engagement results

	Predicted results		
Ground truth	**Low**	**Medium**	**High**
Low	**3**	2	1
Medium	0	**7**	0
High	0	2	**4**

Note: The bold values reflect the number of video samples for which the proposed methodology predicted results are equal to the ground truth results.

TABLE 11 Comparison of the proposed methodology with other related works

Work	Academic affective states through facial expressions	Large offline classroom environment	Student group engagement	Real-time processing
Zaletelj and Košir (2017)	x	x	x	x
Klein and Celik (2017)	x	✓	x	x
Thomas and Jayagopi (2017)	x	x	x	x
Soloviev (2018)	x	x	x	x
Ashwin and Guddeti (2019)	x	✓	✓	x
Zheng, R., et al. (2020)	x	✓	x	✓
Luo, Z., et al., (2020)	x	x	x	✓
Peng, S., et al., (2021)	✓	x	x	✓
Vanneste et al., (2021)	x	x	x	✓
Our methodology	✓	✓	✓	✓

6 | CONCLUSION AND FUTURE WORK

This work proposes an automated framework for the development of AEMS to assist teachers in monitoring students' engagement in a large classroom environment in real-time. The framework recognizes the students' facial expressions from lecture video recordings and analyzes them to estimate overall student group engagement. The proposed framework has two modules. The first is an offline module used as a FER model. The trained FER model has achieved train and test accuracy of 78.70% and 76.90%, respectively. The second is an online module that runs in real-time to classify students' engagement into three levels: Low, medium, and high, and plot them on an engagement level graph. The proposed methodology was validated on 19 classroom lecture video recordings and achieved 73.68% correctness. Our findings suggest that using facial expression cues in estimating and monitoring the students' engagement has advantages over using behavioural cues. In a large classroom environment, regardless of student seating arrangement, facial expressions are less occluded than behavioural cues such as body posture. Also, processing these facial expressions takes less computational overhead when compared with processing behavioural cues in order to estimate overall students' engagement. Hence it is feasible to estimate student group engagement in real-time with adequate computational resources. The proposed methodology helps teachers to monitor the whole classroom student group engagement in real-time and make necessary adjustments to the lecture delivery for improving their engagement and learning outcomes. Moreover, it can also be used as a feedback mechanism to get day-to-day lecture feedback effortlessly; such feedbacks are very much useful to novice teachers for improving their instruction strategies in a short time.

Future work includes improving the FER model accuracy by taking the necessary steps like increasing the dataset size. We would also like to improve robustness of the proposed methodology by combining students' few behavioural cues like the head pose and gaze along with academic affective states to estimate overall students' group engagement. Besides, we would also like to externally validate the proposed system with the trained observer and with students' learning gains once the physical classroom teaching resumes after the current pandemic.

ACKNOWLEDGEMENT

The authors would like to thank all the participants, faculty, and staff members of the computer science and engineering department of VNIT who extended their support in conducting this research. We would also like to thank undergraduate students like Deekshith Etyala, Mahendra

Prajapati, and Joshua Reynolds, who assisted in this research. We would also extend our special thanks to Mrs. Sainaba (psychological science researcher at Pondicherry Central University, India), who extended her support in verification of the facial expression annotations.

CONFLICT OF INTEREST

The authors have declared that there is no conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Chakradhar Pabba https://orcid.org/0000-0001-9135-8216

REFERENCES

Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. A. (2019). Applications of artificial intelligence in transport: an overview. *Sustainability*, *11*(1), 189. http://doi.org/10.3390/su11010189

Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., & Hariri, B. (2014). YawDD: A yawning detection dataset. In Proceedings of the 5th ACM multimedia systems conference (pp. 24-28). https://doi.org/10.1145/2557642.2563678

Al-Nawashi, M., Al-Hazaimeh, O. M., & Saraee, M. (2017). A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments. *Neural Computing and Applications*, *28*(S1), 565–572. http://doi.org/10.1007/s00521-016-2363-z

Amilpur, S., & Bhukya, R. (2020). EDeepSSP: Explainable deep neural networks for exact splice sites prediction. *Journal of Bioinformatics and Computational Biology*, *18*(04), 2050024. http://doi.org/10.1142/s0219720020500249

