*Original Research Article*

EngageNet: A Model for Evaluating Student Engagement through Facial Expression and Behavior Analysis

.

ABSTRACT

|  |
| --- |
| Online learning has emerged as a prominent trend in modern education, driven by its flexibility, accessibility, and capacity to support personalized learning experiences. However, despite these advantages, one of the most pressing challenges it faces lies in maintaining and accurately evaluating the quality of teaching and learning. A particularly critical aspect is the assessment of learner engagement in virtual environments. In fact, traditional approaches to assessing student engagement, which depend on synchronous, face-to-face interaction, frequently prove inadequate in virtual learning environments where such real-time communication between educators and learners is restricted. Therefore, this study introduces a deep learning-based model that combines facial emotion recognition, gaze direction tracking, and eye openness analysis. By integrating these emotional and behavioral characteristics, the model offers a comprehensive and objective approach to assessing learner’s attention throughout online instruction. To support the development and validation of this model, a specialized dataset was proposed, capturing a diverse range of engagement scenarios. Experimental evaluations demonstrate that the proposed method achieves a notable accuracy of 79.76%, underscoring its effectiveness and robustness in capturing learner engagement dynamics. These findings suggest that the model holds strong potential for enhancing the monitoring and personalization of online learning experiences, thereby contributing to improved educational outcomes in virtual classrooms. |

*Keywords: engagement, facial expression, behavior, online learning, deep learning*

1. INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) has significantly accelerated modernization and digitalization, particularly in the field of education. The increasing integration of technology into education is largely driven by AI’s potential to enhance and innovate teaching and learning methodologies. The COVID-19 pandemic, an unforeseen global event, served as a catalyst that expedited the transition from traditional classroom-based learning to online learning environments. This shift has revealed both the potential and inherent challenges of online education. One of the most critical challenges is the substantial decline in learner-instructor interaction due to geographical separation, which reduces engagement and emotional connection, ultimately negatively impacting learning outcomes [1]. Moreover, the accessibility of recorded lectures may contribute to student procrastination, as students might defer watching lectures instead of actively participating in real-time online sessions. This behavior can degrade knowledge retention due to the absence of immediate interactive feedback, which is essential for resolving uncertainties in real time.

The quality of online learning heavily depends on learners' self-discipline. The ability to manage time effectively, maintain motivation, and stay focused on learning tasks becomes more crucial than ever in the absence of direct supervision from instructors, as seen in traditional classrooms. Therefore, capturing and evaluating learners’ behaviors and engagement is essential for enhancing the effectiveness of online learning sessions [2], [3].

To objectively and efficiently assess learner engagement in online learning environments, the application of deep learning models in image and video analysis has emerged as a promising solution. Deep neural network architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformer models have demonstrated superior capabilities in automatically detecting and evaluating learner engagement [4], [5], [6]. Specifically, CNNs can extract fine-grained spatial features from facial expressions, including eye shape, nose, and mouth, while RNNs and Transformer models are particularly effective in capturing temporal variations in facial behavior and expressions, such as sequential emotional changes or gaze shifts.

Leveraging these advantages, this study proposes a comprehensive approach to assessing learner engagement in online education by multidimensionally analyzing facial expressions and behavioral cues. The proposed method utilizes camera-based image data to analyze emotional states, gaze direction, and eye openness. These factors are then integrated to provide a more precise assessment of learners' focus and engagement.

The key contributions of this study include:

* Proposing a model that integrates facial emotion recognition, gaze direction, and eye openness to evaluate learner engagement.
* Constructing an emotion dataset consisting of labeled images and videos representing five emotional states: angry, happy, neutral, sad, and surprise.
* Developing a learner engagement assessment dataset for online learning, categorized into four engagement levels: High Engagement, Normal Engagement, Low Engagement, and Disengagement.
* Conducting experiments and evaluating the proposed model on the constructed dataset to validate its effectiveness.

2. related work

In recent years, assessing learner engagement in online environments has gained significant attention. However, most existing studies still focus primarily on traditional classrooms [7]. Some research efforts employ machine learning models to estimate student attention levels and emotions based on facial expressions [8], [9]. Alruwais et al [10], the authors experimented with a range of machine learning models, including XGBoost, LightGBM, Random Forest, CatBoost, and Multilayer Perceptron, for classifying learner engagement levels in online environments. These models were trained and evaluated on the Open University Learning Analytics Dataset (OULAD). The results demonstrated high accuracy rates, with XGBoost achieving 92.10%, LightGBM and CatBoost reaching 92.23%, Random Forest attaining 92.18%, and Multilayer Perceptron achieving 91.75%. Similarly, Ayouni's study [11] applied machine learning models such as Decision Tree, Support Vector Machine, and Artificial Neural Network to predict learner engagement in online learning, classifying engagement into three states: Not Engaged, Passively Engaged, and Actively Engaged. These models were trained on real-world data from Learning Management Systems (LMS), enabling precise engagement assessment. However, relying solely on LMS data may overlook critical factors influencing engagement, such as course design and teaching style.

The rapid advancement of deep learning has introduced promising approaches for evaluating learner engagement in online education. Deep learning models are increasingly utilized to analyze student attention and emotions in virtual classrooms [12], [13]. Santoni [14] proposed a deep learning-based model incorporating Ensemble Bagging with 1D CNN and 1D ResNet networks to automate online learner engagement assessment. By applying bootstrap aggregating techniques to one-dimensional deep learning networks, the model achieved an accuracy of 93.75% on the DAiSEE dataset. In another study, Gupta [15] introduced a facial expression analysis method that classifies emotions and computes an Engagement Index (EI) to predict two states: "Engaged" and "Not Engaged." Bhardwaj et al. [16] employed CNN-based deep learning models and conducted experiments using the FER-2013 and MES datasets. The study utilized all-CNN, network-in-network CNN, and very deep CNN architectures to classify learner engagement. The accuracy levels for different engagement states were 91.74% for "Not Engaged," 89.55% for "Moderate Engagement," and 95.69% for "High Engagement".

