Indonesian Journal of Electrical Engineering and Computer Science
Vol. 41, No. 3, March 2026, pp. 977~999
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v41.i3.pp977-999

Leveraging CNN to analyze facial expressions for academic engagement monitoring with insights from the multi-source academic affective engagement dataset

Noora C. T., P. Tamil Selvan
Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India

Article Info

Article history:

Received Nov 19, 2023
Revised Nov 3, 2025
Accepted Dec 13, 2025

Keywords:

Academic affective engagement
Emotion recognition
Facial expressions
Multi-source academic affective engagement
Student engagement

ABSTRACT

The dynamics of student engagement and emotional states significantly influence learning outcomes. Positive emotions, stemming from successful task completion, contrast with negative emotions arising from learning struggles or failures. Effective transitions to engagement occur upon problem resolution, while unresolved issues lead to frustration and subsequent boredom. Facial engagement monitoring is crucial for assessing students' attention, interest, and emotional responses during learning. Recent advancements in convolutional neural networks (CNNs) show promise in automatically analyzing facial expressions to infer engagement levels. This study proposes a CNN-based approach utilizing the multi-source academic affective engagement dataset (MAAED) to categorize facial expressions into boredom, confusion, frustration, and yawning. By extracting features from facial images, this method offers an efficient and objective means to gauge student engagement. Recognizing and addressing negative affective states, such as confusion and boredom, is fundamental in creating supportive learning environments. Through automated frame extraction and model comparison, this study demonstrates reduced loss values with improving accuracy, showcasing the effectiveness of this method in objectively evaluating student engagement. Facial engagement monitoring with CNNs, using the MAAED dataset, is pivotal in understanding human behavior and enhancing educational experiences. The CNN model, trained on MAAED annotated facial expressions, accurately classifies engagement categories. Experimental results underscore the CNN-based approach's efficacy in monitoring facial engagement, highlighting its potential to enrich educational environments and personalized learning experiences.

Corresponding Author:

Noora C. T.
Department of Computer Science, Karpagam Academy of Higher Education
Coimbatore, India
Email: nooract@gmail.com

1. INTRODUCTION

The student in a classroom may experience different mixtures of mental states, which are very important factors to reveal the cognitive learning and engagement of students. When the students tend to make mistakes or face failures or struggle at learning may arouse a negative emotional state such as irritation, frustration, and anger on the other hand, if they can complete any task or conquer challenges/difficulties, positive mental states such as delight, excitement, and satisfaction will be the result [1]. According to D'Mello and Graesser [2], in normal cases, the student enters into the learning activity in an engaged and concentrated state it will continue until they reach out in any kind of difficult situation gradually leading to

confusion and boredom. At this point, either one of the transitions will occur. Positive ways of transition will occur when the student returns to the engaged state by resolving the problems he/she faced. Negative ways of transition occur when the problem he/she faced at the time of listening/discussion may not be solved. Gradually the student may stick in such a situation and transition to frustration. if this frustration persisted for some time further leads to boredom. A good teacher should be able to monitor the changes in the mental states of the students during lecturing, thus she can give personalized assistance to the students who felt confusion, frustration, or any other negative emotions. By identifying the problems faced during the discussion, the teacher can able to give further explanation/change the way of teaching in such a way that they can understand easily, thereby maximizing the students learning outcome. Massive open online courses (MOOCs) have brought a revolution in higher education by allowing interested students to pursue their education at their own pace and convenience. The content delivery successful only when it is modulated by real-time student feedback. That key factor is missing in e-learning environments. The automated engagement monitoring methods can easily be employed in such e-learning platforms [3].

Affective computing in education is a growing topic that is increasing in popularity as time goes on. Researchers in this domain employ various methodologies and techniques to capture and interpret emotions in educational settings [2]. Commonly used machine learning models include support vector machines (SVM) [4], convolutional neural networks (CNN) [5], and other deep learning algorithms [6], [7]. Multimodal approaches combine different data types, such as facial expressions, physiological signals [8], [9], and text-based interactions [10], to achieve more accurate emotion detection and analysis. Insights gained from emotion recognition offer valuable understanding of student behaviour and learning experiences across various educational settings, including e-learning, offline classrooms, virtual classrooms, and computer-enabled classrooms. Emotion recognition techniques hold the potential to tailor instructional strategies, offer personalized feedback, and create more engaging educational environments. Integrating affective computing techniques, such as emotion recognition and sentiment analysis, into intelligent tutoring systems enable adaptive instruction based on students' affective states, thereby enhancing engagement, motivation, and overall learning outcomes. Additionally, affective computing plays a crucial role in designing emotionally responsive online learning platforms, which consider students' emotional experiences to create more effective learning environments. The existing literature also focuses on specific aspects of emotions, such as academic emotions, engagement levels, distraction, fatigue, and learning-centred emotions. For instance, Saneiro *et al.* [11] conducted a study on facial emotion recognition (FER) to predict academic performance, highlighting the potential of affective computing in educational assessment. Systematic reviews, like the one conducted by Alameda-Pineda *et al.* [12], emphasize the benefits of integrating affective computing in learning analytics, providing valuable insights for tailoring interventions and supporting student well-being. Nei *et al.* [13] focused on analyzing students' emotional states in online learning environments using text-mining techniques. By analyzing students' written interactions, the researchers gained insights into students' emotional responses, contributing to a deeper understanding of their emotional experiences in online learning. The incorporation of affective computing and emotion recognition techniques has the potential to enhance educational experiences, improve learning outcomes, and provide personalized support based on students' emotional states. By leveraging these techniques, educators can gain a deeper understanding of students' emotional patterns, monitor changes in affective states, and respond proactively to support their learning needs.

Traditional methods for engagement monitoring often rely on manual observation, which is both subjective and time-consuming. However, recent breakthroughs in deep learning methods, notably CNNs [14], have demonstrated significant potential in automatically analyzing facial expressions and inferring engagement levels. In this study, we propose a CNN-based approach for facial engagement monitoring, utilizing a novel dataset named multi-source academic affective engagement dataset (MAAED). Our CNN model is trained on MAAED to classify facial expressions into different engagement categories, such as boredom, confusion, frustration, and yawning. The model can make accurate predictions about students' engagement levels by automatically extracting discriminative features from facial images. The key contributions of this study lie in the creation of MAAED, a unique and rich dataset that reflects real-world academic engagement, and the development of an efficient CNN-based approach for facial engagement monitoring. By harnessing the power of deep learning, the proposed approach aims to provide educators with an objective and efficient method to assess students' engagement during academic activities. This has the potential to enhance educational environments, personalize learning experiences, and enable timely interventions to improve student outcomes.

Eventhough facial expression analysis is a powerful method for detecting emotional responses and assessing student engagement, it is beneficial to integrate multiple modalities into engagement monitoring systems, to gain a more holistic understanding. These additional modalities include eye tracking, which reveals valuable information about students' attention, focus, and information processing during learning

activities [15]. Speech analysis offers insights into students' level of engagement and emotional states by examining speech patterns, tone, and vocal cues [16]. Natural language processing (NLP) techniques applied to students' speech transcriptions or recordings can detect sentiment, engagement, and content understanding [17]. Additionally, it is crucial to consider physiological signals, encompassing heart rate variability, skin conductance, respiration rate, and body temperature. Apart from EEG, which can offer insights into students' emotional states, stresss levels, and cognitive load [18], [19]. Analysing students' interactions with digital learning platforms, including clicks, mouse movements, and touch gestures, provides valuable data on engagement and navigation patterns [20], [21]. Additionally, proximity sensors monitoring students' physical presence in the learning environment offer insights into their level of engagement and participation.

Student engagement is a multifaceted concept, encompassing various dimensions such as behavioral, emotional, cognitive, social, cultural, and contextual engagement. Behavioural engagement involves observable actions like participation and completion of assignments, while emotional engagement focuses on students' feelings and attitudes toward learning. Cognitive engagement relates to mental efforts and involvement in critical thinking and problem-solving. Social engagement emphasizes interactions with peers and teachers, promoting collaboration and communication. Cultural and contextual engagement considers the influence of cultural factors and the learning environment on student engagement, emphasizing inclusivity and supportiveness. Engagement levels can vary from low to medium to high, with each level indicating different levels of interest, motivation, and participation. Recognizing negative affective states like yawning, confusion, boredom, and frustration is crucial, as they can negatively impact student engagement, motivation, and well-being. Addressing these emotions is essential for creating a supportive and effective learning environment. By understanding and catering to individual students' needs, educators can foster a positive classroom atmosphere and promote academic success, helping students to overcome challenges and thrive in their learning experiences. Strategies such as clarification, additional support, relevant and stimulating content, and effective teaching approaches play a vital role in addressing these emotions and enhancing student engagement.

Deep learning techniques, such as CNN, have gained greater popularity in recent studies because of its outstanding results. Ashwin and Guddeti [22] established a strategy based on CNN for hidden engagement analysis using non-verbal cues. Sumer *et al.* [4] use Attention-Net for head pose estimation and Affect-Net for facial expression detection by facial video analysis. Bidwell *et al.* [23] established an automated behavioral analysis system in 2011 to enable teachers to effectively evaluate student behavior. Student engagement is modeled and categorized using many cameras deployed in a third-grade classroom to capture student eye movement patterns. For this purpose, five color cameras and four Microsoft Kinect depth-sensing cameras, SDK known as Pittsburgh Pattern Recognition (PittPatt), were employed. The SDK is used to compute head orientations and gaze targets. The hidden markov model (HMM) is utilized to categorize retrieved sequences of individual student gaze targets as engaged, attentive, or transitional. The expert observation data was utilized to train and evaluate the HMM-based model for automatic engagement detection. The work relied just on the students' gaze targets, which was insufficient to fully comprehend the behavior.

The primary objective of the study is to investigate the role of facial expressions in evaluating student engagement within a classroom environment. Facial expressions provide an immediate and visible means for educators to assess the level of student involvement. However, when dealing with larger groups of students, this method faces significant challenges. As the number of students in a classroom increases, the ability to accurately interpret and analyze facial expressions becomes more complex and less reliable [24]. The diversity of expressions, combined with the difficulty of observing every student at the same time, limits the use of facial cues to assess engagement. As a result, there is a pressing need to investigate and implement alternative, more comprehensive methods of analysis that can assist teachers in better understanding student engagement. Seeking innovative approaches that go beyond facial expression observation alone, considering the dynamics of larger class sizes, becomes crucial in devising a more robust and dependable framework for aiding educators in accurately gauging and enhancing student engagement levels.

