Educational and Information Technologies (2020) 25: 1387–1415
https://doi.org/10.1007/s10639-019-10004-6

# Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks

**Ashwin T. S.**[1] · **Ram Mohana Reddy Guddeti**[1]

Received: 15 May 2019 / Accepted: 2 September 2019 / Published online: 28 October 2019


## Abstract

Predicting the students' emotional and behavioral engagements using computer vision techniques is a challenging task. Though there are several state-of-the-art techniques for analyzing a student's affective states in an e-learning environment (single person's engagement detection in a single image frame), a very few works are available for analyzing the students' affective states in a classroom environment (multiple people in a single image frame). Hence, in this paper, we propose a novel hybrid convolutional neural network (CNN) architecture for analyzing the students' affective states in a classroom environment. This proposed architecture consists of two models, the first model (CNN-1) is designed to analyze the affective states of a single student in a single image frame and the second model (CNN-2) uses multiple students in a single image frame. Thus, our proposed hybrid architecture predicts the overall affective state of the entire class. The proposed architecture uses the students' facial expressions, hand gestures and body postures for analyzing their affective states. Further, due to unavailability of standard datasets for the students' affective state analysis, we created, annotated and tested on our dataset of over 8000 single face in a single image frame and 12000 multiple faces in a single image frame with three different affective states, namely: engaged, boredom and neutral. The experimental results demonstrate an accuracy of 86% and 70% for posed and spontaneous affective states of classroom data, respectively.

**Keywords** Affective computing · Affective states · Convolutional neural network · Classroom environemnt · Facial emotion recognition · Student engagement

✉ Ashwin T. S.
ashwindixit9@gmail.com

Ram Mohana Reddy Guddeti
profgrmreddy@nitk.edu.in

[1] Information Technology Department, National Institute of Technology Karnataka Surathkal, Mangalore, Dakshina Kannada, Karnataka, India

## 1 Introduction

A student's engagement plays a vital role in the teaching-learning process as it is closely associated with the learning rate (D'Mello et al. 2010). The students' engagement is broadly classified into four categories, namely: emotional engagement, behavioral engagement, cognitive engagement, and agentic engagement (Sinclair et al. 2015). The students' affective state recognition can be performed using various techniques such as speech/voice recognition, physiological sensors like heart rate, and pressure sensors (Wang and Ji 2015; DeFalco et al. 2018; Dermeval et al. 2018; Bosch and D'Mello 2017). Using these physiological sensors, the students' affective state analysis becomes obtrusive, and the deployment cost can be high if affective state recognition is performed for all the students present in a classroom. In any synchronous learning environments like a classroom, each student may not get the opportunity to interact with the teacher. Hence, voice/speech recognition for affective state recognition is not recommended. To overcome these limitations, an unobtrusive technique is required for the students' affective state recognition in the classroom environment using non-verbal cues like facial expressions, body postures, and hand gestures captured from the video/image frames from a synchronous learning environment.

There exist several works on emotion detection from video streams or image frames. Machine learning and deep learning techniques are extensively explored in the area of computer vision. Neural network based machine learning techniques are generally used to predict the emotions of students. There are several emotion prediction techniques using SVM (support vector machines), KNN (K-Nearest Neighbor), decision tree learning, association rule learning, rule-based machine learning, etc. (Ashwin et al. 2015; Pao et al. 2005; Girshick et al. 2016). These traditional machine learning approaches use the handcrafted/predefined features for classifying the given input data. For example, predefined features known as AUs (action units) are used to recognize facial expressions. There are 28 AUs corresponding to facial muscles such as inner brow raise, outer brow raise, and so on. A set of action units corresponds to each emotion (happiness should have 6th and 12th AUs only, sadness should have 1st, 4th, and 15th AUs only, etc.). If there is a slight mismatch in the set of action units for the given input data, the algorithm fails to perform better. Hence, these techniques are not robust because of their predefined (handcrafted) features. Whereas, the deep learning method uses the feature learning method where the algorithm extracts the features by itself from a given labeled data. Currently, there are many deep learning based algorithms with better results for pattern recognition problems like computer vision where it learns the features by itself. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are used for multi-modal human action recognition (Krizhevsky et al. 2012; Szegedy et al. 2016). Single-shot multi-box detectors are used for multiple people recognition with object localization and are tested on the dataset of images in the wild (Liu et al. 2016). However, these techniques are not explored for the students' affective content analysis.

An automatic prediction and analysis of a student's engagement is a challenging task. There are several works conducted on recognition of emotional engagement using video affective content analysis in the last few years, and significant progress

has been made in this direction, but it is still in the infancy stage (Wang and Ji 2015; Worsley and Blikstein 2018). This is due to various challenges such as feature representation (i.e., extraction) for characterizing affective video content, emotion annotation for emotional descriptors, visual behavior response or measurement of user's physiological relations with video content and a user's response (Picard and Picard 1997; Kleinsmith and Bianchi-Berthouze 2013). To analyze different affective states, we need a robust technique to solve all corner cases of the students' affective state analysis.

Automatic recognition of the students' affective states helps in analyzing their engagement in intelligent tutoring systems (Sidney et al. 2005). Monitoring both the students' emotional & behavioral engagements and accordingly classifying their affective states is crucial. Facial expressions alone are not sufficient to recognize both the emotional and behavioral engagements of students. There exist a very few algorithms to recognize human emotions from facial expressions, hand gestures, and body postures. The behavioral engagements like eyes barely open, looking away from the system are analyzed using face, hand gesture, and body postures. But the combination of both emotional and behavioral patterns of students to recognize an affective state is not considered in a classroom scenario. There are several methods which recognize the face with high accuracy (Simonyan and Zisserman 2014; Szegedy et al. 2016) but there are limited algorithms which analyze the relationship between low-level features and users' emotional response from their faces. Still, there exists no standard method to recognize the students' affective state for classroom environment based on facial expression, hand gesture, and body posture.

Further, there are no standard datasets available for classroom environment to predict the affective state of students. Though there are several works on computer vision based affective state recognition of students in the e-learning environment, there are limited works on affective state recognition of students in the classroom environment where multiple students are present in a single image frame. This motivates us to address the following issues: (a) Predicting a group engagement score of an entire class; (b) Combining both emotional and behavioral engagements to predict the affective state of the students; (c) Multi-modal affective state recognition of students using their facial expressions, hand gestures and body postures; (d) Recognizing every student present in the single frame with object localization to predict the affective state patterns in the wild, and (e) Creating a standard dataset to train the existing learning techniques for the better design and implementation of predicting the students' affective states.

Hence, we propose a new method using CNN to predict three different affective states of students, namely: engaged, boredom and neutral for the classroom environment. We proposed a hybrid CNN architecture which is a combination of two different CNN models for training and testing of classroom data. The first model (CNN-1) is trained to detect a single person in a single image frame, and the second model (CNN-2) is trained to detect multiple people in a single image frame.

The key contributions of this paper are as follows:

- A novel hybrid CNN model for affective state analysis of students in the classroom environment using,

  - students' facial expression, hand gestures and body postures,
  - group engagement (class) score (for each frame with spontaneous expression (multi-person in a single frame image)), and

- Prediction of affective states using both the emotional and behavioral patterns to classify them into engaged and boredom classes along with neutral.

Rest of the paper is organized as follows. The literature survey is covered in Section 2, followed by Proposed Methodology in Section 3. Section 4 discusses the Results and Analysis. Finally, Section 5 concludes with future directions.

## 2 Literature survey

### 2.1 Students' affective state classification

There are several works on the student's affective state classification. Intelligent Tutoring System (ITS), auto tutor, humanoid robots, and others use the students' facial expressions and other body parts related to behavioral aspects for recognizing their affective states. Learning-centered emotions like anger, boredom, confusion, contempt, curiosity, disgust, eureka and frustrations were used in the Emote-aloud study (Sidney et al. 2005). Constructive and destructive learning affective states like frustration, confused, happy and hopeless were considered for automatic affective state recognition in humanoid robot tutors (Singh et al. 2013). Whitehill et al. (2014) proposed four engagement levels which were mainly behavioral engagements to analyze the students' engagement in a laboratory environment. Zhang et al. (2018) proposed a framework to analyze the students' performance using their behavioral patterns such as buying materials, library entrance, school bus usage, etc. Kim (2018) used students' thermal images to recognize five different engagement levels using fuzzy systems. Holmes et al. (2018) proposed a real-time nonverbal behavior recognition of e-learners. They considered 37 behavioral patterns and analyzed the student's affective state as positive or negative. Further, D'Mello et al. (2007) showed that only confusion, boredom, frustration, delight, and flow emotions were sufficient to analyze the students' emotional engagement in any learning environment. But, this entire work is performed on a single student in a single image frame. Table 1 summarizes the recent work on affective state classification based on the affective descriptors.