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, *44*(5), 427–445. http://doi.org/10.1016/j.jsp.2006.04.002

Ashwin, T. S., & Guddeti, R. M. R. (2019). Unobtrusive behavioral analysis of students in classroom environment using non-verbal cues. *IEEE Access*, *7*, 150693–150709. http://doi.org/10.1109/access.2019.2947519

Ashwin, T. S., & Guddeti, R. M. R. (2020). Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures. *Future Generation Computer Systems*, *108*, 334–348. http://doi.org/10.1016/j.future.2020.02.075

Baltrušaitis, T., Robinson, P., & Morency, L. (2016). Openface: An open source facial behavior analysis toolkit. IEEE Winter Conference on Applications of Computer Vision (Wacv), 1–10. http://doi.org/10.1109/WACV.2016.7477553

Bhardwaj, P., Gupta, P. K., Panwar, H., Siddiqui, M. K., Morales-Menendez, R., & Bhaik, A. (2021). Application of Deep Learning on Student Engagement in e-learning environments. *Computers & Electrical Engineering*, *93*, 107277. http://doi.org/10.1016/j.compeleceng.2021.107277

Bian, C., Zhang, Y., Yang, F., Bi, W., & Lu, W. (2019). Spontaneous facial expression database for academic emotion inference in online learning. *IET Computer Vision*, *13*(3), 329–337. http://doi.org/10.1049/iet-cvi.2018.5281

Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using Video to Automatically Detect Learner Affect in Computer-Enabled Classrooms. *ACM Transactions on Interactive Intelligent Systems*, *6*(2), 1–26. http://doi.org/10.1145/2946837

Bradbury, N. A. (2016). Attention span during lectures: 8 seconds, 10 minutes, or more?. *Advances in Physiology Education*, *40*(4), 509–513. http://doi.org/10.1152/advan.00109.2016

Canedo, D., & Neves, A. J. (2019). Facial Expression Recognition Using Computer Vision: A Systematic Review. *Applied Sciences*, *9*(21), 4678. http://doi.org/10.3390/app9214678

Chakole, J., & Kurhekar, M. (2020). Trend following deep Q-Learning strategy for stock trading. *Expert Systems*, *37*(4). http://doi.org/10.1111/exsy.12514

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *arXiv*, 1405.3531. *abs/1405.3531*, http://arxiv.org/abs/1405.3531

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*, *8*, 75264–75278. http://doi.org/10.1109/access.2020.2988510

Dasari, C. M., & Bhukya, R. (2020). InterSSPP: Investigating patterns through interpretable deep neural networks for accurate splice signal prediction. *Chemometrics and Intelligent Laboratory Systems*, *206*, 104–144. http://doi.org/10.1016/j.chemolab.2020.104144

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. http://doi.org/10.7861/futurehosp.6-2-94

De Villiers, B., & Werner, A. (2016). The relationship between student engagement and academic success. *Journal for New Generation Sciences*, *14*(1), 36–50.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

Di Lascio, E., Gashi, S., & Santini, S. (2018). Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–21. http://doi.org/10.1145/3264913

Dirican, A.C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science*, *3*, 1361–1367. http://doi.org/10.1016/j.procs.2011.01.016

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, *105*(4), 1082–1099. http://doi.org/10.1037/a0032674

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, *22*(4), 53–61. http://doi.org/10.1109/mis.2007.79

D'Mello, S. K., Lehman, B., & Person, N. (2010). Monitoring affect states during effortful problem solving activities. *International Journal of Artificial Intelligence in Education*, *20*(4), 361–389. 10.3233/JAI-2010-012

Dzelzkaleja, L., & Kapenieks, J. (2016). Real-time color codes for assessing learning process. *Procedia - Social and Behavioral Sciences*, *231*, 263–269. http://doi.org/10.1016/j.sbspro.2016.09.101

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3-4), 169–200. http://doi.org/10.1080/02699939208411068

Exeter, D. J., Ameratunga, S., Ratima, M., Morton, S., Dickson, M., Hsu, D., & Jackson, R. (2010). Student engagement in very large classes: the teachers' perspective. *Studies in Higher Education*, *35*(7), 761–775. http://doi.org/10.1080/03075070903545058

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. http://doi.org/10.3102/00346543074001059

Gretchen Greene, M. J., & Phil, C. (2020). The ethics of ai and emotional intelligence. https://www.partnershiponai.org/wp-content/uploads/2020/07/PAI_The-ethics-of-AI-and-emotional-intelligence_073020.pdf.