Cultural differences pose a significant challenge for this study. Ekman [25] cross-cultural research indicates, some cultures openly express emotions, while others conceal their feelings. Acknowledging and understanding these variations can assist teachers in creating inclusive environments. However, few prior studies have addressed cultural diversity in emotional expression. To accommodate these variations and students' emotional experiences, a new dataset was generated by merging five publicly available datasets. This approach aims to develop a universal method for recognizing emotions and monitoring engagement in classroom settings, leveraging cultural differences to enhance accuracy and inclusivity. Another pertinent aspect of the study involves the automated extraction of frames from classroom videos. This extraction process utilized a trained model specifically designed for engagement analysis. From these video frames, only the two most significant emotions from each category were selected, streamlining the frame extraction procedure to enhance both accuracy and overall performance. Researchers investigated the performance of

various pre-trained models for achieving high accuracy and low loss values in a newly built dataset called MAAED. They trained and compared four popular models: visual geometry group16 (VGG16), ResNet-50, ResNet-101, and Inception V3. In addition to accuracy and loss, they also evaluated F1-score, precision, and recall to provide a comprehensive understanding of each model's performance. This approach utilizes a unique dataset, MAAED, to categorize facial expressions into different engagement categories such as boredom, confusion, frustration, yawning, and concentrated aiding educators in assessing student engagement objectively and efficiently rather than it solves the problem of manual selection of peak frames in each category of emotions by considering variety of expression around the world with the help of MAEED. The loss value have been gradually decreased in the proposed model as the accuracy improves.

The foundation of this paper is rooted in its methodology, outlining the methods employed in the study. Following this introduction, subsequent sections delve into detailed processes, analyses, and outcomes explored within the research. The methodology section outlines how the study was conducted, detailing the tools and techniques applied. Following this, the literature review critically evaluates and consolidates pertinent existing research, emphasizing key insights. It then moves on to the integration and analysis of the MAAED dataset, explaining frame preprocessing, model training, and subsequent evaluation using diverse metrics, along with discussing augmentation methods for enhanced performance. A detailed analysis of the findings unfolds, exploring the overall performance metrics of the study. Lastly, potential challenges impacting the study's outcomes are reviewed in the conclusion.

2. METHOD

In this study, we propose a methodology for facial engagement monitoring in educational settings using a CNN. The methodology incorporates the creation of the MAAED, which combines diverse facial expression datasets covering a wide range of emotions and engagement levels relevant to students. The collected datasets are preprocessed to ensure uniformity and consistency. The designed CNN architecture consists of convolutional layers, pooling layers, and fully connected layers, which are trained on the MAAED using suitable optimization algorithms and loss functions for multi-class classification. The trained CNN model is evaluated on a test set using standard evaluation metrics, such as accuracy, precision, recall, and F1-score, as well as a confusion matrix to assess its performance.

CNNs are a category of deep learning models extensively employed in diverse computer vision applications, including FER. CNNs are structured to automatically acquire and extract meaningful features from input data, which greatly enhances their effectiveness in the field of image analysis. In the context of FER, CNN can be trained to detect and classify different facial expressions by learning patterns and features from facial images. The network is trained to recognize crucial facial landmarks, including the eyes, nose, and mouth, and their spatial relationships to capture the distinct features associated with different emotions. The core design of a CNN includes several layers, such as convolutional layers, pooling layers, and fully connected layers. Convolutional layers play a pivotal role in feature extraction by applying filters to the input image, enabling the detection of local patterns and features. The convolution process entails the movement of a filter (K) across the input image (I), where it conducts element-wise multiplications and subsequently aggregates the outcomes by summing them up.

$$(I * K)(t) = \Sigma_a I(a)K(t-a) \tag{1}$$

This essentially means that every pixel in the output is generated by adding together the input pixels, each multiplied by its respective weight defined by the kernel. In two dimensions, this would be,

$$(I * K)(x,y) = \Sigma_a \Sigma_b I(a,b)K(x-a, y-b) \tag{2}$$

In this process, the kernel undergoes element-wise multiplication with the image matrix, and afterwards, the outcomes are summed up. Pooling layers serve to downsample the feature maps, reducing their spatial dimensions while preserving critical information. Two frequently used techniques for this purpose are max pooling and average pooling. Max pooling, as a technique, focuses on extracting the highest value within each window of the feature map, thereby highlighting the most prominent features within the data. In contrast, average pooling computes the average value for each window, giving an equal representation of all features within the window. This spatial reduction carried out by the pooling layers ensures that the most salient features are retained while the overall data size is condensed, making subsequent layers of the CNN more computationally manageable. Finally, the fully connected layers are responsible for classification, mapping the extracted features to specific emotion categories. After the feature extraction process involving the convolutional and pooling layers, the fully connected layers come into play, mapping

the extracted features to specific class categories. Mathematically, a fully connected layer executes a linear transformation and subsequently applies an activation function. If we denote the input to the layer as x (which would be a flattened version of the output from the previous layer), the weights as W, the biases as b, and the output as y, then the linear transformation can be written as:

$$y = Wx + b \tag{3}$$

the weights matrix W and the bias vector b represent the parameters of the fully connected layer that are learned during training. After the linear transformation, an activation function is applied element-wise. The rectified linear unit (ReLU) is a commonly adopted activation function.

$$ReLU(z) = max(0, z) \tag{4}$$

If the fully connected layer is placed as the final layer in the network and is used for multi-class classification, then a SoftMax function is typically applied to the output of the layer to generate probabilities for each class.

$$softmax(zi) = exp\,(zi)/\Sigma\, exp(zj) \tag{5}$$

where the vector z serves as the input to the SoftMax function, zi is the i^{th} element of z, and the denominator is the sum of exp(zj) overall j.

CNN, a specialized class of neural networks can be see in Figure 1, have gained prominence due to their ability to automatically extract meaningful features from images. In the realm of FER, this characteristic proves invaluable as it eliminates the need for manual feature extraction, allowing the model to discern intricate patterns and subtle nuances inherent in facial expressions. Training with a CNN model (Figure 1(a)), several significant obstacles like overfitting, vanishing gradients, and class imbalance in the dataset can negatively impact the model's performance. To mitigate these issues, it becomes necessary to carefully adjust the model's hyper-parameters and apply regularization techniques. Hyper-parameters are elements that guide the learning process of the model. These include aspects such as batch size, kernel size, the choice of loss function, and the optimization algorithm. Adjusting these can help manage the aforementioned challenges effectively. Regularization techniques are a group of strategies aimed at preventing overfitting, a scenario where the model excels with training data but struggles when faced with unfamiliar or unseen data. L1 and L2 regularization, dropout, data augmentation, and early stopping, are some of the common regularization techniques used. In training the FER model, these challenges were managed effectively, resulting in commendable classification accuracy. Hyper-parameter tuning and regularization techniques were employed to optimize the model, thereby leading to a more balanced and accurate classification of facial emotions. The input image undergoes a series of convolutional layers, followed by pooling layers, which progressively reduce the spatial dimensions.

The flattened result from the last pooling layer is fed into fully connected layers, where classification is carried out using the extracted features. The final output layer represents the different emotion classes, such as happiness, sadness, and anger. The key advantage of CNN in FER is their inherent capability to learn and extract relevant features automatically from facial images, thereby eliminating the requirement for manual feature engineering. This trait makes CNNs highly proficient at capturing intricate patterns and subtle details linked to various emotions.

The heart of this study is the MAAED dataset, which combines facial expression data from five publicly available sources worldwide. The dataset creation involves several steps, from gathering diverse expressions to refining the frames. The pivotal success in this process involves leveraging a trained model for automated frame extraction, significantly reducing time compared to manual selection. Although the dataset encompasses five emotions, this study primarily focuses on four specific negative emotions—boredom, frustration, yawning, and confusion—as an initial step. The concentrated frames corresponding to these emotions are not within the scope of this study's considerations. In Figure 1(b) illustrates the MAAED dataset creation method and the engagement classification process based on the MAAED dataset. The detailed process of dataset creation is elaborated in section 4 of this study. Upon training with pre-trained models like VGG16, ResNet50, ResNet101, and InceptionV3, the proposed model demonstrates superior performance across various evaluation metrics including accuracy, loss, and other relevant benchmarks.

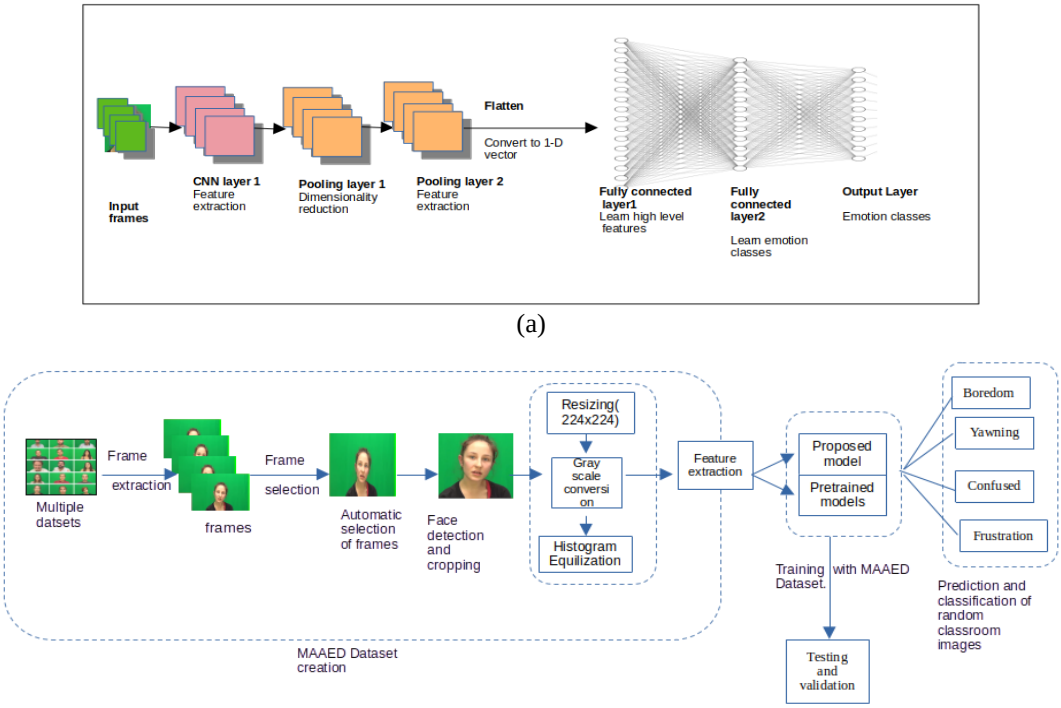


Figure 1. Specialized class of neural networks (a) architecture of CNN and (b) MAAED creation, classification and prediction

3. RELATED WORKS

Understanding and assessing students' engagement levels and emotional states play a crucial role in shaping effective teaching strategies and promoting optimal learning experiences. Each study in Table 1 adopts various methodologies, ranging from traditional machine learning models to sophisticated deep learning architectures. These approaches are designed to capture a wide array of non-verbal cues, including facial expressions, body language, and eye movements. These approaches are designed to decipher students' engagement levels and emotional responses during the learning process. Advanced deep learning methods in certain studies exemplify ongoing progress in education, offering educators deeper insights into human emotions and engagement.