### 2.2 Facial emotion recognition

Several works are available for the classification of categorical facial emotion descriptors, feature selection, feature extraction and classification of emotions (Wang and Ji 2015; Pao et al. 2005; Thomas and Jayagopi 2017). Ortony et al. (1990) proposed an Ortony Clore and Collins (OCC) model with five basic emotions (anger, fear, happiness, joy, & love). Happy et al. (2015) described a model consisting of 10 basic emotions (anger, happiness, fear, sadness, surprise, disgust, love, anticipation, joy, and trust). But few of these emotions were not feasible for learning environments.

**Table 1** Summary of existing works on affective state classification

| Authors | Methodology | Merits | Affective states | Limitations | Environment |
|---|---|---|---|---|---|
| Thomas and Jayagopi (2017) | Facial behavioral cues are used | Classified student's engagement into engaged and distracted | Engaged and distracted | Only frontal face is considered | Classroom |
| D'Mello et al. (2007) | Affective sensitive auto tutor | Classified student's engagement into boredom, confusion, delight, flow and frustration | Learning-centered emotions | Posture analysis using sensors on every chair makes it obtrusive | Auto tutor |
| Balaam et al. (2010) | Subtle Stone is used for engagement analysis | Tangible technology designed to support student's active emotional communication | Ekman's basic emotions | Use of Subtle Stone makes it obtrusive | Classroom |
| Zaletelj and Košir (2017) | Kinect based system for student's engagement | Student's attention monitoring during a lecture using gaze and behavioral cues | Behavior patterns | The range of Kinect is small and cannot be used in large classrooms | Classroom |
| D'Mello et al. (2010) | Monitoring affective states | Temporal dynamics of affective states are analyzed | Learning-centered emotions | Tested for problem solving activities only | E-Learning |
| Klein and Celik (2017) | The WITS intelligent tutoring system | Student's emotion recognition using CNN | Interested and Not-interested | Only behavioural patterns are considered | Classroom |
| Mirko and Dillenbourg (2015) | Students body language | Works for multiple students in a single image frame | low, medium and high attention levels | Only head movement patterns are considered | Classroom |

Chen and Luo (2006) developed a single face emotion recognition system using Euclidean distance in 6D space. On the similar lines, Aiqin and Luo (2006) proposed a single face expression recognition system using rough set theory and the classification was based on the template rule, ANN (Artificial Neural Networks), HMM (Hidden Markov Models), Bayesian and SVM based classifiers. Similar classification rules were also applied to the audio emotion recognition system. Happy et al. (2012) proposed an automated single face landmark detection technique which was similar to the state-of-the-art landmark detection methods, but the feature selection process was static, and it requires a lot of effort for high performance. Further, they proposed a single face emotion recognition system using Gabor descriptor and local binary patterns (LBP), but it failed to produce better results when applied to huge datasets. Li et al. (2015) proposed a deep feature based multi-kernel learning approach for Video Emotion Recognition. This method combines hybrid and multi-modal features for emotion classification, which includes spatial-temporal, LBP-TOP, openEAR, CNN features; but this system was less accurate for large datasets. Table 2 shows the summary of some key existing works on facial expression based affective state recognition. These works were performed in the area of affective computing for emotion recognition of humans, but not specifically for analyzing the students' affective states.

**Mixed emotions** Apart from the use of basic emotions directly, these basic emotions were also combined to get mixed emotions. These mixed emotions (like happily surprised, anger sadness, etc.) were trained and tested on datasets such as CK+ (Extended Cohn-Kanade), MMI, and GEMEP-FERA using Local Binary Patters, SVM, and Open CV (Agarwal and Mukherjee 2018; Valstar et al. 2012). Multimodal mixed emotion recognition using the head, face, hand, and body was performed using Face API, and the data was obtained using Kinect (Patwardhan 2017). But, the study conducted by D'Mello et al. (2007) and D'mello and Graesser (2012) inferred that these mixed and basic emotions were significantly dominated by learning-centered affective states such as engaged and frustrated.

### 2.3 Multi-modal and wild data analysis

Multi-task convolutional neural networks were proposed to address pose-invariant unconstrained face recognition (Yin and Liu 2018). Weighted Mixture Deep Neural Network was used to recognize the basic emotions for CK+ database (which include occluded faces) with 97% accuracy. Further, the authors proved that CNN based architectures outperformed all the handcrafted feature based methods (Yang et al. 2018). Deep Fusion CNN was used for multi-modal expression recognition to classify the basic emotions (Li et al. 2017a). Blur-aware bi-channel deep neural network was proposed for face recognition in the wild (Ding and Tao 2018). Several other works were proposed for face recognition, emotion classification and multi-modal analysis for the data of faces in the wild (Xie and Hu 2019; Zhang et al. 2019; Li et al. 2017b; Ding and Tao 2015; Wu et al. 2018). The base network for most of the above mentioned methods use one among the following architectures: AlexNet (Krizhevsky et al. 2012), VGGNet (Simonyan and Zisserman 2014), ResNet (He

**Table 2** Summary of existing works on affective state recognition techniques

| Authors | Method | Classifiers | Affective descriptors | Accuracy | Demerits | Merits |
|---|---|---|---|---|---|---|
| Arifin and Cheung (2007) | Video segmentation based on pleasure arousal dominance information | 4-level DBNs | Violence, neutral, sadness, fear, amusement and happiness | 0.860 | Accuracy rate is low for some emotions like fear and anger | 4-level DBNs are used for classification of emotions |
| Xu et al. (2008) | Hierarchical affective state analysis based on dimentional emotions namely, arousal and valence | Fuzzy,C-Mean Clustering+ CRF | Intensity, fear, anger, happiness, sadness and neutral | 0.807 | The detection accuracy of fear emotion is less than 50% | Few emotions got better accuracy |
| Kang (2003) | HMMs based affective content detection | AdaBoost, relevance feedback | Fear, sadness and joy | 0.849 | Only frontal face was considered | Accurate in classifying fear, sadness and joy |
| Teixeira et al. (2012) | Determination of emotional content of video clips by low level audiovisual features | HMM | Sadness, happiness, anger, fear, disgust and surprise | 0.770 | Accuracy is low as compared to other classifiers | Uses Markov Model to classify emotions including audio features |
| Irie et al. (2010) | Affective audio-visual words and latent topic driving model for realizing movie affective scene classification | LDA | Joy, acceptance, fear, surprise, sadness, disgust, anger, anticipation | 0.855 | Accuracy is less as compared to other classifiers | Uses LDA with short term memory for better accuracy |

et al. 2016) or GoogleNet (Szegedy et al. 2016). But all these methods were used for crowd data analysis in entertainment domain or for sports data analysis but not for students' affective state recognition and classification. Hence, we proposed a novel hybrid CNN architecture for the students' affective state analysis in the classroom environment. Though there were several works related to affective state recognition of students, very few works were done for the recognition of learning-centered affective state using both emotional and behavioral patterns with some cognitive aspects involved in it. Further, it was observed that there exist no standard dataset for the verification and validation of the students' affective states in a classroom environment. This motivated us to create our dataset for the students' affective state analysis.

Table 3 shows the summary of recent works on affective content analysis performed in various domains. From the existing literature it is clearly evident that the traditional machine learning techniques such as ANNs, SVMs etc. were outperformed by deep learning techniques such as CNN, RNN and LSTM (Long Short-Term Memory) for image frame based emotion recognition (Baveye et al. 2017). Table 3 shows existing works on multi-face emotions recognition for different types of smiles related to happiness. Similarly there were other works in the field of entertainment which were tested on MELD (Multimodal Emotionlines Dataset) and EmotiW dataset using InceptionV3 and VGG16 deep learning architectures (Poria et al. 2018; Guo et al. 2018). But, these architectures were not sufficient to test on learning-centered emotions such as engaged and bored where the students' expressions and behavioural patterns vary significantly when compared to the facial expressions and behavioural patterns of humans obtained from the entertainment filed (DeFalco et al. 2018).