Griol, D., Molina, J. M., & Callejas, Z. (2017). Incorporating android conversational agents in m-learning apps. *Expert Systems*, *34*(4), e12156. http://doi.org/10.1111/exsy.12156

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *Journal of Latex Class Files*. http://arxiv.org/abs/1609.01885

Haddad, R., & Youakim, K. (2014). Google forms: A real-time formative assessment approach for adaptive learning. *Proceedings of the 2014 American Society for Engineering Education, ASEE Annual Conference and Exposition Indiana*. https://peer.asee.org/20540

Jaiswal, D., & Kumar, P. (2020). Real-time implementation of moving object detection in UAV videos using GPUs. *Journal of Real-Time Image Processing*, *17*(5), 1301–1317. http://doi.org/10.1007/s11554-019-00888-5

Joshi, S., Verma, D. K., Saxena, G., & Paraye, A. (2019) Issues in Training a Convolutional Neural Network Model for Image Classification. In: Singh M., Gupta P., Tyagi V., Flusser J., Ören T., Kashyap R. (eds), *Advances in Computing and Data Sciences. ICACDS 2019. Communications in Computer and Information Science*, (Vol. 1046). Springer, Singapore. https://doi.org/10.1007/978-981-13-9942-8_27

Kamath, A., Biswas, A., & Balasubramanian, V. (2016). A crowdsourced approach to student engagement recognition in e-learning environments. *2016 IEEE Winter Conference on Applications of Computer Vision (wacv)*, 1–9. https://doi.org/10.1109/WACV.2016.7477618

Khaire, P., Kumar, P., & Imran, J. (2018). Combining CNN streams of RGB-D and skeletal data for human activity recognition. *Pattern Recognition Letters*, *115*, 107–116. http://doi.org/10.1016/j.patrec.2018.04.035

Kim, Y., Soyata, T., & Behnagh, R. F. (2018). Towards emotionally aware ai smart classroom: current issues and directions for engineering and education. *IEEE Access*, *6*, 5308–5331. http://doi.org/10.1109/access.2018.2791861

Klein, R., & Celik, T. (2017). The wits intelligent teaching system: Detecting student engagement during lectures using convolutional neural networks. In *2017 IEEE International Conference on Image Processing (icip)* (pp. 2856–2860). http://doi.org/10.1109/ICIP.2017.8296804

Krithika, L. B., & Lakshmi, P. G. G. (2016). Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. *Procedia Computer Science*, *85*, 767–776. http://doi.org/10.1016/j.procs.2016.05.264

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. http://doi.org/10.1145/3065386

Kuh, G. D. (2003). What we're learning about student engagement from nsse: benchmarks for effective educational practices. *Change: The Magazine of Higher Learning*, *35*(2), 24–32. http://doi.org/10.1080/00091380309604090

Kumar, A., Kaur, A., & Kumar, M. (2019). Face detection techniques: a review. *Artificial Intelligence Review*, *52*(2), 927–948. http://doi.org/10.1007/s10462-018-9650-2

Lamba, M. S., Rawat, M. A., Jacob, M. J., Arya, M. M., Rawat, M. J., Chauhan, M. V., & Panchal, M. S. (2014). Impact of teaching time on attention and concentration. *IOSR Journal of Nursing and Health Science*, *3*(4), 1–4.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. http://doi.org/10.1038/nature14539

Li, Y., & Lerner, R. M. (2013). Interrelations of behavioral, emotional, and cognitive school engagement in high school students. *Journal of Youth and Adolescence*, *42*(1), 20–32. http://doi.org/10.1007/s10964-012-9857-5

