Citation: Noora C T, Tamil Selvan P. Leveraging CNN to analyse facial expressions for academic engagement monitoring with insights from the multi-source academic affective engagement dataset. *Journal of Harbin Institute of Technology* (*New Series*), 2025, 32(2): 65-79. DOI: 10.11916/j.issn.1005-9113.2024026

Leveraging CNN to Analyse Facial Expressions for Academic Engagement Monitoring with Insights from the Multi-Source Academic Affective Engagement Dataset

Noora C T * *and Tamil Selvan P*

(*Department of Computer Science*, *Karpagam Academy of Higher Education*, *Coimbatore 641021*, *Tamil Nadu*, *India*)

Abstract: The dynamics of student engagement and emotional states significantly influence learning outcomes. Positive emotions resulting from successful task completion stand in contrast to negative affective states that arise from learning struggles or failures. Effective transitions to engagement occur upon problem resolution, while unresolved issues lead to frustration and subsequent boredom. This study proposes a Convolutional Neural Networks (CNN) based approach utilizing the Multi-source Academic Affective Engagement Dataset (MAAED) to categorize facial expressions into boredom, confusion, frustration, and yawning. This method provides an efficient and objective way to assess student engagement by extracting features from facial images. 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. Monitoring facial engagement with CNN using the MAAED dataset is essential for gaining insights into human behaviour and improving educational experiences.

Keywords: emotion recognition; student engagement; facial expressions; academic affective engagement; MAAED

CLC number: TP391, G442 **Document code**: A **Article ID**: 1005-9113(2025)02-0065-15

0 Introduction

The student in a classroom may experience different mental states, which are critical factors in revealing students' cognitive learning and engagement. When students make mistakes, face failures, or struggle with understanding, it may arouse negative emotions, such as irritation, frustration, and anger. On the other hand, if they can complete any task or conquer challenges or difficulties, positive mental states, such as delight, excitement, and satisfaction, will result[1]. According to D' Mello et al.[2], in typical cases, the student enters the learning activity in an engaged and concentrated state. It will continue until they reach out into any difficult situation. Gradually, this leads to confusion and boredom. At this point, either one of the transitions will occur. Positive transition will happen when the student returns to an engaged state by resolving their problems. Negative transition will happen when they face the unresolved problem during listening or discussing. Gradually, the student may stick in such a situation and transit to frustration. If this frustration persists for some time, it will lead to boredom. A good teacher should be able to monitor the changes in the student's mental states during lecturing, and give personalized assistance to the students who feel confused, frustrated, or have any other negative emotions. By identifying the problems faced during the discussion, the teacher can give further explanations or change the teaching so that the stuednts can understand easily, thereby maximising the student's learning outcome. Massive Open Online Courses (MOOCs) have revolutionised higher education by allowing interested students to pursue their education at their own pace and convenience. Content delivery is only effective when combined with real-time student feedback. That critical factor should be included in e-learning environments. Automated

Received 2024-03-01. Published online 2024-12-20.

* Corresponding author: Noora C T, Research Scholar. Email: 22drcs001@kahedu.edu.in.

engagement monitoring methods can easily be employed on 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. Support Vector Machines (SVM)[4], Convolutional Neural Networks (CNN)[5], and other deep learning algorithms[6-7] are among the most widely used models. 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, virtual, 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 enables adaptive instruction based on students' affective states, enhancing engagement, motivation, and overall learning outcomes.

Additionally, affective computing is crucial in designing emotionally responsive for 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] studied Facial Emotion Recognition (FER) to predict academic performance, highlighting the potential of affective computing in educational assessment. Systematic reviews, for example the one undertaken by Alameda-Pineda et al.[12], emphasised the benefits of integrating affective computing in learning analytics, providing valuable insights for tailoring interventions and supporting student well-being. Nie et al.[13] focused on analysing students' emotional states in online learning environments using text-mining techniques. By analysing students' written interactions, the researchers gained insights into students' emotional responses, contributing to a deeper understanding of their emotional experiences in online learning. Incorporating affective computing and emotion recognition techniques can enhance educational experiences, improve learning outcomes, and provide personalised support based on students' emotional states. By leveraging these techniques, educators can better understand 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 are both subjective and time-consuming. However, recent breakthroughs in deep learning methods, notably CNNs[14], have demonstrated significant potential for automatically analysing facial expressions and inferring engagement levels. This study proposes a CNN-based approach for facial engagement monitoring, utilising a novel dataset named MAAED. Our CNN model is trained on MAAED to classify facial expressions into engagement categories, such as boredom, confusion, frustration, and yawning. The model can accurately predict students' engagement levels by extracting discriminative features from facial images. The key contributions of this study lie in the creation of MAAED. This unique and rich dataset reflects real-world academic engagement and the development of an efficient CNN-based approach for facial engagement monitoring. The proposed approach, which uses deep learning, aims to provide educators with an objective and efficient method for assessing students' engagement during academic activities. This can improve educational environments, personalise learning experiences, and provide timely interventions to improve student outcomes.

Even though 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, stress levels, and cognitive load[18], analysing students' interactions with digital learning platforms, including clicks, mouse movements, and touch gestures, provides valuable data on their engagement and navigation patterns[19-21]. 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, including behavioural, emotional, cognitive, social, cultural, and contextual engagement. Behavioural engagement involves observable actions, such as 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 emphasises 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, emphasising inclusivity and supportiveness. Engagement levels can vary from low to medium to high, with each level indicating different levels of interest, motivation, and participation. Recognising negative affective states, such as 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 overcome challenges and thrive in their learning experiences. Strategies, such as clarification, additional support, relevant and stimulating content, and practical teaching approaches, play a vital role in addressing these emotions and enhancing student engagement.