Whitehill *et al.* [26] conducted a thorough analysis of existing computer-vision algorithms for automatic student engagement analysis and recognition. Viola and Jones [27] compared facial features of face patches from various methods such as BoostBF, SVM, Gabor, and CERT toolbox, and did a binary classification of the four types of engagement on the facial expression, and final engagement is estimated from a regression model using the binary classification outputs. Zaletelj and Košir [28] develop a feature set defining both the face and bodily attributes of a student, including gaze point and body posture, using 2D and 3D data received by the Kinect one sensor. Machine learning techniques are used to train classifiers that assess a student's attention levels at various intervals. Krithika and Priya [29] developed a program to identify the emotions of the students by monitoring their head, lip, and eye movements in the e-learning environment. Sahla and Kumar [30] developed a deep CNN technique for classroom emotion detection. A cloud-based facial emotion analysis was conducted (2019) by Boonroungrut *et al.* [31] in facial emotion analysis to find students' emotions in the classroom. The study was conducted among 29 international students by examining their mood changes. Ayvaz *et al.* [32] employ several classification algorithms such as CART, RF, kNN, and SVM to analyze the facial expressions of participants in an e-learning session held over Skype software using the system they built. They eventually concluded that emotions such as happiness, fear, sadness, anger, surprise, and disgust are universally acknowledged in classrooms. Among, the SVM algorithm outperforms others. Recent neurological advancements highlight the connection between learning and emotions. Many studies emphasize the importance of students' emotions during lectures. Acknowledging this strong link between emotions and learning, it's crucial to integrate emotions into education. Understanding and supporting students' feelings leads to improved, customized learning experiences, enhancing both academic performance and students' general well-being.

Table 1. A comparative analysis for understanding performance in diverse learning settings

	Study	Methodology	Categories	Accuracy	E-learning/ Classroom
1.	Whitehill *et al.* [26]	SVM with Gabour features	Not engaged at all, nominally engaged, engaged in task, very engaged, unclear.	76.32	E-learning
2.	Bosch *et al.* [33]	14 different machine learning models like SVM-using facial expressions	Bored, confused, delighted, engaged, frustrated	Individual class accuracy (0.61-0.87)	E-learning
3.	Krithika *et al.* [29]	Facial features like abnormal head and eyes movement machine learning features and 2D, and 3D features from Kinect one camera	Excited, boredom, yawning, drowsiness	N/A	E-learning
4.	Zalatelji *et al.* [28]	Machine learning features and 2D, and 3D features from Kinect one camera	High attention, medium attention, no attention	75.3%	Classroom
5.	Sharma *et al.* [15]	CNN	Student's basic facial expressions	70%	E-learning
6.	Chen *et al.* [34]	Eye tracking with a rule-based system	Happy face, neutral face	N/A	Classroom
7.	Sharma *et al.* [15]	Multimodal (FER, Eye Tracking, Body Language)	Emotions	88%	E-learning
8.	Mukhopadhyay *et al.* [35]	CNN	FER2013 emotions	62	E-learning
9.	Bhardwaj *et al.* [6]	Deep learning	Angry, disgust, fear, happy, sad, surprise and neutral	93.6%	E-learning
10.	Thomas and Jayagopi [36]	Pose, Gaze, AUs	Engaged, distracted	89%	Classroom
11.	Ashwin *et al.* [22]	CNN-based model to analyse non-verbal cues-face, hand gestures, and body postures	Not engaged at all, nominally engaged, engaged in task, very engaged.	71%	Large classroom
12.	Zheng *et al.* [37]	Faster R-CNN	Student behaviors (hand rising, standing, sleeping)	57.6 (mAP)	Classroom
13.	Gupta *et al.* [38]	Multimodal deep neural network	Anger, fear, happiness, sadness, surprise	91.6	E-learning
14.	Ai *et al.* [39]	Deep engagement recognition network	Emotions based on the Daisee dataset	60%	E-learning
15.	Gupta *et al.* [40]	CNN based on AlexNet architecture	Engaged, disengaged	89.60	Large classroom

4. MULTI-SOURCE ACADEMIC AFFECTIVE ENGAGEMENT DATASET- INTEGRATION AND ANALYSIS

There's a significant need for extensive datasets capturing affective states related to student learning, mainly because studies focusing specifically on these states are quite rare. Most research tends to concentrate on general emotions, overlooking the specific emotional states connected to how students learn. Recognizing this gap, especially in negative emotions, researchers noticed the necessity to consolidate multiple datasets. The scarcity of datasets, particularly those highlighting negative emotions, led to the merging of various datasets. This combination allows us to include diverse cultural viewpoints as these datasets originate from different parts of the world. However, it's important to note that this compilation contains both spontaneous and acted emotions, achieving a balance between these two aspects. This equilibrium acknowledges a wide range of emotional expressions, providing a more comprehensive understanding of how emotions relate to students' learning experiences.

The MAAED dataset creation process involved the consideration of five publicly available datasets-YawDD [41], Daisee [42], many faces of confusion in the wild dataset (MFC-Wild) [43], FER 2013 [44] and BAUM-1 [45]. Although the MAAED datasets included the positive emotion "Concentration," this study primarily focused on negative emotions like "Yawning", "Boredom", "Frustration", and "Confusion". The study did not actively incorporate or analyze the positive emotion "Concentration" within its specific context. Monitoring negative emotions in student's affective states is essential for the well-being and mental health of students. These emotions can significantly impact their overall well-being and hinder effective learning. By monitoring negative emotions, educators can identify students in emotional distress and provide appropriate support and interventions. It also enables personalized assistance, early intervention, and the creation of a positive learning environment that promotes both academic success and overall well-being. For detailed specifics about the datasets utilized in this study, including sample sizes, emotion classes, and other pertinent details, please refer to Table 2.

Table 2. A comparison of emotion recognition datasets for MAAED dataset construction

Dataset	Number of samples	Emotion classes	Other details
YawDD	351 video clips	Normal, talking/singing, and yawning	Videos of people yawning in natural settings. Requires manual annotation of yawning instances within videos.
DAiSEE	9,068 video snippets	Boredom, confusion, engagement, frustration	Four levels of labels for each affective state.
MFC-wild	1,000 video clips	Confusion, anger, disgust	Facial expressions were collected from YouTube and Giphy. Ensures representation of different ethnic groups.
FER2013	35,887 images	Anger, disgust, fear, happiness, sadness, surprise, neutral	Facial expression monitoring in real-world variability.
BAUM-1	1,519 videos	Happiness, anger, sadness, disgust, fear, surprise, boredom, contempt, confusion, concentration, curiosity, complaint	Audio-visual affect and mental state recognition.

4.1. Automatic frame extraction and selection

In building the MAAED, frame extraction plays a crucial role in capturing relevant facial expressions. A combination of five publicly available datasets was selected to predict students' engagement levels in academic environments, focusing on negative emotions. The BAUM-1, Daisee,YawDD and MFC datasets provided videos containing different emotions relevant to students' learning affective states. The initial step involved manually selecting peak frames from each emotion category within the datasets. This ensured that each frame accurately represented a specific affective state. This manual selection process was time-consuming, prompting the need for an automated method to streamline frame extraction (Algorithm 1). To address this challenge, two pre-trained models were utilized for distinct purposes. The VGG-face model was employed for feature extraction. Specifically designed for face recognition tasks, this model extracted meaningful and discriminative features from the input frames.

Algorithm 1. Frame extraction

```
Input:
    • Video file path
    • VGGFace model
    • VGG16 pretrined emotion recognition model
    • Number of frames to select num_frames_to_select
    • Output folder path
Output:
Selected frames are saved in the specified output folder
Begin
   1. Initialize peak_emotion_confidence = 0.0, frame_count = 0
   2. Open video capture object with video file path
   3. WHILE video is open:
         a. Read frame
         b. Resize frame to (224, 224)
         c. Convert frame to RGB
         d. Expand frame dimensions for model input
         e. Predict facial features using VGGFace
         f. Predict emotion using the custom model
         g. IF emotion_confidence > peak_emotion_confidence THEN
              - Update peak_emotion_confidence and peak_emotion_frame_index
         h. Increment frame_count
   4.  Release video capture object
   5.  Determine selected_frame_indices using np.linspace (0, frame_count-1,
       num_frames_to_select)
   6. Open video capture object again
   7. FOR each index in selected_frame_indices:
         a. Set video capture to the specific frame index
         b. Read and resize frame
         c. Save frame to output folder
   8.  Release video capture object
   9. Return success message
End
```

The first model employed was the pre-trained VGG-face model [46]. The VGG-face model is a specialized CNN architecture primarily designed for face recognition applications. Built upon the VGG architecture, this model comprises 16 convolutional layers followed by fully connected layers. It utilizes

smaller-sized convolutional filters (3x3) consistently across the network, enabling a deep architecture while maintaining a manageable number of parameters. Trained on an extensive dataset of facial images, VGG-face focuses on learning comprehensive facial features critical for accurate face identification and distinction. Its pre-trained nature and transfer learning capabilities allow for efficient adaptation to specific face-related tasks with limited training data, speeding up training processes and enhancing performance. Known for its depth and hierarchical feature extraction, the model extracts intricate facial details at varying abstraction levels, including shapes, textures, and nuanced facial attributes. Renowned for robustness, it showcases reliability in recognizing faces amidst diverse conditions such as different expressions, poses, lighting variations, and backgrounds. By leveraging the VGG-face model, meaningful and discriminative features were extracted from input frames of videos, enabling further analysis and processing.

The second model used was a separate pre-trained emotion recognition model with VGG 16 architecture. This model is trained specifically to recognize emotions in images and can classify images into different emotion categories. The features extracted from the VGG-face model were inputted into this emotion recognition model. To accomplish this, a modified model was created by combining the VGG-face model with a dense layer responsible for emotion classification. This modified model predicts the emotion class of an input frame by leveraging the extracted features from the VGG-face model. Sequentially utilizing both models benefits from the VGG-face model's capability to extract rich facial features, which are highly relevant for capturing meaningful patterns related to emotions. The separate pre-trained emotion recognition model performs classification on these extracted features, assigning emotion labels to the input frames. This multi-step approach facilitates a more specialized and accurate emotion recognition process than using a single model alone. This approach addresses the limitations of manual frame selection, making it a valuable contribution to the field of emotion recognition.

Utilizing computer vision techniques via the OpenCV library, systematically iterates through videos, resizing frames to match the input specifications of a pre-trained VGG-face model. Subsequently, it employs this model to predict emotions within each frame, identifying the highest confidence level associated with each engagement category by the VGG 16 pretrained model. The algorithm precisely tracks the frame index showcasing the most intense emotional response, extracting a specific number of frames centered around this highlighted peak. These selected frames, encapsulating the pinnacle of emotional response, are then saved to an output directory for further investigation or analysis. This extraction process not only captures crucial moments reflecting heightened engagement category within videos but also streamlines the subsequent exploration and interpretation of these emotionally significant frames for potential deeper insights or diagnostic purposes.

The dataset has further enlarged by incorporating FER 13 dataset. This brings the total number of images to 16,924 for training, 4,725 for testing, and 3,641 for validation. To ensure the accuracy of the dataset, a manual analysis was conducted to verify the correctness of the automatic frame extraction and preprocessing steps. This analysis helped to ensure that the extracted frames and the applied preprocessing techniques resulted in accurate and reliable data for FER.

The comparison Table 3 provides an overview of various research studies concentrating on face detection in single and multiple frames, academic affective states, techniques employed (such as CNN, SVM, and KNN), datasets utilized (like CK+, FER2013, and DAiSEE dataset), and frame extraction methods (such as regular intervals, fixed-rate extraction, and automated peak frame selection). Each study demonstrates distinct focuses, techniques, and datasets, presenting diverse approaches to detecting facial expressions and recognizing affective states in academic contexts. Notably, the proposed model stands out by utilizing a custom dataset (MAAED) and implementing automated peak frame selection methods for frame extraction.