Luo, Z., Jingying, C., Guangshuai, W., & Mengyi, L. (2020). A three-dimensional model of student interest during learning using multimodal fusion with natural sensing technology. *Interactive Learning Environments*, 1–14. http://doi.org/10.1080/10494820.2019.1710852

Mallick, S. (2016). Head pose estimation using opencv and dlib, *26*, 2016. https://www.learnopencv.com/head-pose-estimation-using-opencv-and-dlib/

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. http://doi.org/10.1007/bf02478259

Mogwe, A. W. (2018). Real-time concept feedback in lectures for botho university students. *International Journal of Advanced Computer Science and Applications*, *9*(6), 246–251. http://doi.org/10.14569/ijacsa.2018.090635

Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. *2016 IEEE Winter Conference on Applications of Computer Vision (wacv)*, 1–10. http://doi.org/10.1109/WACV.2016.7477450

Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P. K. D., Dasgupta, N., & Choudhury, P. (2020). Facial emotion detection to assess learner's state of mind in an online learning system. *Proceedings of the 2020 5th International Conference on Intelligent Information Technology*, 107–115. https://doi.org/10.1145/3385209.3385231

Odiri Amatari, V. (2015). The instructional process: a review of flanders' interaction analysis in a classroom setting. *International Journal of Secondary Education*, *3*(5), 43–49. http://doi.org/10.11648/j.ijsedu.20150305.11

Pekrun, R. (2000). A social-cognitive, control-value theory of achievement emotions. In J. Heckhausen (Ed.), *Advances in psychology* (pp. 143–163). Elsevier Science. https://doi.org/10.1016/S0166-4115(00)80010-2

Peng, S., & Nagao, K. (2021). Recognition of students' mental states in discussion based on multimodal data and its application to educational support. *IEEE Access*, *9*, 18235–18250. http://doi.org/10.1109/access.2021.3054176

Pons, G., & Masip, D. (2018). Supervised committee of convolutional neural networks in automated facial expression analysis. *IEEE Transactions on Affective Computing*, *9*(3), 343–350. http://doi.org/10.1109/taffc.2017.2753235

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125. http://doi.org/10.1016/j.inffus.2017.02.003

Ramón, Z.-C., Lucia, B.-E. M., Daniel, L.-H., & Mario, R.-F. J. (2017). Creation of a facial expression corpus from eeg signals for learning centered emotions. *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, 387–390. https://doi.org/10.1109/ICALT.2017.135

Robin Murdoch, P. J. M. F., & Agneta, B.. (2020). Getting emotional. https://www.accenture.com/_acnmedia/PDF-114/Accenture-Responsible-Use-Of-Emotional-AI-Final.pdf.

Rosebrock, A. (2017). Face alignment with opencv and python, *22*, 2017. https://www.pyimagesearch.com/2017/05/22/face-alignment-with-opencv-and-python/

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2018). Improving the prediction of learning outcomes in educational platforms including higher level interaction indicators. *Expert Systems*, *35*(6), e12298. http://doi.org/10.1111/exsy.12298

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. http://doi.org/10.1038/323533a0

Sathik, M., & Jonathan, S. G. (2013). Effect of facial expressions on student's comprehension recognition in virtual educational environments. *SpringerPlus*, *2*(1). http://doi.org/10.1186/2193-1801-2-455

Sharma, P., Joshi, S., Gautam, S., Filipe, V., & Reis, M. J. (2019). Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. *CORR*, 1909.12913. *abs-1909-12913*, http://arxiv.org/abs/1909.12913

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *In Proc. International Conference on Learning Representations*, 1409.1556. http://arxiv.org/abs/1409.1556

Singh, J., & Goyal, G. (2021). Decoding depressive disorder using computer vision. *Multimedia Tools and Applications*, *80*(6), 8189–8212. http://doi.org/10.1007/s11042-020-10128-9

Soloviev, V. (2018). Machine learning approach for student engagement automatic recognition from facial expressions. *Scientific Publications of the State University of Novi Pazar Series A: Applied Mathematics, Informatics and mechanics*, *10*(2), 79–86. http://doi.org/10.5937/spsunp1802079s