1 Literature Review

Understanding and assessing students' engagement levels and emotional states are crucial in shaping effective teaching strategies and promoting optimal learning experiences. Each study[22-25] adopts various methodologies, ranging from traditional machine learning models to sophisticated deep learning architectures. These approaches capture nonverbal cues, such as facial expressions, body language, and eye movements, and decipher students' engagement levels and emotional responses during learning. Advanced deep learning methods in certain studies exemplify ongoing progress in education, offering educators deeper insights into human emotions and engagement.

In 2014, Whitehill et al.[26] thoroughly analysed existing computer-vision algorithms for automatic student engagement analysis and recognition. Those stuies[27-29] compared facial features of face patches from various methods, such as BoostBF, Support Vector Machine (SVM), Gabor, and the 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 et al.[30] developed a feature set defining a student's face and bodily attributes, including gaze point and body posture, using 2D and 3D data received by the Kinect One Sensor. Krithika et al.[31] developed a program to identify the students' emotions by monitoring their head, lip, and eye movements in the e-learning environment. Sahla et al.[32] developed a deep CNN technique for classroom emotion detection, where the work was performed in the classroom for the automatic analysis of the teacher from the video. A cloud-based facial emotion analysis was conducted in 2019 by Boonroungrut et al.[33] in facial emotion analysis to find students' emotions in the classroom, where the researchers examined the mood changes of 29 international students. Ayvaz et al.[34] employed several classification algorithms, such as Classification and Regression Trees (CART), Random Forest (RF), k-Nearest Neighbors (kNN), and Support Vector Machines (SVM), to analyse participants' facial expressions in an e-learning session held over Skype software using the system they built. In classrooms, they eventually concluded that emotions, such as happiness, fear, sadness, anger, surprise, and disgust, are universally acknowledged. Among them, the SVM algorithm outperforms others. Recent neurological advancements highlighted the connection between learning and emotions. Many studies emphasised the importance of students' emotions during lectures. Acknowledging this vital link between emotions and learning, integrating emotions into education becomes crucial. Understanding and supporting students' feelings leads to improved, customised learning experiences, enhancing academic performance and general well-being. Table 1 shows the summary of the existing methodologies.

Table 1 Details of existing methodologies

Study	Methodology	Categories	Accuracy	E-learning/ Classroom
Whitehill et al.[26]	SVM with Gabour features	Not engaged at all, nominally engaged, engaged in task, very engaged, unclear.	76.32%	E-learning
D' Mello et al [2]	14 different machine learning models, for example, SVM - using facial expressions	Bored, confused, delighted, engaged, frustrated	Individual class accuracy (0.61-0.87)	E-learning
Krithika et al.[31]	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
Zalatelji et al.[30]	Machine learning features and 2D, and 3D features from Kinect One Camera	High attention, medium attention, no attention	75.3%	Classroom
Sharma et al.[15]	CNN	Student's basic facial expressions	70%	E-learning
Akhyani et al.[35]	Eye tracking with a rule-based system	Happy face, neutral face	N/A	Classroom
Zheng et al.[36]	Multimodal (FER, eye tracking, body language)	Emotions	88%	E-learning
Mukhopadhyay et al.[37]	CNN	FER 2013 emotions	62%	E-learning
Bhardwaj et al.[6]	Deep learning	Angry, disgust, fear, happy, sad, surprise and neutral	93.6%	E-learning

2 Methodology

Machine learning[18] suggested predicting thyroid disease by focusing on the organ controlled by the hypothalamus. In this study, we proposed 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. Researchers preprocessed the collected datasets to ensure consistency. The designed CNN architecture consists of convolutional, pooling, and fully connected layers trained on the MAAED using suitabled optimization algorithms and loss functions for multi-class classification. Researchers evaluated the trained CNN model on a test set using standard evaluation metrics, such as accuracy, precision, recall, and F1 score, and a confusion matrix, to assess its performance.

CNNs are deep learning models extensively employed in diverse computer vision applications, including facial emotion recognition. CNNs are structured to automatically acquire and extract meaningful features from input data, significantly enhancing their effectiveness in image analysis. One can train CNN to recognize and classify facial expressions based on image patterns and characteristics. The network trains to recognize crucial facial landmarks, including eyes, nose, and mouth, and captures the distinct features associated with different emotions by understanding their spatial relationships. Fig.1 illustrates CNN's architecture. 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.