4.2. Preprocessing

Preprocessing plays a pivotal role in readying data for analysis, particularly in tasks like facial analysis and emotion recognition. Specifically tailored for facial expression datasets extracted from videos, preprocessing encompasses several crucial steps that refine and standardize the input frames. Resizing the frames to a specific dimension ensures uniformity, facilitating consistent analysis across the dataset. Conversion to grayscale simplifies the data while preserving essential facial features, reducing computational complexity without compromising critical visual information. Additionally, normalizing pixel values to a standardized range optimizes data for deep learning models, enhancing convergence during training and enabling models to better generalize across different facial expressions and individuals. Overall, these preprocessing techniques are essential for refining raw data, optimizing its suitability for subsequent analysis, and bolstering the accuracy and reliability of facial analysis and emotion recognition systems.

Table 3. Comparison of the proposed model to previous studies in terms of contributions

Literature	Face detection in a single frame		Academic affective states	Technique	Dataset	Frame extraction
	Single face	Multiple faces				
Wang *et al.* [47]		✓	X	CNN	CK+ and FER2013	___
Yuan [48]	✓	X	✓	MTCNN	Classroom dataset +RAF-DB+ masked dataset	___
Gupta *et al.* [40]	✓	X	X	ResNet-50	RAF-DB+ FER-2013+ OWN Dataset + CK (+)	extracts frames at regular intervals (every 20 sec)
Whitehill *et al.* [26]	✓	X	✓	Boost (BF), SVM, MLR (CERT)	HBCU + UC	frames were extracted at regular intervals and human labelling
Gong and Wei [49]	✓	X	X	KNN	FACS	----
Pabba and Kumar [24]	✓	✓	✓	CNN	Enlarged CSFED	extracts frames at regular intervals (4 frames per second)
Alameda-Pineda *et al.* [12]	✓	X	X	RGB-I3D Network	DAiSEE dataset	15 frames per video
Ai *et al.* [39]	✓	X	X	CNN	DAiSEE dataset	20 frames in each video
Kamath *et al.* [50]	✓	X	X	SVM	DRIShTI WACV 2016	Fixed-rate frame extraction
Mukhopadhyay *et al.* [35]	✓	X	X	CNN	FER2013	1 frame/sec
Thomas and Jayagopi [36]	✓	✓	X	SVM	Custom dataset built by researchers	25 frames per second
Proposed model	✓	✓	✓	CNN	Custom dataset built by researchers (MAAED)	Automated peak frame selection

4.2.1. Face detection

The importance of choosing the right face detection algorithm is crucial in research focused on facial analysis and emotion recognition. Considering the complexities of the research, we found that multi-task cascaded convolutional neural networks (MTCNN), with its multi-stage hierarchical approach, outperformed several other face detection algorithms. Its ability to accurately detect faces across various scales, orientations, and challenging conditions aligns perfectly with our project's requirements. By leveraging stages like proposal network (P-Net), refine network (R-Net), and output network (O-Net), MTCNN excels in identifying a higher number of faces within a single frame, crucial for our goal of accurately recognizing and analyzing facial expressions. This adaptability and robustness of MTCNN play a pivotal role in ensuring the success and effectiveness of our facial analysis and emotion recognition tasks.

The MTCNN algorithm detects faces using cascaded CNNs. It provides bounding box coordinates and landmarks for each detected face, allowing cropping and resizing to 224×224 pixels for focused analysis. The MTCNN algorithm for face detection operates in three major stages: the P-Net, the R-Net, and the O-Net. In the P-Net stage, the convolutional layer is formulated as:

$$Oij = \sum m, \ni (i + m)(j + n)Wmn + b \quad (6)$$

and the ReLU activation function can be represented as:

$$f(x) = max(0, x) \quad (7)$$

The max-pooling operation is given by,

$$Mij = max(Pi: i + s, j: j + s) \quad (8)$$

while the SoftMax function for classification is;

$$pi = \sum jexjexi \quad (9)$$

the R-Net builds on these functions, refining bounding box coordinates using similar layers, and the O-Net further incorporates additional convolutional and fully connected layers for identifying facial landmarks. The final loss function for the MTCNN algorithm combines classification, bounding box, and landmark losses using the formula:

$$L = \lambda 1 Lcls + \lambda 2 Lbox + \lambda 3 Llandmark \quad (10)$$

in this context, L represents the loss function. λ1, λ2, and λ3 are the weights of the classification, bounding box, and landmark losses *Lcls* is the classification loss, *Lbox* is the bounding box loss and *Llandmark* is the landmark loss. Collectively, these mathematical formulations represent the core computations of the MTCNN algorithm, enabling precise face detection and landmark identification, applicable in various real-world scenarios. This remarkable algorithm enhances our understanding of facial expressions and emotions. MTCNN stands as a vital instrument, simplifying our investigation into how people display emotions through their facial features.

4.2.2. Resizing

Resizing an image means changing its size—making it smaller or larger. In computer tasks like facial analysis or emotion recognition, resizing images is essential for consistency. It helps make all images the same size, making it easier for computers to understand and process them uniformly. This process also affects how quickly the computer can work and the level of detail in the images. This work focused on resizing all frames to the standardized dimensions of 224x224, improves model performance, and efficiency, and facilitates transfer learning.

4.2.3. Grey-scale conversion

Converting images to grayscale holds multiple advantages in the context of facial analysis and emotion recognition. This conversion process eliminates color information, emphasizing essential structural and textural elements present in facial images. By removing color variations, grayscale images become more standardized, allowing algorithms to focus more precisely on the intrinsic features of the face, such as lines, shapes, and contrasts.

Moreover, grayscale conversion simplifies the computational workload by reducing the complexity of the data. It decreases the image's file size, streamlining the analysis process and enhancing computational efficiency. This simplification facilitates quicker processing, making it easier for algorithms to recognize and extract key facial features necessary for accurate emotion recognition. Additionally, the absence of color distractions in grayscale images aids in reducing noise and enhancing the clarity of facial attributes. The focused attention on structural details in grayscale images assists in robustly identifying emotional cues, contributing significantly to the accuracy and reliability of emotion recognition systems.

4.2.4. Train-test-validation splitting

Dataset splitting into training, testing, and validation sets (60%, 20%, and 20%) aids in model training, evaluation, and fine-tuning. It prevents overfitting and allows performance monitoring, hyper-parameter tuning, and model selection. Overall, these pre-processing steps enhance the dataset, enable effective facial analysis, and support robust emotion recognition and engagement monitoring. Table 4 presents a comprehensive overview of the MAAED following the completion of necessary pre-processing steps.

Table 4. Dataset overview after preprocessing

Engagement category	Train (60)	Test (20)	Validation	Total
Boredom	3387	1130	1129	5646
Concentrating	2322	774	774	3870
Yawning	2742	914	914	4570
Confused	3063	1021	1021	5105
Frustration	2974	991	992	4957

4.3. Training and hyperparameter overview

The newly developed CNN model is designed for FER using the MAAED dataset. It consists of convolutional layers, batch normalization, activation functions, max-pooling layers, and fully connected layers with regularization. The model leverages these layers to extract meaningful features from facial

images and accurately classify emotions. The MAAED dataset was trained using four popular pre-trained models: VGG16, Inception V3, ResNet50 and ResNet101. These models have been widely used in computer vision tasks and have shown excellent performance in various domains. During the training process, the models were fine-tuned on the MAAED dataset to adapt for the specific task of FER and student engagement classification. The results of the evaluation of the pre-trained models on the MAAED dataset are very promising. The models achieved high precision and recall values for individual engagement and their effectiveness in correctly identifying and classifying different levels of student engagement. These results suggest models of complex patterns and features associated with student engagement from facial expressions.

Table 5 exhibits the architecture of the CNN model utilized for the study. The model is organized into five main blocks, each containing a sequence of layers performing specific operations. The model takes as input a 48x48 pixel RGB image. Block 1 starts with a Conv2D layer that employs 32 filters of size 3×3 to the input image, followed by a MaxPooling2D layer that decreases the spatial dimensions of the output from the Conv2D layer. Block 2 and Block 3 follow a similar structure but with an increasing number of filters applied in the Conv2D layers. Each block doubles the number of filters from the previous block, using 64 and 128 filters respectively. Block 4 further increases the complexity of the model by applying 256 filters in its Conv2D layers. Block 5 begins with a Flatten layer, which reshapes the multi-dimensional output from Block 4 into a one-dimensional array. This is followed by three dense layers, which perform the final classification. The initial two dense layers utilize 'ReLU' activation and incorporate L1L2 regularization, batch normalization, and a dropout rate of 0.5 to mitigate the risk of overfitting. The final dense layer uses a 'SoftMax' activation function to output the class probabilities. This architecture, with its hierarchical structure and increasing complexity, allows the model to effectively learn from the input images and perform accurate image classification. The use of regularization techniques and dropout also helps to prevent overfitting, ensuring that the model generalizes well to unseen data.

Table 5. Detailed architecture of the proposed CNN model

Block	Layer type	Details
	Input layer	48×48×3 (RGB image)
Block 1	Conv2D	32 kernels (3×3), stride 1×1, padding 'same', activation 'ReLU', batch normalization
	MaxPooling2D	Pool size 2×2, stride 2×2, padding 'valid'
Block 2	Conv2D	64 kernels (3×3), stride 1×1, padding 'same', activation 'ReLU', batch normalization
	Conv2D	64 kernels (3×3), stride 1×1, padding 'same', activation 'ReLU', batch normalization
	MaxPooling2D	Pool size 2×2, stride 2×2, padding 'valid'
Block 3	Conv2D	128 kernels (3×3), stride 1×1, padding 'same', activation 'ReLU', batch normalization
	Conv2D	128 kernels (3×3), stride 1×1, padding 'same', activation 'ReLU', batch normalization
	MaxPooling2D	Pool size 2×2, stride 2×2, padding 'valid'
Block 4	Conv2D	256 kernels (3×3), stride 1×1, padding 'same', activation 'ReLU', batch normalization
	Conv2D	256 kernels (3×3), stride 1×1, padding 'same', activation 'ReLU', batch normalization
	MaxPooling2D	Pool size 2×2, stride 2×2, padding 'valid'
Block 5	Flatten	-
	Dense	512 neurons, activation 'ReLU', L1L2 regularization (L1=0.01, L2=0.01), batch normalization, dropout 0.5
	Dense	256 neurons, activation 'ReLU', L1L2 regularization (L1=0.01, L2=0.01), batch normalization, dropout 0.5
	Dense	num_classes neurons, activation 'softmax'

Hyper-parameters in machine learning are predefined parameters that control the behaviour and characteristics of a learning algorithm. They are set before the learning process and are not learned from the data. The selection and optimization of hyper-parameters, known as hyper-parameter tuning, are crucial for achieving optimal model performance and generalization. Table 6 represents the hyper-parameter configuration for the emotion recognition model used in the study. In the realm of FER employing CNN, the model consists of convolutional layers, batch normalization, ReLU activation, max pooling for downsampling, fully connected layers with batch normalization, ReLU activation, and dropout regularization. The output layer utilizes SoftMax activation for class probabilities. The training process involves the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric. Careful selection and tuning of hyper-parameters are essential to avoid overfitting or underfitting and to build effective and well-performing machine learning models.