To summarize, automatic prediction of the students' affective states were less explored in the classroom environment. The existing studies predict the students' emotional and behavioral engagement separately in both e-learning and classroom environments. The use of multi-modality for recognizing the students' affective states was very less explored in the literature (though there were claims in other domains that the multi-modality will increase the performance of affect detection). An automatic prediction of group-level students' engagement was not explored in the literature. Finally, there exists no standard dataset to train, test, and validate the machine learning/deep learning models in classroom environments. This motivated us to propose a hybrid convolutional neural network architecture to automatically predict the students' affective states in the classroom environment along with the group level engagement score prediction for each frame obtained from the classroom environment. We also created a students' dataset for classroom environment using engaged, bored, and neutral affective states to analyze both the students' behavioral and emotional engagement.

## 3 Proposed methodology

The proposed methodology includes the database creation and the students' affective state recognition using convolutional neural networks. The database creation

**Table 3** Summary of works on image based affective content analysis

| Authors | Method | Tested on | Emotional descrptors | Domain | IFC | M | GA |
|---|---|---|---|---|---|---|---|
| Gupta et al. (2016) | InceptionV3 model | DAiSEE dataset | Boredom, confusion, engagement, frustration | Education | SP | No | No |
| Lopes et al. (2017) | Modified CNN | CK+, JAFFE and BU-3DFE | Ekman's basic emotions | Education | SP | No | No |
| Huang et al. (2019) | Long short term memory | DAiSEE Dataset | Boredom, confusion, engagement, frustration | Education | SP | No | No |
| Hayashi (2019) | Facial action coding system | 21 Japanese University students | Ekman's basic emotions | Education | SP | No | No |
| Ramirez et al. (2019) | Decision Trees, Data obtained from Kinectv2 | 16 undergraduate students | Engagement, frustration | Education | MP | No | Yes |
| Tiam-Lee and Sumi (2019) | WEKA and OpenFace | 73 Students | Boredom, confusion, engagement, frustration | Education | SP | YES | No |
| Subramanian et al. (2018) | Linear SVMs | ASCERTAIN | Engagement, liking, familiarity | Entertainment | SP | YES | No |
| Liu and Jiang (2019) | Particle swarm optimization, kNNs | 10 Adults from shooting team | Happiness, sadness, anger, fear | Sports | SP | YES | No |
| (Huang et al. 2018) | Information aggregation, decision fusion | HAPPEI and GAFF | Neutral, small smile, large smile, small laugh, large laugh and thrilled | Real-Life activities | MP | YES | Yes |

IFC - Image Frame Content; M - Multimodality; GA - Group Analysis

SP - Single Person in a Single Image Frame; MP - Multiple Person in a single image frame

includes details such as affective state classification, affective state's labels & definitions and data annotation performed on both posed and spontaneous data. Further, we included the data augmentation to achieve the robustness of the proposed model and thus, we create two different datasets for the single and multiple students in a single image frame. Figure 1 shows the flow of the proposed architecture and the detailed explanation is provided in the following subsections.

## 3.1 Database creation

### 3.1.1 Affective state classification

Existing works include recognizing a student's facial expressions and classifying them into Ekman's basic emotions (Ekman 1992). The study investigated by D'Mello et al. (2010) showed that learning-centered emotions such as flow, boredom, and frustrated are more dominant and regularly observed in students than Ekman's basic emotions. Whitehill et al. (2014) considered the body posture and eye movement along with facial expressions and classified the engagement into four types engagement levels. The different types of engagement are monitored during the learning

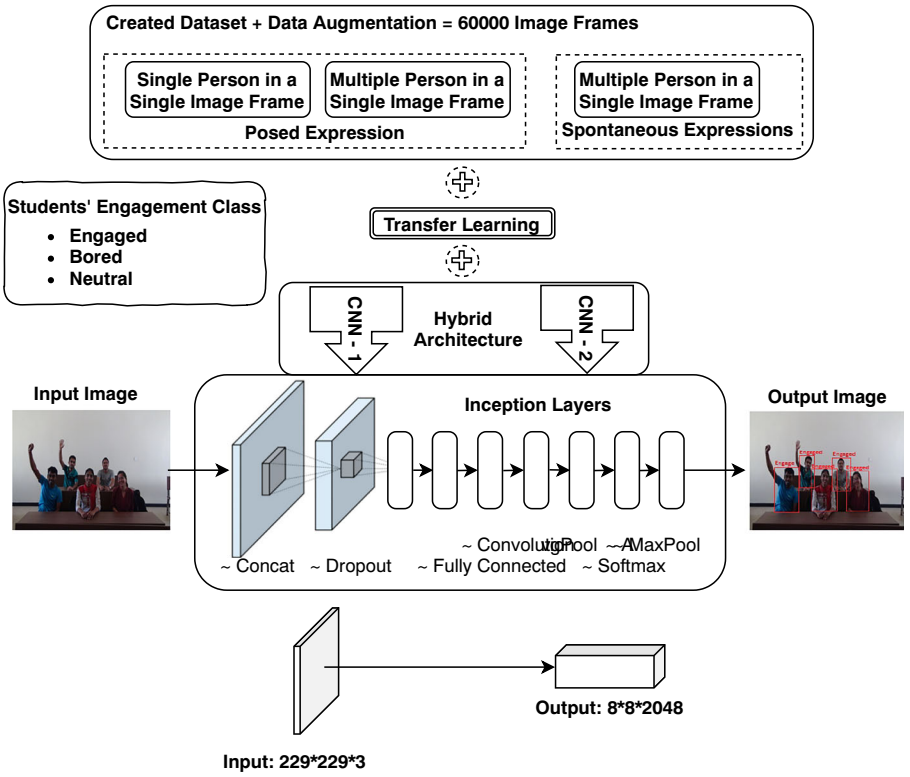


**Fig. 1** Proposed students' affective state recognition architecture

process, and corresponding affective trajectories are generated. The observed results show that flow or engaged students can finally end up getting bored if the confusion or frustration is not addressed, as shown in Fig. 2: Russell's core affective framework (D'Mello 2012). Hence, we considered the engaged/flow, neutral, and boredom affective state for this study. Analyzing the students' emotions alone is not sufficient to classify among the above mentioned affective states. Behavioral patterns such as looking away from the task, eyes barely open, and complete lean on the desk are also beneficial to analyze the affective states. The facial expressions contain more information about the emotions, and hand gestures, and body postures contain more information about the behavioral patterns. Therefore, along with facial expressions for emotions, we also used the hand gestures and body postures to analyze the behavioral patterns to classify the affective states. So, the proposed classification includes both behavioral and emotional engagements of the students, along with some cognitive aspects involved in it. Based on this, we tried to classify the student's engagement into three states, namely: engaged, boredom, and neutral.

### 3.1.2 Labels and definitions

Learning-centered emotions and the behavioral patterns (head and eye movements) are considered as mentioned in D'Mello et al. (2010) and Whitehill et al. (2014) respectively. But, these works were on the single person in a single image frame. Also, the behavioral patterns mentioned in Whitehill et al. (2014) did not include the

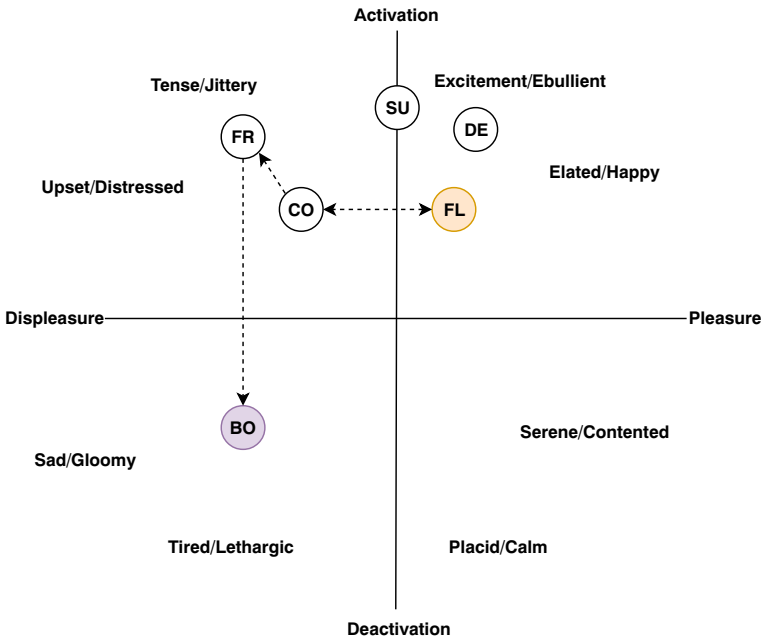


**Fig. 2** Russell's core affective framework (FRustrated, COnfusion, FLow, BOredom, DElight, SUrprise) (D'Mello 2012)

body postures and hand gestures of multiple students in a single image frame. Further, a few emotions are recognized accurately using both hand gestures and facial expressions instead of using only facial expressions. Hence, in this proposed classification, we modified the existing classification standards by adding the hand gesture and the body posture components of multiple students in a single image frame. The label details are mentioned below.