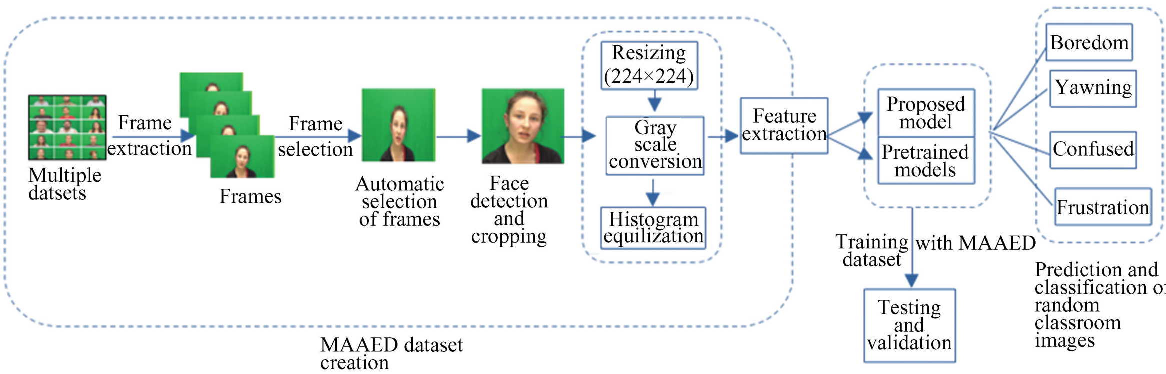


Fig.1 Architecture of CNN

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

where $\boldsymbol{I}$ represents the input data or image over which the convolution operation is performed. It can be a 1D signal, 2D image, or multi-dimensional array. $\boldsymbol{K}$ represents the convolution filter or kernel, which is a smaller matrix used to extract specific features (such as edges or textures) from the input image. t denotes the position in the output where the convolution operation is evaluated. It represents the temporal or spatial location depending on the input's dimensions. a represents the index variable used for iterating over the input data during the convolution operation. $\boldsymbol{I} * \boldsymbol{K}$ (convolution result): The output resulting from the convolution operation between the input image I and the filter K. Σ_a: The summation operator, indicating that the element-wise multiplications between $I(a)$ and $K(t - a)$ are summed to compute the convolution output at position t. The equation describes how the convolution operation is computed at each position t by sliding the filter K across the input I, performing element-wise multiplications, and aggregating the results.

This means that the output generates every pixel by adding the input pixels, each multiplied by its respective weight defined by the kernel. In two dimensions, it would be shown in Eq.(2).

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

here, (x,y) represents the coordinates of the output pixel, while (a,b) are the indices iterating over the input image pixels covered by the kernel. The filter kernel K, with its weights, shifts across the input image I both horizontally and vertically, computing the output by summing the weighted contributions of input pixels for every location. This 2D convolution operation effectively captures spatial features such as edges, textures, and patterns from the input data.

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. As a technique, Max Pooling 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 by the pooling layers retains the most salient features. At the same time, 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 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 in Eq.(3).

$$y = \boldsymbol{W}x + \boldsymbol{b} \qquad (3)$$

The weight matrix $\boldsymbol{W}$ and the bias vector $\boldsymbol{b}$ represent the parameters of the fully connected layer that we learned during training. After applying the linear transformation, we apply the activation function element-wise. The Rectified Linear Unit (ReLU) is a commonly adopted activation function.

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

The final layer in the network places the fully connected layer and uses it for multi-class classification. In that case, a softmax function is typically applied to the output of the layer to generate probabilities for each class.

$$\mathrm{softmax}(z_i) = \exp(z_i)/\Sigma\exp(z_j) \tag{5}$$

where the vector z serves as the input to the softmax function, z_i is the ith element of z, and the denominator is the sum of exp (z_j) overall j.

CNN, a specialized class of neural networks, has gained prominence due to its ability to extract meaningful features from images automatically. This characteristic proves invaluable in the realm of FER as it eliminates the need for manual feature extraction, allowing the model to discern intricate patterns and subtle nuances inherent in facial expressions. In training with a CNN model, several significant obstacles, such as overfitting, vanishing gradients, and class imbalances in the dataset, can negatively impact the model's performance. Adjusting the model's hyper-parameters and carefully applying regularization techniques become necessary to mitigate these issues. 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 challenges above effectively. Regularization techniques are a group of strategies to prevent overfitting, a scenario where the model excels with training data but struggles when faced with unfamiliar or unseen data. Some of the standard regularisation techniques used are L1 and L2 regularization, dropout, data augmentation, and early stopping. 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, leading to a more balanced and accurate classification of facial emotions. The input image underwent a series of convolutional layers, followed by pooling layers, which progressively reduce the spatial dimensions.

The last pooling layer's flattened result is fed into fully connected layers for classification using the extracted features. The final output layer represents the different emotion classes: happiness, sadness, and anger. The critical advantage of CNN in facial emotion recognition is their inherent capability to learn and extract relevant features automatically from facial images, thereby eliminating manual feature engineering requirements. This trait makes CNNs 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.

3 MAAED-Integration and Analysis

There is a significant need for extensive datasets capturing affective states related to student learning, mainly because studies focusing specifically on these states are rare. Most research concentrates on general emotions, overlooking the specific emotional states connected to how students learn. Recognizing this gap, especially in negative emotions, researchers noticed the necessity of consolidating 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 is important to note that this compilation contains 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 involves the consideration of five publicly available datasets: YawDD[38], Daisee[39], Many Faces of Confusion in the Wild dataset (MFC-Wild)[40], FER 2013[41], and BAUM-1[42]. Although the MAAED datasets includs the positive emotion concentration, this study primarily focused on negative emotions such as yawning, boredom, frustration, and confusion. The study did not actively incorporate or analyze the positive emotion of concentration within its specific context. Monitoring negative emotions in student's affective states is essential for their well-being and mental health. 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 may cotribute to academic success and well-being.