4.4. Image augmentation

In the context of computer vision and the realm of deep learning, image augmentation is a powerful technique that plays a vital role in enhancing the performance and generalization of machine learning models. It involves applying various transformations to original images, creating augmented versions of the data, and significantly increasing the size and diversity of the training dataset. The primary objective of image augmentation is to expose the model to a wider range of variations and perspectives that could occur in real-

world scenarios. By presenting the model with augmented data during training, it becomes more robust and less prone to overfitting to the specific patterns present in the training set. The augmentation process incorporates a variety of techniques, each designed to introduce specific variations to the images. Table 7 is a closer look at the image augmentation techniques used in this specific implementation: Image augmentation involves several techniques to enhance the diversity and robustness of a training dataset for computer vision and deep learning tasks. Rescaling (1.0 / 255.0) normalizes pixel values in images from a 0-255 range to a 0-1 range, thereby improving the learning speed and stability. The shear range parameter, set to 0.2, shifts each point in a fixed direction, preserving parallel lines to generate realistic distortions. A zoom range of 0.2 randomly magnifies images within a specified range, enabling the model to recognize different pattern scales. The horizontal flip option, set to True, mirrors the image along its vertical axis, which is particularly useful for model generalization when dealing with a small training dataset. A rotation range of 20 degrees randomly rotates the image, allowing the model to learn and recognize patterns at different orientations. Width and height shift ranges (both set to 0.2) move the image pixels in the horizontal and vertical directions respectively, enabling the model to recognize the object of interest even if it is not centered in the image. Together, these techniques work in tandem to mitigate overfitting and improve the model's capacity to generalize from the training data to unseen data.

Table 6. Hyper-parameter configuration for the emotion recognition model used in the study

Hyper-parameter	Value	Description
Number of filters	32, 64, 128, 256	The count of filters applied in each convolutional layer.
Kernel size	3×3	The kernel size in each convolutional layer.
Pooling size	2×2	The pooling window size utilized in each max pooling layer.
Units within the fully connected layers	512, 256	The unit count for each fully connected layer.
Regularization strength	0.01	L1 and L2 regularization strength enforced in the fully connected layers.
Dropout rate	0.5	percentage of neurons randomly turned off during training
Activation function	ReLU	The mathematical function that determines the output of a neuron based on its input
Optimizer	Adam	Adjust the model's parameters and minimize the loss function
Loss function	Categorical cross-entropy	The loss function is used to evaluate the model.
Metrics	Accuracy, Loss, F1-score, precision, recall	The metrics that are used to track the performance of the model.

Table 7. Image augmentation techniques used

Augmentation technique	Value
Rescaling	1.0 / 255.0
Shear	0.2
Zoom	0.2
Rotation range	20
Width shift range	0.2
Height shift range	0.2

4.5. Evaluation metrics

Evaluation metrics are employed to appraise the performance and efficiency of a variety of systems, models, algorithms, or processes. These metrics offer quantitative measures that facilitate the assessment of a system or model's performance and also allow for comparisons between different methodologies. A confusion matrix (Figure 2), is a structured table that depicts the quantities of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by a classification model. This matrix provides a thorough breakdown of the model's performance.

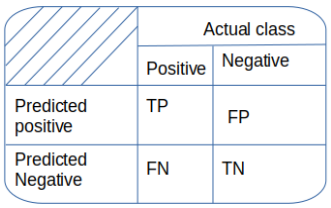


Figure 2. Structure of confusion matrix

Among the evaluation metrics, accuracy is the most frequently utilized for evaluating classification models. Accuracy quantifies the ratio of accurate predictions made by the model and is calculated by dividing the number of correct predictions by the total predictions. Despite its simplicity in interpretation, accuracy may not offer a comprehensive understanding of model performance, particularly when handling imbalanced datasets.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (11)$$

Precision, in binary classification tasks, quantifies the proportion of correctly predicted positive instances among all instances predicted as positive. This metric emphasizes the accuracy of positive predictions. A high precision value signifies a low occurrence of FP, indicating the model's proficiency in accurately identifying positive instances.

$$Precision = TP/(TP + FP) \quad (12)$$

Recall, sometimes referred to as sensitivity or the TP rate, serves as another metric employed in binary classification. It quantifies the proportion of correctly predicted positive instances among all actual positive instances. Recall assesses the model's capability to identify all positive instances. A high recall value indicates a low occurrence of FN, signifying the model's effectiveness in capturing positive instances.

$$Recall = TP/(TP + FN) \quad (13)$$

The F1-score represents the harmonic mean of precision and recall, offering a consolidated metric that strikes a balance between these two aspects. This score is especially valuable in scenarios involving imbalanced datasets, as it takes into account both FP and FN. The F1-score's scale ranges from 0 to 1, with 1 representing the optimal value. A higher F1-score signifies a superior equilibrium between precision and recall.

$$F1 = 2 * (Precision * Recall)/(Precision + Recall) \quad (14)$$

5. RESULTS AND ANALYSIS

The proposed CNN model has high recall and precision scores on all emotion classes. This suggests that the model can accurately identify the emotions in facial expressions, even when the expressions are subtle or ambiguous. The proposed model exhibits overall performance on the MAAED dataset, with precision, recall, and F1-score values are close to 0.96. It shows slightly better accuracy in classifying boredom and yawning confusion and frustration. The macro average, precision, recall, and F1-score are all 0.96, indicating consistent performance across all classes.

The weighted average precision, recall, and F1-score are also 0.96, indicating that the model is imbalance. The model demonstrates impressive performance across various evaluation metrics, showcasing its effectiveness in image classification. With an accuracy of 97%, it consistently classifies images with a high level of accuracy. The precision for each category is also commendable, indicating that the model excels at correctly identifying images within each class. Moreover, the model exhibits high recall, successfully capturing the vast majority of images that truly belong to a particular class.

The F1-score, which combines precision and recall, further emphasizes the model's ability to strike a balance between these metrics (Table 7). An encouraging aspect is that the model does not suffer from overfitting, as evidenced by its high accuracy in the validation data. This indicates that it generalizes well and can perform reliably on unseen data. Additionally, the model's efficiency is noteworthy, as it swiftly classifies images within a time frame of just 10 milliseconds. This characteristic makes it suitable for real-time applications where quick image processing is vital. Interpretability is another strength of the model, as it can be understood and debugged effectively. This implies that users have the capability to understand how the model makes decisions and can provide explanations for its predictions to others. The model showcases its robustness by demonstrating the capacity to adapt to new data and maintain consistent performance, even when the dataset contains noise or variations. This robustness is essential for the model to consistently and effectively perform in real-world situations. In Figure 3, the proposed CNN model's positive predictions are showcased using the MAAED dataset. These predictions indicate instances where the model correctly identifies and classifies the engagement level from randomly selected facial images.

Figure 3. Positive predictions of the model

The confusion matrix (Figure 4) shows the performance of a CNN system for facial engagement monitoring. The values in the matrix represent the percentage of people in each category. For example, the value in the TP cell is 0.93, which means that 93% of the people who are confused, made correct identification by the model. The value in the FP cell is 0.01, which means that 1% of the people who are not confused are incorrectly identified by the model as being confused. However, there is some room for improvement. The FN rate is not as high as it could be. This means that the model is missing some of the confused people. If the model could improve its FN rate, it would be able to identify more people who are confused. Overall, the confusion matrix shows that the model is performing well. However, there is some room for improvement.

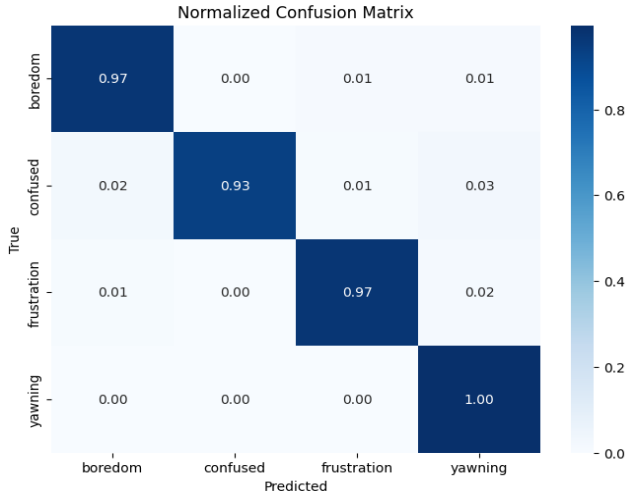


Figure 4. Normalized confusion matrix of the proposed model

The graphs (Figure 5) offer a holistic view of how well each model performs in accurately identifying and classifying various engagement levels. Notably, the MAAED-based proposed CNN stands out, demonstrating superior performance across the different evaluation metrics. The information in the Tables 8 and 9 shows that the suggested CNN model performs really well at recognizing quick facial expressions linked to different feelings like confusion, boredom, yawning, and frustration. When we compare it to well-known models like VGG16 and Inception v3, proposed model consistently does better in understanding these expressions. It's good at noticing small details in faces that other general models might miss. This success is because our model was trained on specific data related to these expressions, maybe from a larger or more varied set of examples. This suggests our CNN model could be a useful tool for quickly spotting how engaged someone is, and we should explore using it more in different areas.