*Engaged:* The engaged label includes both emotional and behavioral engagement aspects such as looking at the teacher/board, taking notes, listening, and discussions with the teacher and the standard facial expression action units of engaged/flow emotion.
*Boredom:* It includes looking away from the board/teacher, eyes barely open or completely closed and obviously not thinking about the task, leaning on the desk with heads down and the standard facial expression action units of boredom emotion.
*Neutral:* If there are no behavioral patterns and affective states recognized, with no facial expression, then it is labeled as neutral.

#### 3.1.3 Data annotation

**Data annotation for posed expressions** To collect the posed expressions and behavioral patterns from the students, we mentioned the situation and the required affective state expected from them. We collected 2 minutes of video clips for all the variants of each affective state. We observed that the affective states were at its peak for about 2 to 8 seconds for each variant of an affective state in a 2 minutes video. We collected those peak frames and got manually verified by the student.

**Annotations for classroom data in the wild** We also collected the students' 1-hour spontaneous (expressions and behaviors) classroom data with 24fps. The collected data is preprocessed by deleting the blurred and repeated frames. Here, repeated frames are those frames which had the same expressions and behaviors from all the students present in the single image frames. E.g., there were occasions where we got two successive image frames which were exactly the same. After pre-processing, we got 2400 image frames, and these image frames were manually annotated by three annotators (Sidney et al. 2005). One annotator was the student himself who did self-annotation, and the other two were expert annotators who followed the guidelines and definitions given for each label. We calculated Cohen's $\kappa$ to check the reliability among annotators, and we found that these annotators reliably agree when discriminating three different affective states with Cohen's $kappa = 0.59$. Figure 3 shows the sample image snapshot of spontaneous classroom data.

### 3.2 Affective state detection of multiple students in a classroom environment

Here, we used two different methods for identifying the affective state of the students, which are as follows.

**Fig. 3** Sample image frame of students' spontaneous expressions and behavioral patterns obtained from classroom data

- Method1: - Detecting all the students in the image, classifying their affective states separately using each student's multi-modal data and then taking the collective average of all the students' affective states.
- Method2: - One complete image frame with all the multi-modal features of all students is considered to classify the group engagement score of that image frame.

### 3.2.1 Building convolutional neural network model

The functioning of convolutional neural networks is similar to that of a bunch of neurons collectively processing the input image and analyzing the data using axons, dendrites, and synapse. Similarly, CNN uses hidden layers, fully connected layers, and a classifier to classify the given input image frame data.

As shown in Fig. 1, filters are convolved on the input image, and their dot product is calculated. These filters are used until they reach the first fully connected layer. Then ReLu is used as an activation function after every convolutional layer (CONV) or pooling layer (POOL). Finally, after the last fully connected (FC) layer, we used the softmax classifier to get the probability distribution for different classes which represents scores between 0 and 1. Figure 1 shows an overall view of the proposed CNN model where an image is given to the model, and then its pixels are used as an input with their corresponding RGB (Red Green and Blue) values. This input feature (pixels) goes through a series of layers (CONV, Relu, Pool, and FC) and finally, it classifies the image as a class in terms of its score. Here the students' image is given as input and the output is the collective average of all of their affective states.

We used the inception v3 (Szegedy et al. 2016) model for affective state classification of students by providing a 299x299x3 RGB image to the model. To reduce the training time, we trained only the final layer of inception v3 model for affective state categories (engaged, boredom and neutral). We built two separate convolutional neural network models, CNN-1 for single student in a single image frame and CNN-2 for multiple students in a single image frame where CNN-2 consists of few extra layers. By combining both of these we emerged with the proposed hybrid architecture which yielded better accuracy of results.

The two CNN models are:

- CNN-1 - It is trained for classifying the affective state of a single student in a single image frame (Fig. 4).
- CNN-2 - It is trained for classifying the affective state of multiple students in a single image frame (Fig. 4). Further, the number of layers is increased by 20 from the base Inception-V3 model. The hyperparameters are fine-tuned separately.

**Classification of the affective state of students for entire image using convolutional neural network** The students' spontaneous data is analyzed using a hybrid model which is a combination of both CNN-1 and CNN-2 models. The number of layers in the hybrid model is the same as that of CNN-2 model. The last two layers of CNN-1 model are also added in the hybrid model along with their weights. The model is trained for 299x299x3 image with RGB color values so that we can downscale or upscale the image accordingly. An input image of size 299x299x3 is given to the hybrid model, and it generates scores for the three corresponding affective states. The class with the highest score is considered for the overall affective state of students in the classroom.

**Collective average affective state score** The students' posed data present in multiple people in a single frame image has one affective state for the entire image, but the students' spontaneous classroom data has different affective states in a single image frame. Hence we used feature fusion to calculate the same. The multi-modal (here, it is an intra-image multi-modality where the features of different students with their facial expressions, hand gestures and body postures present within that image frame are considered) feature fusion vector $V_f$ for any pixel $p_i$ and normalized prediction

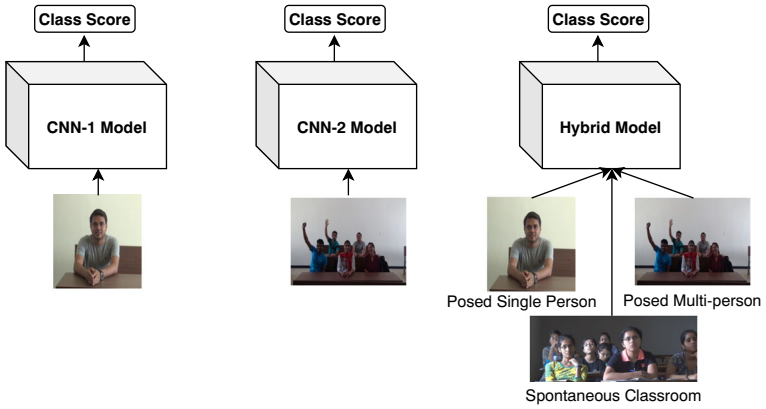


**Fig. 4** Class score prediction using proposed architectures

vector $N_{P_i}$ uses normalized predicted probability distribution $N_{P_i,a}$ of class $a$ using the softmax function (1).

$$N_{P_i,a} = \frac{e^{W_a^T V_f}}{\sum_{i \in classes} e^{W_i^T V_f}} \tag{1}$$

Where $W$ is the temporary weight matrix used to learn the features. The training generally converges in $T = 4000$ epocs. The final collective average affective state score $A_{S_i}$ is given by (2).

$$A_{S_i} = arg\ max\ N_{P_i,a}\ where\ a \in classes \tag{2}$$

### 3.3 Data augmentation

The key to the robust deep learning model is the high quality data. But, it is a challenge to obtain such data. One better way to address this issue is the augmentation of datasets. Due to lack of data available for affective state analysis, we used data augmentation. Data augmentation has increased training data size by 10-fold. Following are the different data augmentation techniques which we performed on our datasets.

- channel_shift_range: Random channel shifts of the image.
- zca_whitening: Applies ZCA whitening to the image.
- rotation_range: Random rotation of image with a degree range.
- width_shift_range: Random horizontal shifts of the image with a fraction of total width.
- height_shift_range: Random vertical shifts of the image with a fraction of total height.
- shear_range: Shear intensity of the image where the shear angle is in the counter-clockwise direction as radian.
- zoom_range: Random zoom of the image where the lower value is 1-room_range and upper value is 1+zoom_range.
- fill_mode: If any of constant, nearest, reflect or wrap are filled according to the given mode, if any points outside the boundaries of the input.