3.1 Automatic Frame Extraction and Selection

In building the MAAED, frame extraction is crucial 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, thus ensuring that each frame accurately represented a specific affective state. This manual selection process took up much time, prompting the need for an automated method to streamline frame extraction in Algorithm 1. To address this challenge, we utilized two pre-trained models for distinct purposes. The proposed model utilized the VGG face model 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
- VGG face model
- VGG16 pretrained 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:

Read frame

Resize frame to (224, 224)

Convert frame to RGB

Expand frame dimensions for model input

Predict facial features using VGG face model

Predict emotion using the custom model

If emotion_confidence > peak_emotion_confidence

Then update peak_emotion_confidence and peak_emotion_frame_index

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:

Set video capture to the specific frame index

Read and resize frame

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[43]. The VGG face model is a specialized convolutional neural network designed for face recognition. Based on the VGG architecture, it consists of 16 convolutional layers followed by fully connected layers. The model uses small convolutional filters (3×3), which have 9 parameters per filter for a single-channel input. For multi-channel inputs, such as RGB images, the filter parameters increase proportionally (e.g., 27 parameters for three channels), plus one bias term per filter. This consistent use of 3×3 filters enables a deep network that captures detailed facial features while keeping the parameter count efficient.

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 categorizes these extracted features, assigning emotion labels to the input frames. This multi-step approach facilitates a more specialized and accurate emotion recognition process than a single model alone. This approach addresses the limitations of manual frame selection, making it a valuable contribution to 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 pre-trained 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 captures crucial moments reflecting the heightened engagement category within videos. It streamlines the subsequent exploration and interpretation of these emotionally significant frames for potential deeper insights or diagnostic purposes.

Incorporating the FER 13 dataset further enlarges the dataset, bringing the total number of images to 16924 for training, 4725 for testing, and 3641 for validation. Conducting a manual analysis ensured the accuracy of the dataset by verifying 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 facial emotion recognition.

3.2 Pre-processing

Pre-processing plays a pivotal role in preparing data for analysis, particularly in task, such as facial analysis and emotion recognition. Specifically tailored for facial expression datasets extracted from videos, pre-processing 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 pre-processing 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.

3.2.1 Face detection

Choosing a suitable face detection algorithm is crucial in facial analysis and emotion recognition research. Considering the complexities of the research, we found that MTCNN (Multi-Task Cascaded Convolutional Neural Networks), 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, such as P-Net, R-Net, and O-Net, MTCNN excels, in identifying more faces within a single frame, which is crucial for our goal of accurately recognizing and analysing facial expressions. The adaptability and robustness of MTCNN are crucial factors in ensuring the success and effectiveness of facial analysis and emotion recognition tasks.

The MTCNN algorithm detects faces using cascaded convolutional neural networks. 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 Proposal Network (P-Net), the Refine Network (R-Net), and the Output Network (O-Net). The P-Net stage formulates the convolutional layer as follows:

$$O_{ij} = \sum_{(m,n)} I\,(i+m)(j+n)\,W_{mn} + b \tag{6}$$

where O_{ij}: The output value at position (i,j); I: The input feature map; W_{mn}: The weight of the convolutional kernel at position (m,n); b: The bias term added to the convolution result.

The ReLU activation function can be represented as:

$$f(x) = \max(0,x) \tag{7}$$

where $f(x)$ is the output of the ReLU activation function; x is the input value.

The max-pooling operation is given as follows:

$$M_{ij} = \max(P_{i:i+s,j:j+s}) \tag{8}$$

where M_{ij}: The result of the max-pooling operation at position (i,j); $P_{i:i+s,j:j+s}$: The region of the input being pooled, defined by a sliding window of size s.

While the softmax function for classification is

$$p_i = \frac{e^{x^i}}{\sum_j e^{x^j}} \quad (9)$$

where p_i: The probability output for class i; x_i: The input value for class i before applying the softmax function; $\sum_j e^{x^j}$: The normalization term, summing the exponentials of all input values x_j.

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.

$$L = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{box}} + \lambda_3 L_{\text{landmark}} \quad (10)$$

In this context, L represents the loss function. λ_1, λ_2, and λ_3 are the weights of the classification, bounding box, and landmark losses. L_{cls} is the classification loss, L_{box} is the bounding box loss, and L_{Landmark} is the landmark loss. These mathematical formulations collectively 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 is a vital instrument, simplifying our investigation into how people display emotions through facial features.

3.2.2 Resizing

Resizing an image means changing its size, making it smaller or larger. Resizing images is essential for consistency in computer tasks, such as facial analysis or emotion recognition. 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 224 × 224 pixels, improving model performance and efficiency and facilitating transfer learning.

3.2.3 Grayscale conversion

Converting images to grayscale holds multiple advantages in the context of facial analysis and emotion recognition. This conversion process eliminates colour information, emphasizing essential structural and textural elements in facial images. By removing colour 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 vital facial features necessary for accurate emotion recognition. Additionally, the absence of colour 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.

3.2.4 Train-test-validation splitting

Segmenting the dataset into training, testing, and validation sets (60%, 20%, and 20%) supports model training, evaluation, and fine-tuning tasks. It prevents overfitting and allows performance monitoring, hyperparameter tuning, and model selection. These pre-processing steps enhance the dataset, enable facial analysis, and support robust emotion recognition and engagement monitoring. Table 2 presents a comprehensive overview of the MAAED after completing the necessary pre-processing steps.

Table 2 Dataset overview after pre-processing

Engagement category	Train*	Test#	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

Note: The symbol "*" represents that 60% of the total datasets are allocated for training; The symbol "#" represents that 20% of the total datasets are allocated for testing; The symbol "+" represents that 20% of the total datasets are allocated for the fine-tuning task.