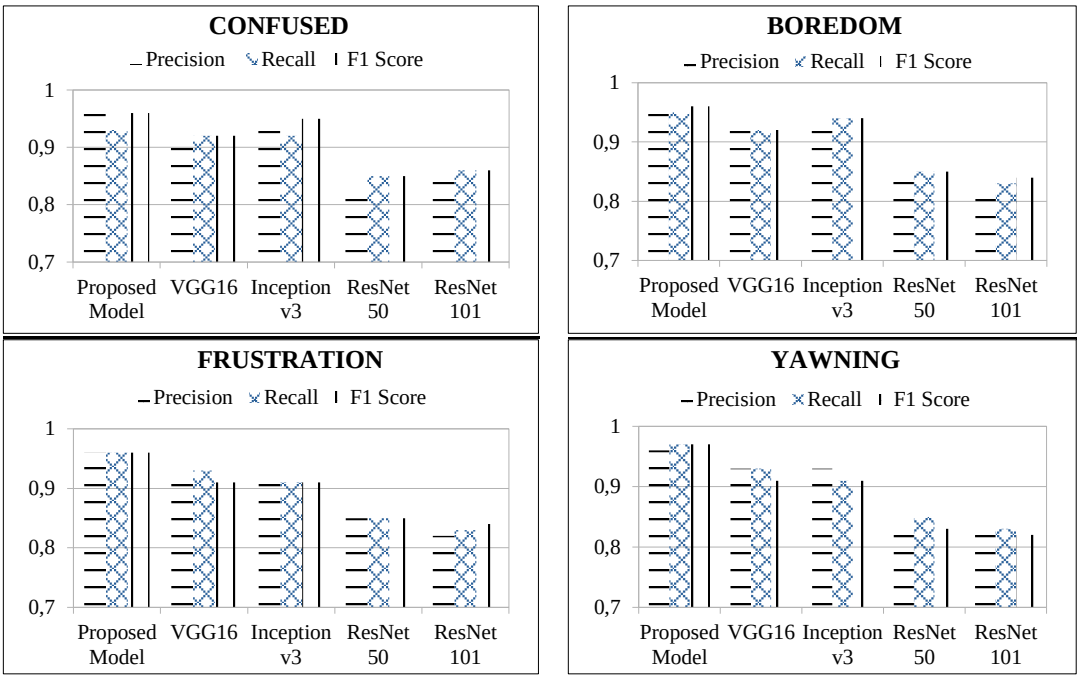


Figure 5. Graphical analysis of evaluation metrics for different engagement categories

Table 8. Classification report of the proposed model

Emotion	Precision	Recall	F1-score	Support
Boredom	0.96	0.97	0.965	1208
Confused	0.97	0.92	0.945	1084
Frustration	0.95	0.97	0.96	1287
Yawning	0.96	0.96	0.96	1146
Accuracy	–	–	0.97	4725
Macro Avg	0.96	0.96	0.96	4725
Weighted Avg	0.96	0.96	0.96	4725

Table 9. Comparative analysis of precision, recall, and F1-scores for different engagement categories

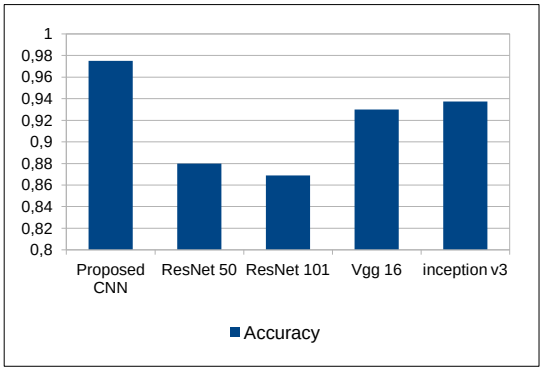
	Confused			Boredom			Yawning			Frustration		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Proposed CNN	0.97	0.93	0.96	0.95	0.95	0.96	0.96	0.97	0.97	0.96	0.96	0.96
VGG16	0.92	0.92	0.92	0.93	0.92	0.92	0.93	0.93	0.91	0.93	0.93	0.91
Inception v3	0.94	0.92	0.95	0.94	0.94	0.94	0.93	0.91	0.91	0.91	0.91	0.91
ResNet 50	0.83	0.85	0.85	0.84	0.85	0.85	0.84	0.85	0.83	0.85	0.85	0.85
ResNet 101	0.86	0.86	0.86	0.82	0.83	0.84	0.82	0.83	0.82	0.82	0.83	0.84

Figure 6 illustrates the loss and accuracy analysis for the proposed CNN, ResNet50, ResNet101, and VGG16 models. As shown in Figure 6(a), the proposed CNN achieves higher accuracy compared to the other models, indicating more effective learning and improved prediction performance. Figure 6(b) further shows that the proposed CNN yields lower loss values, reflecting better convergence during training. Among the pre-trained models, VGG16 demonstrates relatively lower loss and higher accuracy, indicating more accurate predictions compared to ResNet50 and ResNet101. In contrast, ResNet50 and ResNet101 do not exhibit competitive performance and may require further optimization to enhance their learning and prediction capabilities. Overall, the proposed CNN demonstrates competitive performance and, with additional training, has the potential to achieve even higher accuracy.

Accuracy (%)

Epoch	Proposed CNN	ResNet 50	ResNet 101	VGG 16	Inception V3
1	68	57.4	48	74	68.75
5	76	72	53	76	78.12
10	75	78	57	79	81.25
15	74	80	61	83	84.38
20	86	80	63	84	86.38
25	85	81	65	90	85
30	91	82	72	92	87.5
35	94	84	79	93	90.62
40	94	85	86	93	92
45	96	88	86	93	92.6
50	97.5	88	87	93	93.75

(a)



Loss

Epoch	Proposed CNN	ResNet 50	ResNet 101	VGG 16	Inception V3
1	3.188590	0.973	0.876	0.199	0.1375
5	1.216084	0.5111	0.743	0.177	0.1264
10	0.580635	0.4439	0.488	0.147	0.1102
15	0.536500	0.3988	0.432	0.141	0.1098
20	0.423535	0.3613	0.412	0.136	0.1091
25	0.332	0.3484	0.46	0.131	0.1079
30	0.251	0.3306	0.395	0.130	0.1088
35	0.209	0.3154	0.392	0.127	0.1078
40	0.176	0.3015	0.39	0.125	0.1082
45	0.113	0.2894	0.389	0.123	0.1087
50	0.037	0.2845	0.39	0.123	0.108

(b)

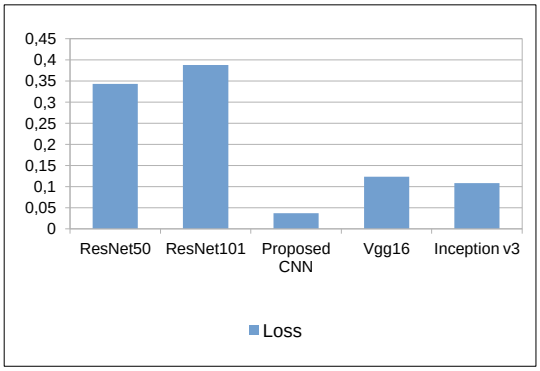


Figure 6. Performance comparison of various models in terms of (a) accuracy and (b) loss

Accuracy and loss performance of the proposed model during training as shown in Figure 7. The training results graph reveals (Figure 7(a)) an upward trend in the model's accuracy, culminating in a plateau after approximately 50 epochs. This pattern may signify that the model has reached a point where it is no longer gaining substantial insights from the training data, a phenomenon that can lead to overfitting. In the context of deep learning, overfitting occurs when a model becomes excessively adapted to the training data, leading to difficulties in performing effectively on new, unseen data valuable metric to assess overfitting is the gap between training accuracy (performance on training data) and validation accuracy (performance on new data). A wide gap usually signals overfitting, but in this case, the relatively narrow difference suggests that the overfitting is not extreme. While the results indicate that the model is relatively well-balanced and not severely overfitting, the plateau in accuracy after 50 epochs does imply that there may be room for improvement. Additional training data or adjustments to the training process might help the machine to learn more effectively, maximizing its performance on both known and unknown data.

The initial accuracy (Figure 7(b)) of approximately 85% indicated a reasonably good starting point for the model's predictive performance, yet suggested room for improvement. Notably, there was a disparity between training and validation accuracies, hinting at potential mild overfitting, where the model memorized the training data more than it generalized to new, unseen examples. However, after the introduction of augmentation techniques aimed at diversifying the dataset and enhancing the model's ability to generalize, significant improvements were observed. Post-augmentation, the model showcased enhanced generalization capabilities, as evidenced by a reduction in overfitting tendencies. This improvement was reflected in a decrease in the disparity between training and validation accuracies, indicating that the model was better equipped to make accurate predictions on unseen data. Consequently, the comparative analysis before and after augmentation highlighted how these techniques positively influenced the model's ability to generalize beyond the training dataset, resulting in a more robust and accurate predictive performance.

In Figure 7(c) demonstrates the loss curves observed during the training and validation sessions of the proposed model. The graph depicts how the loss values evolve over the course of training across 50 epochs. The convergence of the loss curves for both training and validation sets signifies that the model has effectively learned the underlying patterns within the data without excessively fitting to the training set. This convergence demonstrates good generalization capabilities, indicating that the model can make accurate predictions not only on the training data but also on unseen or validation data. This loss metric assesses

the disparity between the actual distribution Y and the predicted distribution Y^. Mathematically, it is defined as:

$$L(Y, \hat{Y}) = -\Sigma_i^n Y_i log(\hat{Y}_i) \quad (15)$$

here, Yi represents the actual label for class i, often encoded as one-hot, while Y^i signifies the predicted probability associated with class i. In simple terms, the categorical cross-entropy loss quantifies how much the predicted probabilities differ from the actual labels.

A lower loss value indicates a better model prediction, and conversely, a higher value signifies a poor prediction. The logarithmic function imposes a substantial penalty when the model makes highly confident yet incorrect predictions. Impact of L1L2 regularization on model performance over several epochs after applying augmentation techniques. There are two distinct curves: “Before L1L2” and “After L1L2” (Figure 7(d)).

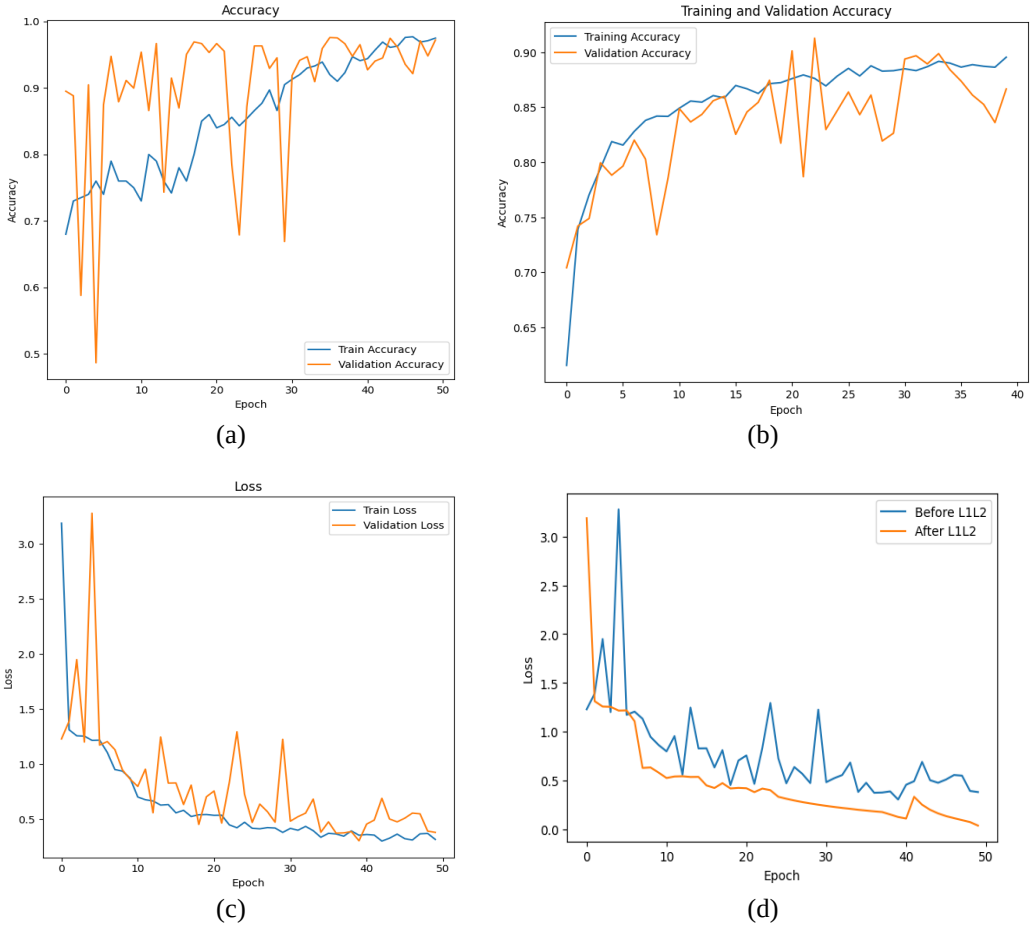


Figure 7. Accuracy and loss graph of proposed model; (a) accuracy of proposed model after L1L2 regularization and augmentation, (b) accuracy before augmentation, (c) train and validation loss of proposed model, and (d) loss curve before and after L1L2

As epochs progress, the loss value for the “After L1L2” curve noticeably decreases compared to its counterpart, underscoring the efficacy of L1L2 regularization. This method discourages the model from having excessively large weights, thus mitigating the problem of overfitting, which occurs when a model becomes overly focused on its training set and struggles to perform effectively on new datasets. The graph clearly demonstrates the value of L1L2 regularization in enhancing a model’s generalization capabilities and performance.