**Table 4** Types of data augmentation used

| Type of augmentation | Augmentation value |
|---|---|
| channel_shift_range | 20 |
| zca_whitening | TRUE |
| rotation_range | 40 |
| width_shift_range | 0.2 |
| height_shift_range | 0.2 |
| shear_range | 0.2 |
| zoom_range | 0.2 |
| horizontal_flip | TRUE |
| fill_mode | Nearest |


**Fig. 5** Sample images of dataset-1: single student in a single image frame

– horizontal_flip: Randomly flip the inputs horizontally. Table 4 shows the details of different data augmentations performed on our dataset.

### 3.4 Variants of datasets created

We created a dataset considering two types of image inputs consisting of 50 Indian students each for both training and testing.

– Dataset-1: It contains 24000 posed images of 50 students with data augmentation for classification. Dataset consisting of images of a single student in a single image frame with three different affective states, namely: engaged (Fig. 5a), boredom (Fig. 5b and c) and neutral (Fig. 5d) as shown in Fig. 5. Every affective state has 8000 image frames each.
– Dataset-2: It contains 36000 images with data augmentation for classification of multiple students sitting in the classroom and their affective states are being classified into three different classes, namely: boredom (Fig. 6a and b), neutral (Fig. 6c), and engaged (Fig. 6d) with 12000 images each as shown in Fig. 6. Dataset-2 also contains 2 hours of classroom video with 2400 students' images with spontaneous expressions.

Collection of posed data sets was necessary as it facilitated the training of the proposed architecture. It was also observed that the posed datasets for single and multiple students in a single frame image considered for training the proposed architecture


**Fig. 6** Sample images of dataset-2: multiple students in a single image frame

increased the overall accuracy by 18% while testing the spontaneous classroom data (results are shown in Table 10).

These 24000 images of dataset-1 and 36000 images of dataset-2 contain the images with data augmentation as mentioned in Table 5.

## 4 Results and analysis

### 4.1 Experimental set-up

For the current study we used $8^{th}$ Generation Intel® Core™ $i7-4510U$ Processor, 8GB RAM, and 2GB NVIDIA® GeForce® $840M$.

Table 6 shows the training setup attributes for both CNN-1 and CNN-2 models where each attribute has its corresponding values used for training both the models.

Retraining of the last pool layer (pooled three layers) took around 2 hours for three classes of 50 students with three affective states. Each training set consists of

**Table 5** Details of created datasets for posed affective states

| Affective states | Number of single students in a single image frames with data augmentation | Number of multiple students in a single image frames with data augmentation |
|---|---|---|
| Engaged | 8000 | 12000 |
| Board | 8000 | 12000 |
| Neutral | 8000 | 12000 |

**Table 6** Training set-up for CNN-1 and CNN-2 models

| Attribute | Details | Attribute | Details |
|---|---|---|---|
| Validating and testing batch size | 100 images | Optimize | Adaptive momentum |
| Epochs | 4000 | Loss | Categorical entropy |
| Learning rate | 0.1 | Weight initialization | Pre Trained weights of Imagenet v3 model trained over imageNet. We fine tune that model by retraining bottleneck layer |
| Classifier at Bottleneck layer | Softmax classifier | Testing set | 10% of training dataset |
| Number of classes | 3 | Validating set | 10% of training dataset |

800 images of students, with a different facial expressions, hand gestures and body postures.

### 4.2 Performance evaluation of posed data

Convolutional neural networks work effectively for object classification in an image, even for detecting faces, gestures and the affective states of humans. The performance evaluation of posed data is carried out using Method1 mentioned in Section 3.2. Initially, we got similar results for both the methods (Method1 and Method2), but we did not get better results when the number of students present in a single image frame was more than one. Table 7 shows the accuracy of different affective states for our proposed hybrid architectures.

The performance evaluations were carried out for both CNN-1 and CNN-2 models individually, but the proposed hybrid architecture performed better than the individual models. The overall results of the proposed hybrid model is shown in Table 10.

**Table 7** Performance evaluation of posed data

| Model | Affective states | Test accuracy (%) | Average test accuracy (%) |
|---|---|---|---|
| CNN-1 | Engaged | 95.2 | 94.0 |
| | Boredom | 93.1 | |
| | Neutral | 93.7 | |
| CNN-2 | Engaged | 96.2 | 95.6 |
| | Boredom | 95.6 | |
| | Neutral | 95.0 | |
| Hybrid | Engaged | 96.2 | 95.8 |
| | Boredom | 96.1 | |
| | Neutral | 95.0 | |

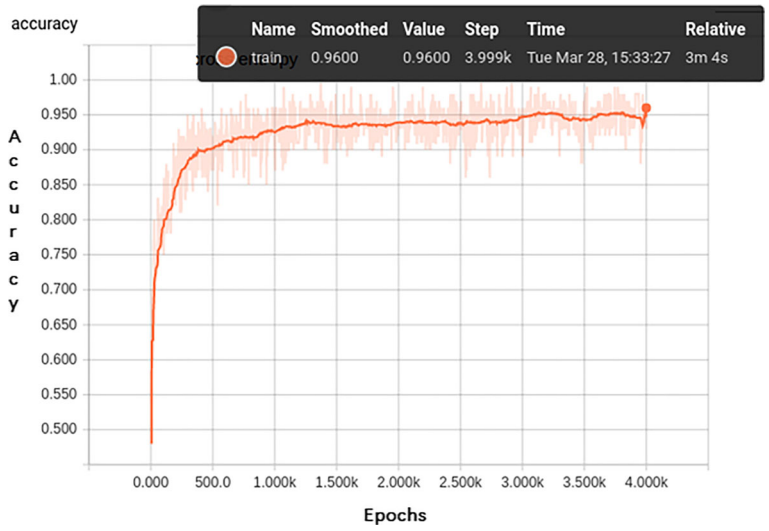


**Fig. 7** Accuracy curve w.r.t epochs for training CNN-1 model

The proposed CNN-1 model with the dataset of single students images got a final test accuracy of 94%.

The accuracy for three different affective states, i.e., engaged, boredom, and neutral are 95.2% , 93.1% and 93.7%, respectively. We also observed that training and validation accuracy was improving with each step or epoch, and reached the saturation after 1500 epochs as shown in Figs. 7 and 8. At the end of 4000 epochs, we got cross entropy of 0.1459 for training and 0.2045 for validation as shown in Figs. 9