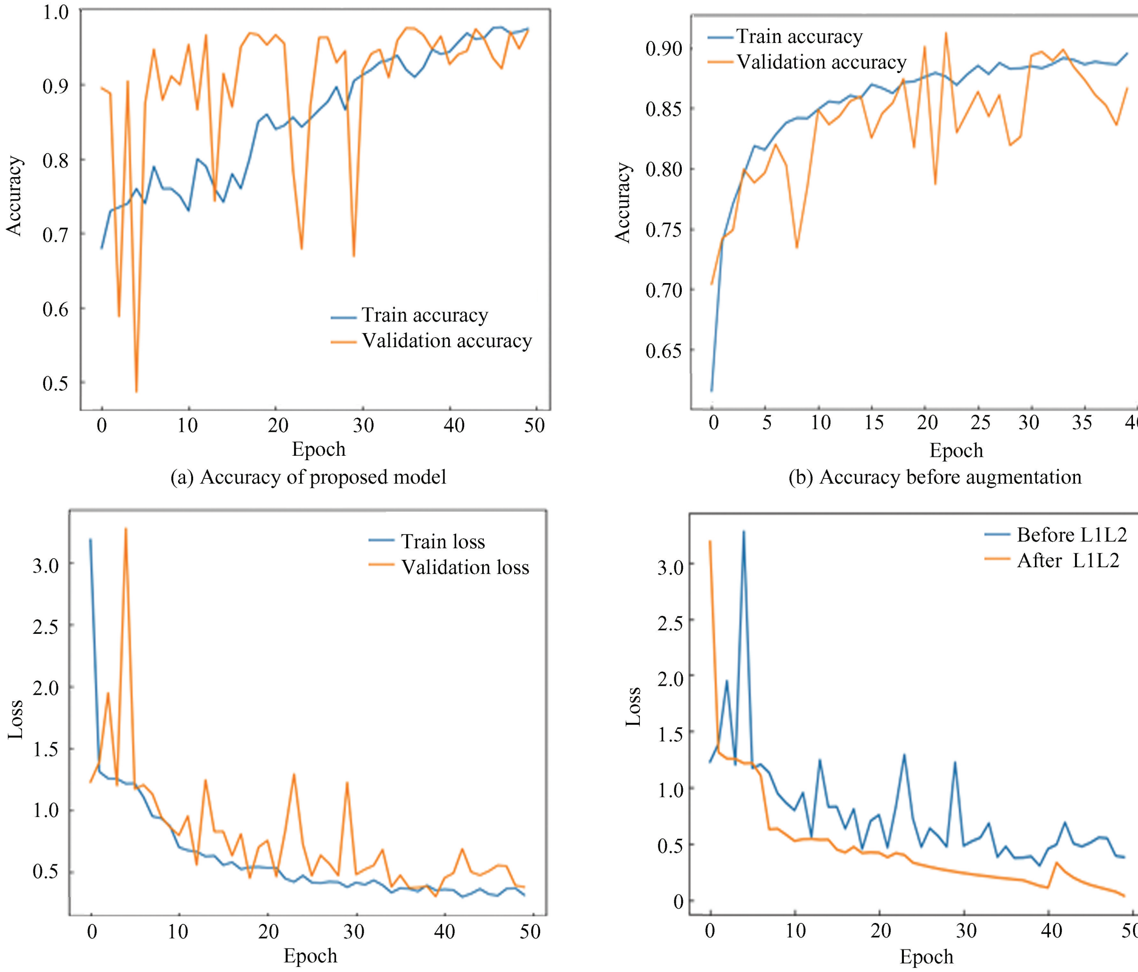
4 Results Analysis

The proposed CNN model demonstrates high recall and precision scores across all emotion classes, indicating its ability to accurately detect emotions in facial expressions, even in subtle or ambiguous expressions. The proposed model exhibits overall performance on the MAAED dataset, with precision, recall, and F1-score values close to 0.96. It shows slightly better accuracy in classifying boredom, 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 imbalanced. 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 high 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 balance these metrics, as shown in Table 3. 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. The model's efficiency is also noteworthy, as it swiftly classifies images within a time frame of just 10 ms. 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 can 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 Fig. 2, 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. The confusion matrix demonstrates how effective the CNN system is in facial engagement monitoring. Each cell in the matrix shows the percentage of individuals in different categories. For instance, the true positive cell's value of 0.93 indicates that the model correctly identified 93% of confused individuals. On the other hand, the value of 0.01 in the false positive cell indicates that the model correctly classifies 1% of those needful clarification is confused. Nonetheless, there is potential for enhancement, particularly in increasing the false negative rate, suggesting the need for the model to detect and include more individuals experiencing confusion. Overall. In contrast, the model performs satisfactorily. There is scope for improvement, particularly in minimizing the false negative rate to identify better those experiencing confusion.