Thomas, C., & Jayagopi, D. B. (2017). Predicting student engagement in classrooms using facial behavioral cues. *Proceedings of the 1st ACM Sigchi International Workshop on Multimodal Interaction for Education*, 33–40. https://doi.org/10.1145/3139513.3139514

Tonguç, G., & Ozaydın Ozkara, B. (2020). Automatic recognition of student emotions from facial expressions during a lecture. *Computers & Education*, *148*, 103797. http://doi.org/10.1016/j.compedu.2019.103797

Vanneste, P., Oramas, J., Verelst, T., Tuytelaars, T., Raes, A., Depaepe, F., & Van den Noortgate, W. (2021). Computer vision and human behaviour, emotion and cognition detection: a use case on student engagement. *Mathematics*, *9*(3), 287. http://doi.org/10.3390/math9030287

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, *1*.I-511–I-518. 10.1109/CVPR.2001.990517

Wang, S., & Ji, Q. (2015). Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing*, *6*(4), 410–430. http://doi.org/10.1109/taffc.2015.2432791

Wei, Q., Sun, B., He, J., & Yu, L. (2017). BNU-LSVED 2.0: Spontaneous multimodal student affect database with multi-dimensional labels. *Signal Processing: Image Communication*, *59*, 168–181. http://doi.org/10.1016/j.image.2017.08.012

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86–98. http://doi.org/10.1109/taffc.2014.2316163

Wiley, V., & Lucas, T. (2018). Computer vision and image processing: a paper review. *International Journal of Artificial Intelligence Research*, *2*(1), 22. http://doi.org/10.29099/ijair.v2i1.42

Yolcu, G., Oztel, I., Kazan, S., Oz, C., & Bunyak, F. (2020). Deep learning-based face analysis system for monitoring customer interest. *Journal of Ambient Intelligence and Humanized Computing*, *11*(1), 237–248. http://doi.org/10.1007/s12652-019-01310-5

Zaletelj, J., & Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. *EURASIP Journal on Image and Video Processing*, *2017*(1). https://doi.org/10.1186/s13640-017-0228-8

Zhalehpour, S., Onder, O., Akhtar, Z., & Erdem, C. E. (2017). BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Transactions on Affective Computing*, *8*(3), 300–313. http://doi.org/10.1109/taffc.2016.2553038

Zhang, H., Xiao, X., Huang, T., Liu, S., Xia, Y., & Li, J. (2019). An novel end-to-end network for automatic student engagement recognition. *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 342–345. https://doi.org/10.1109/ICEIEC.2019.8784507

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. http://doi.org/10.1109/lsp.2016.2603342

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE Multimedia*, *19*(2), 4–10. http://doi.org/10.1109/mmul.2012.24

Zheng, R., Jiang, F., & Shen, R. (2020). Intelligent student behavior analysis system for real classrooms. *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 9244–9248. https://doi.org/10.1109/ICASSP40776.2020.9053457

AUTHOR BIOGRAPHIES

Chakradhar Pabba received a Bachelor of Technology degree in Computer Science and Engineering (CSE) from JNTU, Hyderabad in 2008, and a Master of Technology degree in Computer Science from the National Institute of Technology, Trichy in 2010. He is currently pursuing a Ph.D. in CSE at the Visvesvaraya National Institute of Technology, Nagpur, India. His research interests include affective computing, deep learning, and computer vision.

Praveen Kumar is an assistant professor in the CSE department at VNIT in Nagpur, India. In 2005, he received his B.Tech in CSE from the Indian Institute of Technology Roorkee. In 2008, he received his Ph.D. from the same institute. Following that, he worked as a postdoctoral fellow at the University of Missouri-Columbia, USA. Dr. Kumar's research interests are in parallel computation, deep learning, and intelligent

video surveillance. He has many papers published in international journals and top conferences. For his research and technical achievements, he has won awards from CSTUP, INAE, IEEE ICISIP conference, AICTE.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Pabba, C., & Kumar, P. (2022). An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems*, *39*(1), e12839. https://doi.org/10.1111/exsy.12839