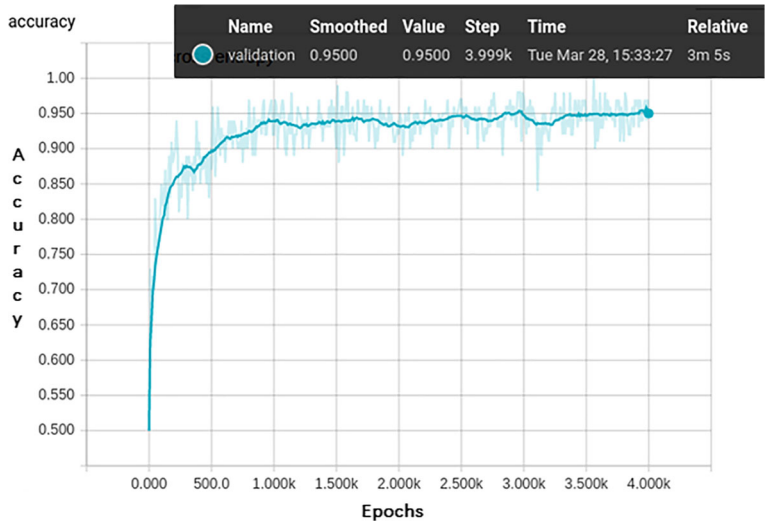


**Fig. 8** Accuracy curve w.r.t epochs for validation CNN-1 model

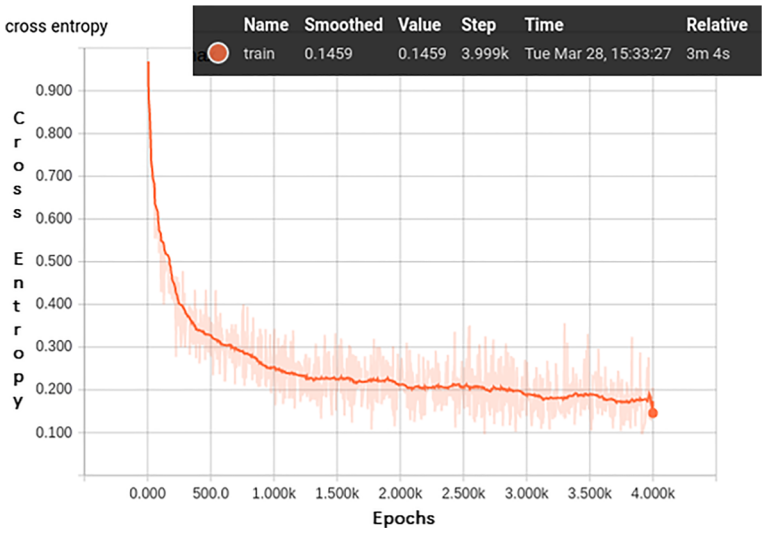


**Fig. 9** Accuracy curve w.r.t cross entropy for training CNN-1 model

and 10. Similarly, we performed the experiment on CNN-2 model and the results of test accuracy are shown in Table 7.

An off-the-shelf CNN training method is used to overcome the problem of high computational training cost, and this process is referred to as transfer learning. Here, we used Inception-V3 model which is trained over ImageNet dataset, we used the weights to this model by training the last layer (pool 3 layers) of inception model as

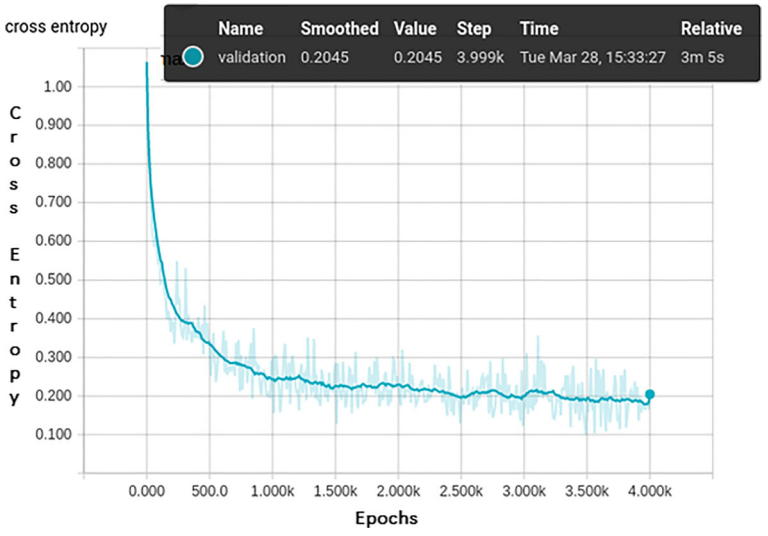


**Fig. 10** Accuracy curve w.r.t cross entropy for validation CNN-1 model

**Table 8** Time and accuracy obtained for the proposed methods

| Proposed methods | Training time (Hrs) | Testing Accuracy (%) | Testing time (ms) |
|---|---|---|---|
| CNN-1 | 1.5 | 94.0 | 1200 |
| CNN-2 | 2.0 | 95.6 | 1200 |

shown in Fig. 1. Training took 1.5 hours for the CNN-1 model and 2 hours for the CNN-2 model as shown in Table 8. This helped us to reduce the training time and thus produce higher accuracy using Inception-V3 model for transfer learning.

We considered three different parameters for both the CNN-1 and CNN-2 models. Table 8 gives the details about the testing time for each model when given an image as input. Training time represents the time taken for training the model for a given set of images. Testing accuracy demonstrates the performance of our model while classifying the given images. As the overall running time for CNN-1 is lesser than that of CNN-2, it can further be considered for use in e-learning and flipped classroom environments.

### 4.3 Performance evaluation of spontaneous data

**Testing on spontaneous classroom data** The proposed techniques are tested on spontaneous classroom data, where the videos are recorded during the actual classroom lectures. The test data is student-independent and classroom-independent data (the students/classroom in test data are different from training data). Table 9 shows the classification accuracies of affective states obtained by the three different models. The recognition of boredom affective state is better when CNN-1 model is used when compared to that of CNN-2, whereas CNN-2 performs better than CNN-1 for the engaged affective state. Hence, we used the hybrid of CNN-1 and CNN-2 models and obtained better accuracy as shown in Table 9.

**Table 9** Comparison of proposed affective state classification techniques with spontaneous classroom data

| Model | Affective States | Accuracy (%) |
|---|---|---|
| CNN-1 | Engaged | 61.4 |
| | Bored | 66.3 |
| | Neutral | 70.0 |
| CNN-2 | Engaged | 68.3 |
| | Bored | 63.1 |
| | Neutral | 71.1 |
| Hybrid | Engaged | 68.7 |
| | Bored | 66.8 |
| | Neutral | 73.7 |

**Table 10** Overall results of the proposed model

| Performance metrics | Affective state classification | | | | | |
|---|---|---|---|---|---|---|
| | CNN-1 Model | CNN-2 Model | Hybrid model | CNN-1 model | CNN-2 model | Hybrid model |
| | Posed dataset | | | Spontaneous classroom dataset | | |
| Average Accuracy | 0.84 | 0.85 | 0.86 | 0.65 | 0.67 | 0.70 |
| Average Recall | 0.87 | 0.86 | 0.89 | 0.66 | 0.71 | 0.72 |
| Average Precision | 0.91 | 0.88 | 0.91 | 0.69 | 0.78 | 0.77 |
| Average F1-score | 0.85 | 0.84 | 0.84 | 0.60 | 0.63 | 0.62 |
| AUC | 0.88 | 0.89 | 0.90 | 0.63 | 0.68 | 0.69 |

### 4.4 Overall results

The overall performance evaluation of the proposed method with the created database is shown in Table 10. The CNN-1, CNN-2, & hybrid models are tested on student-independent & classroom-independent 10-fold cross-validation, and the observed results are mentioned in Table 10. It is observed from Table 10 that both CNN-1 and CNN-2 are necessary for better classification of student's affective states. The hybrid model gave almost the same performance as that of CNN-1 and CNN-2 models for posed dataset but performed better than CNN-1 and CNN-2 models for the spontaneous dataset. F1-score is less for spontaneous data as the affective states are not equally distributed in the spontaneous classroom data.

The created dataset also contains images from different camera positions (to make the recognition process more robust). Figure 11 is a sample image frame from the created dataset with different camera position/angle and was tested with the proposed method. It is observed from Fig. 11 that the affective states are classified correctly for each student and also the overall class score for that frame was also calculated (top left corner in Fig. 11). The proposed model is also tested on a few images of classroom subset data of Imagenet database (Deng et al. 2009) and a sample image

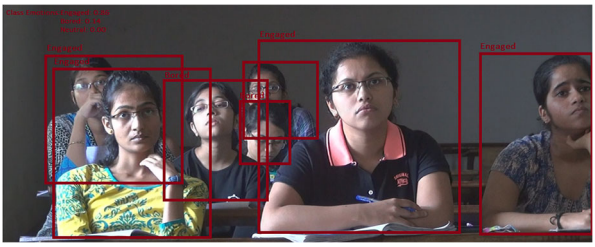


**Fig. 11** Screenshot of the sample tested image of created dataset


**Fig. 12** Screenshot of the sample tested image of ImageNet dataset

snapshot is shown in Fig. 12. Since the ImageNet database does not contain the annotations for students' affective states, we were not able to compare the results with the ground truth.

### 4.5 Comparison with state-of-the-art

**State-of-the-art deep learning architectures** Deep learning methods are popular and give better results for multi-modal affective content classification. The existing deep learning techniques use the following architectures as backbones, which include, AlexNet, ResNet, VGGNet and Inception-V3. These architectures were tested on the created database and we obtained an accuracy of 51% for AlexNet (Krizhevsky et al. 2012), 53% for ResNet (He et al. 2016), 61% for VGGNet (Simonyan and Zisserman 2014). The results were encouraging for a single person in a single image frame but failed to perform better than the proposed model for images obtained from the classroom environment.