Table 3 Comparative analysis of precision, recall and F1 scores for different engagement categories

Model	Metric	Confused	Boredom	Yawning	Frustration
Proposed CNN	Precision	0.97	0.95	0.96	0.96
	Recall	0.93	0.95	0.97	0.96
	F1 Score	0.96	0.96	0.97	0.96
VGG16	Precision	0.92	0.93	0.93	0.93
	Recall	0.92	0.92	0.93	0.93
	F1 Score	0.92	0.92	0.91	0.91
Inception v	Precision	0.94	0.94	0.93	0.91
	Recall	0.92	0.94	0.91	0.91
	F1 Score	0.95	0.94	0.91	0.91
ResNet 50	Precision	0.83	0.84	0.84	0.85
	Recall	0.85	0.85	0.85	0.85
	F1 Score	0.85	0.85	0.83	0.85
ResNet 101	Precision	0.86	0.82	0.82	0.82
	Recall	0.86	0.83	0.83	0.83
	F1 Score	0.86	0.84	0.82	0.84



(a) Accuracy of proposed model
(b) Accuracy before augmentation
(c) Train and validation loss of proposed model
(d) Loss curve before and after L1L2

Fig.2 Accuracy and loss graph of proposed model

Figs.2(c) and (d) demonstrate the loss curves observed during the training and validation sessions of the proposed model. The graph depicts how the loss values evolve throughout training across 50 epochs. The convergence of the loss curves for 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 the model's ability to accurately predict both the training data and unseen or validation data. This loss metric assesses the disparity between the actual distribution Y and the predicted distribution $\hat{Y}$. Mathematically, define it as in Eq.(11).

$$L(Y,\hat{Y}) = -\sum_i^n Y_i \log(\hat{Y}_i) \tag{11}$$

here, Y_i represents the actual label for class i, often encoded as one-hot, while $\hat{Y}_i$ signifies the predicted probability associated with class i. 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[44]. The logarithmic function imposes a substantial penalty when the model makes highly confident yet incorrect predictions—the impact of L1L2 regularization on model performance over several epochs after applying augmentation techniques. There are two distinct curves of loss before L1L2 and after L1L2, as illustrated in Fig.2(d).

L1 and L2 refer to two types of regularization techniques used to prevent overfitting and improve the generalization of machine learning models, particularly in neural networks.

1) L1 regularization (Lasso regularization)

Mechanism: L1 regularization adds a penalty equal to the absolute value of the magnitude of coefficients to the loss function. Mathematically, it can be represented Eq.(12):

$$\text{Loss} = \text{OriginalLoss} + \lambda \sum |w_i| \quad (12)$$

where w_i are the weights of the model, and λ is a regularization parameter that controls the strength of the penalty.

Effect: L1 regularization encourages sparsity in the model by driving some of the weights to zero. This is particularly useful for feature selection, as it effectively eliminates less important features by assigning their coefficients a value of zero.

Applications: L1 regularization is used when feature selection or creating simpler models is desirable, as it can simplify models without significantly sacrificing performance.

2) L2 regularization (ridge regularization)

Mechanism: L2 regularization adds a penalty equal to the square of the magnitude of coefficients to the loss function. Mathematically, it is expressed as:

$$\text{Loss} = \text{OriginalLoss} + \lambda \sum w_i^2 \quad (13)$$

where w_i are the weights of the model, and λ is the regularization parameter.

Effect: L2 regularization discourages large weight values, making the model more stable and reducing the risk of overfitting. Unlike L1 regularization, L2 does not force the weights to be zero but rather keeps them small, leading to a smoother model.

Applications: L2 regularization is widely used when all input features are potentially relevant and the goal is to ensure model stability and avoid overfitting.

3) L1L2 regularization (elastic net)

Combination: L1L2 regularization combines both L1 and L2 regularization techniques. The combined penalty is:

$$\text{Loss} = \text{OriginalLoss} + \lambda_1 \sum |w_i| + \lambda_2 \sum w_i^2 \quad (14)$$

here, λ_1 and λ_2 are the regularization parameters that control the contribution of the L1 and L2 penalties, respectively.

Advantages: The combination of L1 and L2 regularization benefits from both techniques, allowing for feature selection (due to L1) and maintaining overall stability (due to L2). It is particularly useful for high-dimensional data where some features may be more relevant than others.

5 Contributions

The present work introduces several novelties in the recognition field and its application in academic environments. Firstly, a new dataset was explicitly curated to predict the engagement level of students in educational settings. This dataset holds excellent value as it facilitates training and evaluating emotion recognition models customized to the unique demands of academic contexts. Secondly, the researchers adopted a two-step approach for extracting frames from video datasets for emotion recognition. This innovative strategy utilizes two models, enhancing accuracy and automating the frame extraction process. Compared to relying on just one model, the outcome is a more effective and exact method. Beyond emotion recognition, the study demonstrates the potential applications in educational domains. It highlights how these technological advancements can 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. The proposed model's performance comparison with baseline models on the MAAED dataset consistently demonstrates positive outcomes across various pre-trained models and our proposed model, as shown in Fig.3.

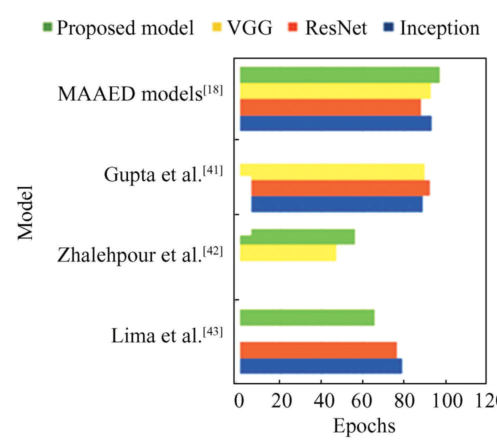


Fig.3 Analysing the accuracy of the proposed model

The MAAED dataset proves highly effective in identifying student affective states, which is evident from the overall solid results. Notably, the ResNet 50 pre-trained model exhibits exceptional proficiency when applied to Gupta et al.[41], Daisee dataset, surpassing its performance on MAAED. Conversely, VGG and Inception V3 are slightly advantageous, particularly within the MAAED dataset. Meanwhile, Zhalehpour et al.'s[42] research showcases AI's practical implementation with automated detection capabilities, successfully identifying behaviors, such as hand-raising, standing, and sleeping among students. Thus, AI proves valuable in understanding various classroom dynamics. It is worth noting that Lima et al.[43] focused on six basic emotion categories rather than student learning-based emotions, conducting experiments on CK+, FER-2013, and SFEW datasets with pre-trained 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.