This model showcases innovation through its integration of a MAAED, reflecting a comprehensive effort to encompass diverse emotional expressions prevalent in educational environments. Notably, the model

introduces the concept of "Automated Peak Frame Selection," indicating a strategic approach to frame extraction that targets significant emotional moments, potentially capturing the essence of engagement. In essence, the proposed model offers a tailored and promising avenue for understanding the nuanced emotional dynamics within academia, underlining its potential significance in advancing the field of emotion recognition research.

6. CONTRIBUTIONS

The present work introduces several novelties in the field of recognition and its application in academic environments. Firstly, a new dataset was curated specifically for predicting the engagement level of students in educational settings. This dataset holds great value as it facilitates the training and evaluation of emotion recognition models tailored to the unique demands of academic contexts. Secondly, the researchers adopted a two-step approach for extracting frames from video datasets to for emotion recognition. This innovative strategy utilizes two models, enhancing accuracy and automating the frame extraction process. Compared to relying just on one model, the outcome is a method that is more effective and exact. Beyond emotion recognition, the study demonstrates the potential applications in educational domains. It highlights how these technological advancements can lead to improve the design of educational materials and personalize learning experiences. Moreover, the ability to identify students who may be facing academic challenges opens up new possibilities for targeted student support and intervention. Proposed model's performance comparison with baseline models on the MAAED dataset consistently demonstrates positive outcomes across various pretrained models and our proposed model (Figure 8).

ACCURACY (%) COMPARISON OF PROPOSED WORK

MAAED based models	VGG	93
	ResNet	88
	Inception	93.5
	Proposed model	97.5
Gupta *et al.* (2021)	Inception	89.11
	ResNet	92.32
	VGG	90.14
Shen *et al.* (2020)	VGG	47
	Proposed model	56
Dukic *et al.* (2022)	Inception	78.97
	ResNet	76.456
	Proposed model	65.57

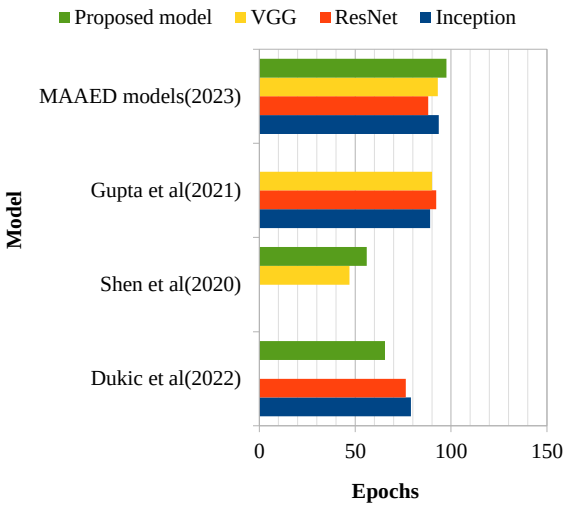


Figure 8. Analyzing the accuracy of the proposed model

The MAAED dataset proves highly effective in identifying student affective states, evident from the overall strong results obtained. Notably, the ResNet 50 pretrained model exhibits exceptional proficiency when applied to Gupta *et al.* [38] daisee dataset, surpassing its performance on MAAED. Conversely, VGG and Inception V3 show a slight advantage, particularly within the MAAED dataset. Meanwhile, Zheng *et al.* [37] research showcases AI's practical implementation with automated detection capabilities, successfully identifying behaviors like hand-raising, standing, and sleeping among students. This highlights AI's usefulness in understanding various classroom dynamics. It's worth noting that Dukić and Krzic [51] focused on six basic emotion categories rather than student learning-based emotions, conducting experiments on CK+, FER-2013, and SFEW datasets with pretrained models. Researchers' opinions suggest that instead of solely focusing on evaluating performance on individual datasets, prioritizing datasets centered around specific emotion learning could yield more beneficial results. In conclusion, the presented work contributes significantly to the broader field of education and student support by bringing together novel dataset creation, a two-step approach for accurate recognition, and diverse practical applications (Table 6). Its impact extends beyond emotion recognition, promising to enhance the academic experience and foster student success in various educational settings [37].

7. DISCUSSION

The current CNN models for emotion recognition in the classroom are limited in their ability to capture the full range of emotions that students experience. They typically focus on predicting and classifying negative emotions. However, classrooms are rich environments where a spectrum of emotions is present, including positive emotions such as engagement, interest, and excitement. Figure 9 illustrates a pertinent example of this limitation. Despite the classroom scene encompassing individuals actively participating and engaged, the current model, which has been primarily trained to recognize and predict specific emotions, may not possess the capacity to accurately forecast the behaviours of active participants consequently, these instances of active participation might be misclassified as erroneous predictions due to the model's limited exposure to such scenarios. However, it is noteworthy that the model may still demonstrate proficiency in accurately predicting other distinct emotions that it has been trained on. This observation underscores the specialized nature of the model's capabilities, emphasizing the need for a more comprehensive training regimen that encompasses a broader spectrum of emotions and behaviours to enhance its predictive accuracy across diverse classroom dynamics. Future CNN models should be enlarged to include positive feelings in order to better comprehend the emotional dynamics in the classroom. This would allow the models to identify moments of active participation and enthusiasm, which are important indicators of student engagement. The models could also be used to identify students who are experiencing positive emotions and then provide them with opportunities to share their excitement with their classmates. Every student stands to gain from a more positive and collaborative learning environment as a result of this. In addition to incorporating positive emotions, future CNN models should also be expanded to use multimodal data. This would allow the models to get a more comprehensive understanding of students' emotions by taking into account of their facial expressions, body language, tone of voice, and words. Finally, future CNN models should be made more interpretable. This would allow educators to better understand how the models are making their predictions, and it would make it easier to use the models to inform educational decision-making. By incorporating positive emotions, using multimodal data, and making the models more interpretable, future CNN models for emotion recognition in the classroom could become powerful tools for educators. They possess the capability to improve learning environments, rendering them more supportive and engaging for every student.

The study brings new ideas by creating a special dataset and using a unique method to recognize emotions in students. However, it has some important limitations. It only looks at faces and doesn't consider other modalities, like their voice or body language. Also, it doesn't talk about problems if the face is partly covered or the picture is unclear. This might make the results less accurate in real classrooms. Another critical concern is the sensitivity of the model to occlusions or unclear facial frames. The study fails to address how occlusions or unclear facial images might impact the accuracy of the emotion recognition models. This limitation raises doubts about the model's robustness in real-world scenarios where facial obstructions or poor image quality are common. Additionally, the study doesn't adequately address the need for real-time processing. The lengthy processing time required for accurate recognition remains a significant obstacle that needs to be substantially improved for practical implementation in real-time educational settings. The exclusive reliance on facial recognition, oversight of biases, insensitivity to occlusions, long processing times, and the absence of real-time image analysis underscore the need for further development and refinement in future studies to ensure the practicality and efficacy of these technologies in real educational scenarios.

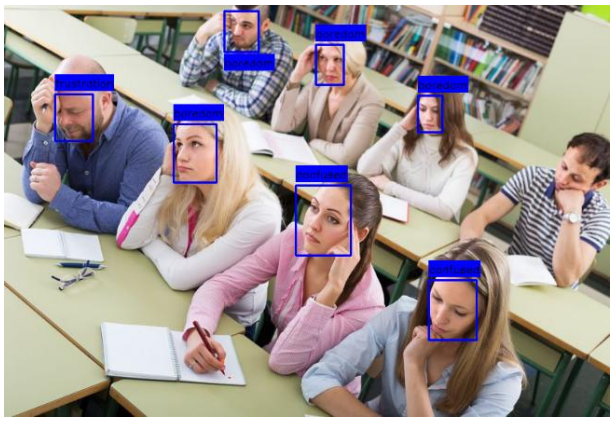


Figure 9. Prediction on a random classroom image

8. ETHICAL CONSIDERATIONS

The researchers involved in this study meticulously adhered to ethical standards while utilizing publicly available datasets. Prior to dataset integration, comprehensive permissions were obtained to ensure compliance with legal and ethical requirements. The research team followed established guidelines and protocols for data acquisition, ensuring proper consent, authorization, and acknowledgment of the original dataset sources. All necessary measures were taken to respect copyright, protect user privacy, and maintain data confidentiality throughout the research process. The researchers have diligently documented the sources and permissions obtained, ensuring full compliance and ethical considerations in using these datasets.

9. CONCLUSION

This study proposes a CNN-based approach using the MAAED for facial engagement monitoring. The CNN model accurately classifies engagement categories like Boredom, Confused, Frustration, and Yawning, extracting discriminative features for precise predictions. The contributions include creating MAAED, a rich dataset reflecting academic engagement, and developing an objective method for assessing student engagement. This has the potential to enhance educational environments and improve student outcomes through personalized learning experiences. The future work of this study includes, further expanding the MAAED dataset to include more diverse facial expressions and engagement levels. We also plan to explore other deep-learning techniques for facial engagement monitoring, such as attention-based models and recurrent neural networks. We are confident that our efforts hold the promise of making a substantial impact on educational technology and improving the learning experiences of students.

REFERENCES

[1] S. Craig, A. Graesser, J. Sullins, and B. Gholson, "Affect and learning: an exploratory look into the role of affect in learning with AutoTutor," *Journal of Educational Media*, vol. 29, no. 3, pp. 241–250, Oct. 2004, doi: 10.1080/1358165042000283101.

[2] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learning and Instruction*, vol. 22, no. 2, pp. 145–157, Apr. 2012, doi: 10.1016/j.learninstruc.2011.10.001.

[3] R. Wang, J. Cao, Y. Xu, and Y. Li, "Learning engagement in massive open online courses: a systematic review," *Frontiers in Education*, vol. 7, Dec. 2022, doi: 10.3389/feduc.2022.1074435.

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

[5] S. S. Khan, A. Abedi, and T. Colella, "Inconsistencies in the definition and annotation of student engagement in virtual learning datasets: a critical review," *arXiv*, 2022, [Online]. Available: http://arxiv.org/abs/2208.04548.