**Students' affective state classification methods** There are limited recent studies on student engagement. Whitehill et al. (2014) used Gabor features with SVM and obtained an AUC of 0.729 to analyze the behavioral patterns. But this result was by testing on a single person in a single frame image. Zaletelj and Košir (2017) used the Kinect sensor & KNN and obtained an accuracy of 0.753 in the classroom environment. Though the Kinect considers multiple people in a single image frame, the range of capturing students is less. Hence, the students' detection accuracy decreases if the number of students are more than 10 in a single image frame. Kahu (2013) and Bosch et al. (2016) used deep instance learning and WEKA (Waikato Environment for Knowledge Analysis) tools, but it is already evident from the literature that the handcrafted features are less efficient for faces in the wild. It is difficult to directly compare the proposed methodology with the existing works as the datasets, and the multimodalities are different, in spite of which our results are comparable and more robust in terms of AUC and accuracy.

### 4.6 Application and generalizability

This study proposes a group engagement score for each frame, which allows us to visualize the affective state of students throughout the class. This can be used as feedback to improve the teaching-learning process for novice teachers for further improvement. Further, the proposed method can be used to improve and personalize the teaching-learning process by proving the affective content as feedback to both students and teachers. Also, the proposed method helps us to study/map for any possible correlation among the students' affective states and their performance in tests. In webinars and large classrooms, the proposed method can be used to assist the faculty member as it will be challenging to see every student throughout the class.

The proposed method can be used in intelligent tutoring systems to make it more personalized by providing the affective content as feedback. Recently developed auto-tutors detect the students' affective states and accordingly responds. The affective state prediction techniques used in the Auto tutors can be replaced by the proposed method due to its robustness. Further, the affective content analysis can be used as an immediate feedback to modify the teaching strategy accordingly.

In addition to its application in the education domain, the proposed multi-modal affective state predictor can be used in the entertainment domain as well to see the users' experience while watching movies, advertisements, and so on. The users' experience in shopping malls and other places can also be predicted. Although the proposed affective state predictor can be used in any application that requires the boredom and engaged affective states, every domain requires a new set of contextual features to be reengineered for better prediction.

The deep learning techniques are quite robust and take raw input images (preprocessing of images are not required) for classification of affective states. The proposed method is quite robust to predict the affective states of students in e-learning, flipped classroom, classroom, and webinar environments as the method is well trained with data augmentation, and the use of multi-modality makes it work quite effectively in all the learning environments. It is observed from Section 4.4 Fig. 12 that the proposed method is more robust as it is tested on students of ImageNet dataset which contains students of different age groups with diverse cultural and regional backgrounds. In other domains such as entertainment, business and shopping, the new set of contextual features needs to be added and reengineered to perform better in predicting the engaged and bored affective states along with the neutral.

### 4.7 Significance

Affective state classification with both behavioral and emotional engagements analysis was one of the contributions where we practically exhibited the feasibility of detecting the bored and engaged affective states along with the neutral. It was difficult to analyze the students' engagement using just face in a classroom environment as there were cases where the faces were occluded but the hand gestures and body postures were important to judge the behavioral pattern. Also, the use of only the body postures or hand gestures was not sufficient to classify into the considered three

affective states. To overcome this issue we used multi-modal affective state analysis using students' facial expressions, hand gestures and body postures.

A novel deep learning architecture to analyze the students' affective states is another contribution. There were complexities in capturing the images to perform the affective state classification caused due to uncontrolled distractions. These are facial expressions and behavioral patterns of students in the wild. Only the annotated spontaneous expressions training data was not sufficient to get better accuracy, hence we also used students' posed data which contains both single and multiple students in a single image frame for training.

The aggregation of the group (entire class is considered as one group) affective states was another challenge. To overcome this, we used the feature fusion technique in the deep learning architecture to analyze the group affective state of students. The last contribution is the created dataset. Creation of students' posed data is easy when compared to labeling faces in the wild. To address this issue, we used multiple annotators and Cohen's Kappa to judge the reliability among the multiple annotators for each image frame obtained from the classroom environment.

### 4.8 Limitations

Similar to most of the research, the proposed methodology also has a few limitations. The current study focusses only on three affective states, few learning centered emotions like frustrated, and confused are not considered. The majority of students present in the created dataset are Indians. Hence, the working of the proposed model may differ when we test on other than Indian students. The proposed multimodal analysis is performed on spontaneous data obtained from a regular classroom environment. This will vary if we consider computer-enabled classrooms or game-based learning classroom environments. This study considers only the emotional and behavioral engagements as the engagement detection is performed by using only the image frames. This limits us to analyze the cognitive engagement of students, for example, the student can be with the not-engaged physical patterns, but he/she may be engaged. These aspects are not considered in this study.

## 5 Conclusion

The current study explored the students' affective states in the classroom environment. Both emotional and behavioral engagements are considered to predict the students' affective states such as engaged and boredom along with the neutral. The multi-modal analysis is performed using the students' facial expression, hand gesture, and body posture to increase the robustness of the method. Since the classroom image frame data contains multiple students in every image, we predicted a group engagement score for image frame data using the feature fusion technique. We proposed a deep learning-based hybrid CNN model to predict the students' affective states. A deep learning model should get trained well, and no standard datasets are available for analyzing students affective states in the classroom environment. Hence, we cre-

ated a dataset with two types of image input using the student's facial expression, hand gestures, and body postures. Dataset-1 consists of a single student in a single image frame, and dataset-2 consists of multiple students in a single image frame. For dataset-2, we also collected students' spontaneous expressions and their behavior in the classroom environment. Manual annotation was carried out by three annotators for annotating three different affective states, namely: engaged, boredom, and neutral. We obtained the reliability among annotators (Cohen's $\kappa = 0.59$) for spontaneous classroom data. We proposed both CNN-1 and CNN-2 models, for the affective state recognition of single and multiple students in a single image frame. We obtained an accuracy of 94% and 95.6% for the proposed CNN-1 and CNN-2 models, respectively, for posed classroom data. Further, we obtained 70% accuracy using student & classroom independent 10-fold cross-validation for the proposed hybrid model, which is a combination of CNN-1 and CNN-2 for spontaneous classroom data. The proposed models outperformed the existing state-of-the-art techniques on both posed and spontaneous datasets.

The proposed model can be further extended for real-time affective state analysis of students in both e-learning and classroom environments. The affective state classifications can be further explored for other learning-centered emotions. Object localization and student identification can be introduced to enhance the performance of auto tutor by making the teaching-learning process more personalized. The data obtained from the proposed method can also be used to map to various student assessments such as diagnostic, formative, and summative assessments and to check for any possible correlation among them. The proposed method can also be explored for different users to know the user experience in various other domains such as entertainment, marketing, healthcare, and so on.

**Acknowledgments** The authors wish to thank undergraduate, postgraduate and doctoral research students of Information Technology Department, National Institute of Technology Karnataka Surathkal, India for their help in creating the affective database for both e-learning and classroom environments.

**Compliance with Ethical Standards** The experimental procedure, participants and the course contents used for the experiment are approved by the Institutional Ethics Committee (IEC) of National Institute of Technology Karnataka. The video recordings of the subjects were included in the experiment only after they gave written consent for the use of their videos for this research experiment. All the subjects were also agreed to use their facial expressions, hand gestures and body postures for all the process involved in the completion of the entire project.

## References

Agarwal, S., & Mukherjee, D.P. (2018). Synthesis of realistic facial expressions using expression map. *IEEE Transactions on Multimedia*, *21*(4), 902–914.

Aiqin, Z., & Luo, Q. (2006). Study on e-learning system model based on affective computing. In *International conference on information and automation, 2006. ICIA 2006* (pp. 245–249): IEEE.

Arifin, S., & Cheung, P.Y. (2007). A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information. In *Proceedings of the 15th ACM international conference on multimedia* (pp. 68–77): ACM.

Ashwin, T., Jose, J., Raghu, G., Reddy, G.R.M. (2015). An e-learning system with multifacial emotion recognition using supervised machine learning. In *2015 IEEE seventh international conference on technology for education (T4E),* (pp. 23–26): IEEE.

Balaam, M., Fitzpatrick, G., Good, J., Luckin, R. (2010). Exploring affective technologies for the classroom with the subtle stone. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1623–1632): ACM.