6 Conclusions

This study proposes a CNN-based approach using the MAAED for facial engagement monitoring. The CNN model accurately classifies engagement categories, such as 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. It could 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. The proposed method has potential applicability in educational technology and shows potential to enhance students' learning experiences.

References

[1] Craig S, Graesser A, Sullins J, et al. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. Journal of Educational Media, 2004, 29(3): 241-250. DOI: 10.1080/1358165042000283101.

[2] D'Mello S, Graesser A. Dynamics of affective states during complex learning. Learning and Instruction, 2012, 22(2): 145-157. DOI: 10.1016/j.learninstruc.2011.10.001.

[3] Wang R, Cao J, Xu Y, et al. Learning engagement in massive open online courses: A systematic review. Frontiers in Education, 2022, 7: 1074435. https://www.frontiersin.org/articles/10.3389/feduc.2022.1074435

[4] Sümer Ö, Goldberg P, D'Mello S, et al. Multimodal engagement analysis from facial videos in the classroom. IEEE Transactions on Affective Computing, 2023, 14(2): 1012-1027. DOI: 10.1109/TAFFC.2021.3127692.

[5] Khan S S, Abedi A, Colella T. Inconsistencies in the Definition and Annotation of Student Engagement in Virtual Learning Datasets: A Critical Review. 2023, arXiv:2208.04548. DOI:10.48550/arXiv.2208.04548.

[6] Bhardwaj P, Gupta P K, Panwar H, et al. Application of deep learning on student engagement in e-learning environments. Computers and Electrical Engineering, 2021, 93: 107277. DOI: 10.1016/j.compeleceng.2021.107277.

[7] Perumal B, Nagaraj P, Thulasi Sai Narsimha Charan, et al. Student engagement detection in classroom using deep CNN-based learning approach. 2023 8th International Conference on Communication and Electronics Systems (ICCES). Piscataway: IEEE, 2023:1233-1238. DOI: 10.1109/ICCES57224.2023.10192809.

[8] Bustos-López M, Cruz-Ramírez N, Guerra-Hernández A, et al. Wearables for engagement detection in learning environments: A review. Biosensors, 2022, 12: 509. DOI:10.3390/bios12070509.

[9] Apicella A, Arpaia P, Frosolone M, et al. EEG-based measurement system for monitoring student engagement in learning 4.0. Scientific Reports, 2022, 12: 5857. DOI: 10.1038/s41598-022-09578-y.

[10] Green G R. Text Based Discussions: An Approach to Teach Reading Comprehension. Culminating Experience Projects. 166. Allendale: Grand Valley State University, 2022. https://scholarworks.gvsu.edu/gradprojects/166

[11] Saneiro M, Santos O C, Salmeron-Majadas S, et al. Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. The Scientific World Journal, 2014, 2014:484873. DOI:10.1155/2014/484873.

[12] Alameda-Pineda X, Staiano J, Subramanian R, et al. SALSA: A Novel Dataset for Multimodal Group Behavior Analysis, 2015, arXiv: 1506.06882. DOI: 10.48550/arXiv.1506.06882.

[13] Nie Y, Luo H, Sun D. Design and validation of a diagnostic MOOC evaluation method combining AHP and text mining algorithms. Interactive Learning Environments, 2021, 29(2): 315-328. DOI: 10.1080/10494820.2020.1802298.

[14] Hu Y, Jiang Z, Zhu K. An optimized CNN model for engagement recognition in an E-learning environment. Applied Sciences, 2022, 12(16): 8007. DOI: 10.3390/app12168007.

[15] Sharma P, Joshi S, Gautam S, et al. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. Technology and Innovation in Learning, Teaching and Education, TECH-EDU 2022. Berlin: Springer, 2022, 1720:.52-68. DOI: 10.1007/978-3-031-22918-3_5.

[16] Conner J, Posner M, Nsowaa B. The relationship

between student voice and student engagement in urban high schools. The Urban Review, 2022,54(5):755-774. DOI:10.1007/s11256-022-00637-2.

[17] Slater S, Ocumpaugh J, Baker R, et al. Using natural language processing tools to develop complex models of student engagement. 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). Piscataway: IEEE, 2017: 542-547. DOI: 10.1109/ACII.2017.8273652.

[18] Gu Z, Zarubin V C, Mickley Steinmetz R, et al. Heart rate variability in healthy subjects during monitored, short-term stress followed by 24-hour cardiac monitoring. Frontiers in Physiology, 2022,13:897284. DOI:10.3389/fphys.2022.897284.

[19] Fragueiro A, Debroize R P, Coutrot A, et al. Pilot study: Eye-tracking and skin conductance to monitor task engagement during bimodal neurofeedback. Hal Open Science, 2023, inserm-04107747. https://inserm.hal.science/inserm-04107747v1/document.

[20] Nabil R H, Rupai A A A, Barid M, et al. An intelligent examination monitoring tool for online student evaluation. Malaysian Journal of Science and Advanced Technology, 2022:122-130. DOI:10.56532/mjsat.v2i3.62.