[6] P. Bhardwaj, P. K. Gupta, H. Panwar, M. K. Siddiqui, R. Morales-Menendez, and A. Bhaik, "Application of deep learning on student engagement in e-learning environments," *Computers & Electrical Engineering*, vol. 93, p. 107277, Jul. 2021, doi: 10.1016/j.compeleceng.2021.107277.

[7] B. Perumal, P. Nagaraj, T. S. Narsimha Charan, Y. V. S. Saideepak, C. V. Vignesh Reddy, and S. Nagendra, " Student Engagement Detection in Classroom using deep CNN-based learning approach," in *Proceedings of the 8th International Conference on Communication and Electronics Systems, ICCES 2023*, Jun. 2023, pp. 1233–1238, doi: 10.1109/ICCES57224.2023.10192809.

[8] M. Bustos-López, N. Cruz-Ramírez, A. Guerra-Hernández, L. N. Sánchez-Morales, N. A. Cruz-Ramos, and G. Alor-Hernández, "Wearables for engagement detection in learning environments: a review," *Biosensors*, vol. 12, no. 7, p. 509, Jul. 2022, doi: 10.3390/bios12070509.

[9] A. Apicella, P. Arpaia, M. Frosolone, G. Improta, N. Moccaldi, and A. Pollastro, "EEG-based measurement system for monitoring student engagement in learning 4.0," *Scientific Reports*, vol. 12, no. 1, p. 5857, Apr. 2022, doi: 10.1038/s41598-022-09578-y.

[10] G. R. Green, "Text based discussions: an approach to teach reading comprehension comprehension," 2022, [Online]. Available: https://scholarworks.gvsu.edu/gradprojects.

[11] M. Saneiro, O. C. Santos, S. Salmeron-Majadas, and J. G. Boticario, "Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches," *The Scientific World Journal*, vol. 2014, pp. 1–14, 2014, doi: 10.1155/2014/484873.

[12] X. Alameda-Pineda *et al.*, "SALSA: a novel dataset for multimodal group behavior analysis," *arXiv*, 2015.

[13] Y. Nie, H. Luo, and D. Sun, "Design and validation of a diagnostic MOOC evaluation method combining AHP and text mining algorithms," *Interactive Learning Environments*, vol. 29, no. 2, pp. 315–328, Feb. 2021, doi: 10.1080/10494820.2020.1802298.

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

[15] P. Sharma *et al.*, "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning," in *Communications in Computer and Information Science*, vol. 1720 CCIS, 2022, pp. 52–68.

[16] J. Conner, M. Posner, and B. Nsowaa, "The relationship between student voice and student engagement in urban high schools," *The Urban Review*, vol. 54, no. 5, pp. 755–774, Dec. 2022, doi: 10.1007/s11256-022-00637-2.

[17] S. Slater, J. Ocumpaugh, R. Baker, M. V. Almeda, L. Allen, and N. Heffernan, "Using natural language processing tools to develop complex models of student engagement," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, Oct. 2017, vol. 2018-Janua, pp. 542–547, doi: 10.1109/ACII.2017.8273652.

[18] Z. Gu, V. C. Zarubin, K. R. M. Steinmetz, and C. Martsberger, "Heart rate variability in healthy subjects during monitored, short-term stress followed by 24-hour cardiac monitoring," *Frontiers in Physiology*, vol. 13, Jun. 2022, doi: 10.3389/fphys.2022.897284.

[19] A. Fraguiero, A. Coutrot, E. Bannier, and C. Cury, "Pilot study : eye-tracking and skin conductance to monitor task engagement during bimodal neurofeedback," 2023.

[20] R. H. Nabil, A.-A.-A. Rupai, M. Barid, A. Sami, and M. N. Hossain, "An intelligent examination monitoring tool for online student evaluation," *Malaysian Journal of Science and Advanced Technology*, pp. 122–130, 2022, doi: 10.56532/mjsat.v2i3.62.
[21] D. Preuveneers, G. Garofalo, and W. Joosen, "Cloud and edge based data analytics for privacy-preserving multi-modal engagement monitoring in the classroom," *Information Systems Frontiers*, vol. 23, no. 1, pp. 151–164, Feb. 2021, doi: 10.1007/s10796-020-09993-4.
[22] T. S. Ashwin and R. M. R. Guddeti, "Unobtrusive behavioral analysis of students in classroom environment using non-verbal cues," *IEEE Access*, vol. 7, pp. 150693–150709, 2019, doi: 10.1109/ACCESS.2019.2947519.
[23] J. Bidwell, H. F.-B. R. Methods, and U. 2011, "Classroom analytics: measuring student engagement with automated gaze tracking," *Researchgate.Net*, doi: 10.13140/RG.2.1.4865.6242.
[24] C. Pabba and P. Kumar, "An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition," *Expert Systems*, vol. 39, no. 1, Jan. 2022, doi: 10.1111/exsy.12839.
[25] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *Nebraska Symposium on Motivation*, vol. 19, pp. 207–283, 1971.
[26] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, Jan. 2014, doi: 10.1109/TAFFC.2014.2316163.
[27] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004, doi: 10.1023/B:VISI.0000013087.49260.fb.
[28] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features," *Eurasip Journal on Image and Video Processing*, vol. 2017, no. 1, p. 80, Dec. 2017, doi: 10.1186/s13640-017-0228-8.
[29] L. B. Krithika and L. Priya, "Student Emotion recognition system (SERS) for e-learning improvement based on learner concentration metric," *Procedia Computer Science*, vol. 85, pp. 767–776, 2016, doi: 10.1016/j.procs.2016.05.264.
[30] K. S. Sahla and T. S. Kumar, "Classroom teaching assessment based on student emotions," in *Advances in Intelligent Systems and Computing*, vol. 530, 2016, pp. 475–486.
[31] C. Boonroungrut, T. T. Oo, and K. One, "Exploring classroom emotion with cloud-based facial recognizer in the Chinese beginning class: A preliminary study," *International Journal of Instruction*, vol. 12, no. 1, pp. 947–958, Jan. 2019, doi: 10.29333/iji.2019.12161a.
[32] U. Ayvaz, H. Gürüler, and M. O. Devrim, "Use of facial emotion recognition in e-learning systems," *Information Technologies and Learning Tools*, vol. 60, no. 4, pp. 95–104, Sep. 2017, doi: 10.33407/itlt.v60i4.1743.
[33] N. Bosch, "Detecting student engagement: human versus machine." *In Proc. of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16). Association for Computing Machinery*, USA, 2016, pp. 317–320. https://doi.org/10.1145/2930238.2930371.
[34] Peng, X., Xu, Q., Chen, Y. *et al.* An eye tracking study: positive emotional interface design facilitates learning outcomes in multimedia learning?. *Int J Educ Technol High Educ* **18**, 40 (2021). https://doi.org/10.1186/s41239-021-00274-x.
[35] M. Mukhopadhyay, S. Pal, A. Nayyar, P. K. D. Pramanik, N. Dasgupta, and P. Choudhury, "Facial emotion detection to assess learner's state of mind in an online learning system," in *ACM International Conference Proceeding Series*, Feb. 2020, pp. 107–115, doi: 10.1145/3385209.3385231.
[36] C. Thomas and D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues," in *MIE 2017 - Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education, Co-located with ICMI 2017*, Nov. 2017, vol. 2017-November, pp. 33–40, doi: 10.1145/3139513.3139514.
[37] R. Zheng, F. Jiang, and R. Shen, "Intelligent student behavior analysis system for real classrooms," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2020, vol. 2020-May, pp. 9244–9248, doi: 10.1109/ICASSP40776.2020.9053457.
[38] S. Gupta, P. Kumar, and R. Tekchandani, "A multimodal facial cues based engagement detection system in e-learning context using deep learning approach," *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 28589–28615, Jul. 2023, doi: 10.1007/s11042-023-14392-3.
[39] X. Ai, V. S. Sheng, C. Li, and Z. Cui, "Class-attention video transformer for engagement intensity prediction," *arXiv*, 2022, [Online]. Available: https://github.com/mountainai/cavt.
[40] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11365–11394, Mar. 2023, doi: 10.1007/s11042-022-13558-9.
[41] S. Shirmohammadi, "YawDD: Yawning detection dataset," *IEEE*, 2020. https://ieee-dataport.org/open-access/yawdd-yawning-detection-dataset (accessed Dec. 27, 2023).
[42] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: towards user engagement recognition in the wild," *arXiv*, 2022, [Online]. Available: http://arxiv.org/abs/1609.01885.
[43] S. Akhyani, M. Abbasi, M. Chen, and A. Lim, "Towards Inclusive HRI: Using Sim2Real to address underrepresentation in emotion expression recognition," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2022-October, pp. 9132–9139, 2022, doi: 10.1109/IROS47612.2022.9982252.
[44] L. Zahara, P. Musa, E. P. Wibowo, I. Karim, and S. B. Musa, "The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi," in *2020 5th International Conference on Informatics and Computing, ICIC 2020*, Nov. 2020, pp. 1–9, doi: 10.1109/ICIC50835.2020.9288560.
[45] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: a spontaneous audio-visual face database of affective and mental states," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, Jul. 2017, doi: 10.1109/TAFFC.2016.2553038.
[46] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Procedings of the British Machine Vision Conference 2015*, 2015, pp. 41.1-41.12, doi: 10.5244/c.29.41.
[47] Y. Wang, Y. Li, Y. Song, and X. Rong, "Facial expression recognition based on auxiliary models," *Algorithms*, vol. 12, no. 11, p. 227, Oct. 2019, doi: 10.3390/a12110227.
[48] Q. Yuan, "Research on classroom emotion recognition algorithm based on visual emotion classification," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–10, Aug. 2022, doi: 10.1155/2022/6453499.
[49] B. Gong and J. Wei, "Quantitative analysis of facial expression recognition in classroom teaching based on FACS and KNN classification algorithm," in *Proceedings of the 2022 International Conference on Educational Innovation and Multimedia Technology (EIMT 2022)*, Dordrecht: Atlantis Press International BV, 2023, pp. 663–671.

[50] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–9, doi: 10.1109/WACV.2016.7477618.
[51] D. Dukić and A. S. Krzic, "Real-time facial expression recognition using deep learning with application in the active classroom environment," *Electronics (Switzerland)*, vol. 11, no. 8, 2022, doi: 10.3390/electronics11081240.

BIOGRAPHIES OF AUTHORS

Noora C. T. assistant professor at the Department of Computer Science, Malabar College of Advanced Studies, Vengara (Calicut University) Malappuram, Kerala. She received her Master's degree in Computer Application from Calicut University in 2018. Currently, she is pursuing her Ph.D. in the Department of Computer Science at Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India. Her research interests encompass affective computing and deep learning networks. Her work focuses on the intersection of emotions and technology. She can be contacted at email: nooract@gmail.com.

Dr. P. Tamil Selvan completed his Ph.D. in Computer Science from Karpagam Academy of Higher Education in 2017. He is working as associate professor in Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore. His experience is 13.5 years. He has presented more than 25 papers in International and national Journals and Conference. His research interests are data mining and networking. He can be contacted at email: tamilselvancs@kahedu.edu.in.