Baveye, Y., Chamaret, C., Dellandréa, E., Chen, L. (2017). Affective video content analysis: a multidisciplinary insight. *IEEE Transactions on Affective Computing*, *9*(4), 396–409.

Bosch, N., & D'Mello, S. (2017). The affective experience of novice computer programmers. *International journal of artificial intelligence in education*, *27*(1), 181–206.

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 (TiiS)*, *6*(2), 17.

Chen, J., & Luo, Q. (2006). Research on e-learning system model based on affective computing. In *7th international conference on computer-aided industrial design and conceptual design, 2006. CAIDCD'06* (pp. 1–4): IEEE.

DeFalco, J.A., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S., Lester, J.C. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*, *28*(2), 152–193.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848.

Dermeval, D., Paiva, R., Bittencourt, I.I., Vassileva, J., Borges, D. (2018). Authoring tools for designing intelligent tutoring systems: a systematic review of the literature. *International Journal of Artificial Intelligence in Education*, 1–49.

Ding, C., & Tao, D. (2015). Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia*, *17*(11), 2049–2058.

Ding, C., & Tao, D. (2018). Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 1002–1014.

D'Mello, S. (2012). Monitoring affective trajectories during complex learning. In *Encyclopedia of the sciences of learning* (pp. 2325–2328): Springer.

D'mello, S., & Graesser, A. (2012). Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(4), 23.

D'Mello, S., Picard, R.W., Graesser, A. (2007). Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, 22(4).

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.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200.

Girshick, R., Donahue, J., Darrell, T., Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *38*(1), 142–158.

Guo, X., Zhu, B., Polanía, L.F., Boncelet, C., Barner, K.E. (2018). Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In *Proceedings of the 2018 on international conference on multimodal interaction* (pp. 635–639): ACM.

Gupta, A., D'Cunha, A., Awasthi, K., Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. arXiv:160901885.

Happy, S., George, A., Routray, A. (2012). A real time facial expression classification system using local binary patterns. In *2012 4th international conference on intelligent human computer interaction (IHCI)* (pp. 1–5): IEEE.

Happy, S., Patnaik, P., Routray, A., Guha, R. (2015). The indian spontaneous expression database for emotion recognition. *IEEE Transactions on Affective Computing*.

Hayashi, Y. (2019). Detecting collaborative learning through emotions: an investigation using facial expression recognition. In *International conference on intelligent tutoring systems* (pp. 89–98): Springer.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Holmes, M., Latham, A., Crockett, K., O'Shea, J.D. (2018). Near real-time comprehension classification with artificial neural networks: Decoding e-learner non-verbal behavior. *IEEE Transactions on Learning Technologies*, *11*(1), 5–12.

Huang, T., Mei, Y., Zhang, H., Liu, S., Yang, H. (2019). Fine-grained engagement recognition in online learning environment. In *2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC)* (pp. 338–341): IEEE.

Huang, X., Dhall, A., Goecke, R., Pietikäinen, M., Zhao, G. (2018). Multimodal framework for analyzing the affect of a group of people. *IEEE Transactions on Multimedia*, *20*(10), 2706–2721.

Irie, G., Satou, T., Kojima, A., Yamasaki, T., Aizawa, K. (2010). Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. *IEEE Transactions on Multimedia*, *12*(6), 523–535.

Kahu, E.R. (2013). Framing student engagement in higher education. *Studies in Higher Education*, *38*(5), 758–773.

Kang, H.B. (2003). Affective content detection using hmms. In *Proceedings of the eleventh ACM international conference on multimedia* (pp. 259–262): ACM.

Kim, P.W. (2018). Ambient intelligence in a smart classroom for assessing students' engagement levels. *Journal of Ambient Intelligence and Humanized Computing*, 1–6.

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): IEEE.

Kleinsmith, A., & Bianchi-Berthouze, N. (2013). Affective body expression perception and recognition: a survey. *IEEE Transactions on Affective Computing*, *4*(1), 15–33.

Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Li, H., Sun, J., Xu, Z., Chen, L. (2017a). Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia*, *19*(12), 2816–2831.

Li, J., Liu, L., Li, J., Feng, J., Yan, S., Sim, T. (2017b). Towards a comprehensive face detector in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, W., Abtahi, F., Zhu, Z. (2015). A deep feature based multi-kernel learning approach for video emotion recognition. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 483–490): ACM.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21–37): Springer.

Liu, Y., & Jiang, C. (2019). Recognition of shooter's emotions under stress based on affective computing. *IEEE Access*, *7*, 62338–62343.

Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, *61*, 610–628.

Ortony, A., Clore, G.L., Collins, A. (1990). *The cognitive structure of emotions*. Cambridge: Cambridge University Press.

Pao, T.L., Chen, Y.T., Yeh, J.H., Chang, Y.H. (2005). Emotion recognition and evaluation of mandarin speech using weighted d-knn classification. In *Proceedings of the 17th conference on computational linguistics and speech processing* (pp. 203–212).

Patwardhan, A.S. (2017). *Multimodal mixed emotion detection, 2017 2nd international conference on communication and electronics systems (ICCES)*, (pp. 139–143): IEEE.

Picard, R.W., & Picard, R. (1997). *Affective computing* Vol. 252. Cambridge: MIT Press.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv:181002508.

Mirko, R.L.K., & Dillenbourg, P. (2015). Translating head motion into attention-towards processing of student's body-language.

Ramirez, L., Yao, W., Chng, E., Schneider, B. (2019). Toward instrumenting makerspaces: Using motion sensors to capture students' affective states and social interactions in open-ended learning environments, pp. 639–642.

Sidney, K.D., Craig, S.D., Gholson, B., Franklin, S., Picard, R., Graesser, A.C. (2005). Integrating affect sensors in an intelligent tutoring system. In *Affective interactions: the computer in the affective loop workshop at* (pp. 7–13).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:14091556.

Sinclair, J., Butler, M., Morgan, M., Kalvala, S. (2015). Measures of student engagement in computer science. In *Proceedings of the 2015 ACM conference on innovation and technology in computer science education* (pp. 242–247): ACM.

Singh, A., Karanam, S., Kumar, D. (2013). Constructive learning for human-robot interaction. *IEEE Potentials*, *32*, 13–19.

Subramanian, R., Wache, J., Abadi, M.K., Vieriu, R.L., Winkler, S., Sebe, N. (2018). Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, *9*(2), 147–160.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Teixeira, R.M.A., Yamasaki, T., Aizawa, K. (2012). Determination of emotional content of video clips by low-level audiovisual features. *Multimedia Tools and Applications*, *61*(1), 21–49.

Thomas, C., & Jayagopi, D.B. (2017). Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education* (pp. 33–40): ACM.

Tiam-Lee, T.J., & Sumi, K. (2019). Analysis and prediction of student emotions while doing programming exercises. In *International conference on intelligent tutoring systems* (pp. 24–33): Springer.

Valstar, M.F., Mehu, M., Jiang, B., Pantic, M., Scherer, K. (2012). Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics Part B (Cybernetics)*, *42*(4), 966–979.

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.

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.

Worsley, M., & Blikstein, P. (2018). A multimodal analysis of making. *International Journal of Artificial Intelligence in Education*, *28*(3), 385–419.

Wu, H., Zhang, K., Tian, G. (2018). Simultaneous face detection and pose estimation using convolutional neural network cascade. *IEEE Access*, *6*, 49563–49575.

Xie, S., & Hu, H. (2019). Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Transactions on Multimedia*, *21*(1), 211–220.

Xu, M., Jin, J.S., Luo, S., Duan, L. (2008). Hierarchical movie affective content analysis based on arousal and valence features. In *Proceedings of the 16th ACM international conference on multimedia* (pp. 677–680): ACM.

Yang, B., Cao, J., Ni, R., Zhang, Y. (2018). Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access*, *6*, 4630–4640.

Yin, X., & Liu, X. (2018). Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, *27*(2), 964–975.

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), 80.

Zhang, S., Pan, X., Cui, Y., Zhao, X., Liu, L. (2019). Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access*.

Zhang, X., Sun, G., Pan, Y., Sun, H., He, Y., Tan, J. (2018). Students performance modeling based on behavior pattern. *Journal of Ambient Intelligence and Humanized Computing*, *9*(5), 1659–1670.