[21] Preuveneers D, Garofalo G, Joosen W. Cloud and edge based data analytics for privacy-preserving multi-modal engagement monitoring in the classroom. Information Systems Frontiers, 2021, 23(1): 151-164. DOI: 10.1007/s10796-020-09993-4.

[22] Ashwin T S, Guddeti R M R. Unobtrusive behavioral analysis of students in classroom environment using non-verbal cues. IEEE Access, 2019,7:150693-150709. DOI: 10.1109/ACCESS.2019.2947519.

[23] Henrie C R, Halverson L R, Graham C R. Measuring student engagement in technology-mediated learning: A review. Computers & Education, 2015,90:36-53. DOI: 10.1016/j.compedu.2015.09.005.

[24] Pabba C, Kumar P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert Systems, 2022, 39(1): e12839. DOI: 10.1111/exsy.12839.

[25] Ekman P. Universals and cultural differences in facial expressions of emotion. Nebraska Symposium on Motivation, 1971,19:207-283.

[26] Whitehill J, Serpell Z, Lin Y C, et al. The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing, 2014, 5(1): 86-98. DOI: 10.1109/TAFFC.2014.2316163.

[27] Viola P, Jones M J. Robust real-time face detection. International Journal of Computer Vision, 2004,57:137-154. DOI: 10.1023/B:VISI.0000013087.49260.fb.

[28] Khan A B F, Kamalakannan K, Ahmed N S S. Integrating machine learning and stochastic pattern analysis for the forecasting of time-series data. SN Computer Science, 2023, 4: Article number 484. DOI: 10.1007/s42979-023-01981-0.

[29] Littlewort G, Whitehill J, Wu T, et al. The computer expression recognition toolbox (CERT). 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG). Piscataway: IEEE, 2011: 298-305. DOI: 10.1109/FG.2011.5771414.

[30] Zaletelj J, Kosir A. Predicting students' attention in the classroom from Kinect facial and body features. EURASIP Journal on Image and Video Processing, 2017, 2017: Article number 80. DOI: 10.1186/s13640-017-0228-8.

[31] Krithika L B, Lakshmi Priya G G. Student emotion recognition system (SERS) for e-learning Improvement based on learner concentration metric. Procedia Computer Science, 2016, 85: 767-776. DOI: 10.1016/j.procs.2016.05.264.

[32] Sahla K S, Kumar T S. Classroom teaching assessment based on student emotions. Intelligent Systems Technologies and Applications 2016. Berlin: Springer, 2016,530:475-486. DOI: 10.1007/978-3-319-47952-1_37.

[33] Boonroungrut C, Oo T, One K. Exploring classroom emotion with cloud-based facial recognizer in the chinese beginning class: A preliminary study. International Journal of Instruction, 2019,12(1):947-058. DOI: 10.29333/iji.2019.12161a.

[34] Ayvaz U, Gürüler H, Devrim M O. Use of facial emtion recognition in e-learning systems. Information Technologies and Learning Tools, 2017, 60(4): 2076-8184. DOI: 10.33407/itlt.v60i4.1743.

[35] Akhyani S, Boroujeni M A, Chen M, et al. Towards inclusive HRI: Using Sim2Real to address underrepresentation in emotion expression recognition. 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2022. DOI: 0.1109/IROS47612.2022.9982252.

[36] Zheng R, Jiang F, Shen R. Intelligent student behavior analysis system for real classrooms. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2020: 9244-9248. DOI: 10.1109/ICASSP40776.2020.9053457.

[37] Mukhopadhyay M, Pal S, Nayyar A, et al. Facial emotion detection to assess learner's state of mind in an online learning system. Proceedings of the 2020 5th International Conference on Intelligent Information Technology. New York: ACM Press, 2020: 107-115. DOI: 10.1145/3385209.3385231.

[38] Abtahi S, Omidyeganeh M, Shirmohammadi S, et al. YawDD: A yawning detection dataset. Proceedings of ACM Multimedia Systems. New York: ACM Press, 2014:24-28. DOI: 10.1145/2557642.2563678.

[39] Gupta A, D'Cunha A, Awasthi K, et al. DAiSEE: Towards User Engagement Recognition in the Wild.

2022, arXiv: 1609.01885. DOI: 10.48550/arXiv.1609.01885.

[40] Zahara L, Musa P, Prasetyo Wibowo E, et al. The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi. 2020 Fifth International Conference on Informatics and Computing (ICIC). Piscataway: IEEE, 2020: 1-9. DOI: 10.1109/ICIC50835.2020.9288560.

[41] Gupta S, Kumar P, Tekchandani R. A multimodal facial cues based engagement detection system in e-learning context using deep learning approach. Multimedia Tools and Applications, 2023, 82(18): 28589-28615. DOI: 10.1007/s11042-023-14392-3.

[42] Zhalehpour S, Onder O, Akhtar Z, et al. BAUM-1: A spontaneous audio-visual face database of affective and mental states. IEEE Transactions on Affective Computing, 2016, 8(3): 300-313. DOI: 10.1109/TAFFC.2016.2553038

[43] Lima M, Ahmed K R, Jahan N, et al. Deep learning based approach for detecting student engagement through facial emotions. 2024 International Conference on Data Science and Network Security (ICDSNS). Piscataway: IEEE, 2024: 1-6. DOI: 10.1109/ICDSNS62112.2024.10691098.

[44] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition. Proceedings of the British Machine Vision Conference 2015, Glasgow: BMVA Press, 2015: 41.1-41.12. DOI: 10.5244/C.29.41.