Overall, these studies highlight the significant potential of machine learning and deep learning models in online learner engagement assessment, demonstrating impressive accuracy and performance. However, ensuring reliable engagement evaluation requires careful attention to input data processing. Future research should focus on integrating diverse data sources and refining processing algorithms to provide a more comprehensive and effective solution for online learning environments.

3. dataset

Currently, datasets for evaluating student engagement in online learning environments still have several limitations, such as a lack of diversity in facial angles and restricted public availability. Therefore, we have decided to build a dataset specifically designed to assess learner engagement in online learning. To enhance the dataset's practicality and applicability, in addition to behavioral features, we have incorporated emotional states, including *angry*, *happy*, *neutral*, *sad*, and *surprise*. This addition allows the dataset to more comprehensively reflect learners' psychological states during online participation, thereby increasing its real-world applicability.

To build the dataset for this study, we conducted an *IELTS Reading* online class on the Google Meet platform, inviting students aged 18 to 22 to participate. The session was recorded using OBS Studio at HD resolution with a frame rate of 60 FPS. As a result, we collected an engagement dataset comprising 1,294 video clips, each with a duration of 10 seconds. Specifically, the dataset contains 413 videos labeled as *Disengagement*, 91 videos labeled as *Low Engagement*, 763 videos labeled as *Normal Engagement*, and 27 videos labeled as *High Engagement*.

For the emotion recognition training dataset, we extracted facial expressions from the recorded videos, ensuring that each selected frame contained one of the five emotional states (*angry*, *happy*, *neutral*, *sad* *and* *surprise*). The facial images were cropped and resized to 224×224 pixels in RGB format. To ensure fairness in model training and evaluation, individuals appearing in the engagement assessment dataset were excluded from the emotion recognition dataset. The details of the engagement dataset and the emotion dataset are presented in Table 1 and Table 2, respectively.

**Table 1.** **The details of 10-second video samples for each engagement level**

|  |  |
| --- | --- |
| **Engagement level** | **Total** |
| Disengagement  Low Engagement | 413  91 |
| Normal Engagement | 763 |
| High Engagement | 27 |

**Table 2.** **Number of images for each emotion in the Train and Test sets of the facial emotion dataset.**

|  |  |
| --- | --- |
| **Emotion** | **Quantity** |
| Angry  Happy  Neutral  Sad | 500  764  711  182 |
| Surprise | 455 |
| Total | 2612 |

4. PROPOSED METHOD

This study proposes a method for evaluating student engagement in online learning based on facial expression and behavioral analysis. An overview of the model architecture is illustrated in Figure 1. Initially, the model captures images from the student's camera. A facial emotion recognition model is then employed to determine the student’s emotional state. For behavioral analysis, the input remains the camera images, which are processed using a model to detect gaze direction and eye openness. Finally, both emotional and behavioral features are combined to calculate the student's engagement level during the online learning session.

A person standing next to a diagram

AI-generated content may be incorrect.

**Fig. 1. EngageNets’ architecture**

**4.1. Emotion Recognition Model**

In this study, we propose the LaCER model for facial emotion recognition based on landmark features. Figure 3 illustrates the overall architecture of LaCER. First, input images with a resolution of 224×224 are processed and facial landmarks are extracted using the Face Mesh model from MediaPipe. The extracted and normalized landmarks are then sequentially passed through the layers of the LaCER model, with the final output being the predicted probabilities for each emotion category. Additionally, we conduct experiments on three pixel-based emotion recognition models, including ResNet50 [17], VGG19 [18], and MobileNetV1 [19], utilizing pre-trained weights on the ImageNet dataset.

4.1.1 Landmarks Feature Extraction

A person taking a selfie

AI-generated content may be incorrect.

**Fig. 2.** **202 Facial Landmark Points**

Landmarks provide a robust geometric representation that is resistant to environmental factors such as background and lighting conditions. By focusing on key facial structures, the model can more accurately recognize emotion-related variations. In this study, we utilize the Face Mesh module from MediaPipe to detect faces and extract 468 landmark points from input images. However, only 202 critical landmarks are retained, corresponding to five key facial regions: eyes, eyebrows, mouth, nose, and chin, as these areas convey the most distinctive emotional information. Figure 2 shows 202 landmark points on the face. The selection of these landmarks optimizes feature extraction, ensuring the model focuses on the most relevant facial expression variations.

Before being fed into the model, the data undergoes a two-step normalization process to enhance stability and accuracy in facial emotion recognition. In the first step, the absolute coordinates of the landmark points are converted into relative coordinates to minimize the impact of viewpoint differences in the input images. Specifically, for each landmark with initial coordinates, we select a reference point , which is the first point in the feature list. The first point, namely the tip of the nose, is the point with index 1, the initial point in the list. Subsequently, all remaining points are adjusted to obtain new coordinates as defined in equations (1) and (2):





The next step involves normalizing the data to the range [−1,1]. Since the coordinates still contain large values, normalization is performed by identifying the maximum coordinate valuesfrom the processed set of coordinates obtained in the previous step. The final normalized coordinatesare then computed using the formulas (3) and (4):



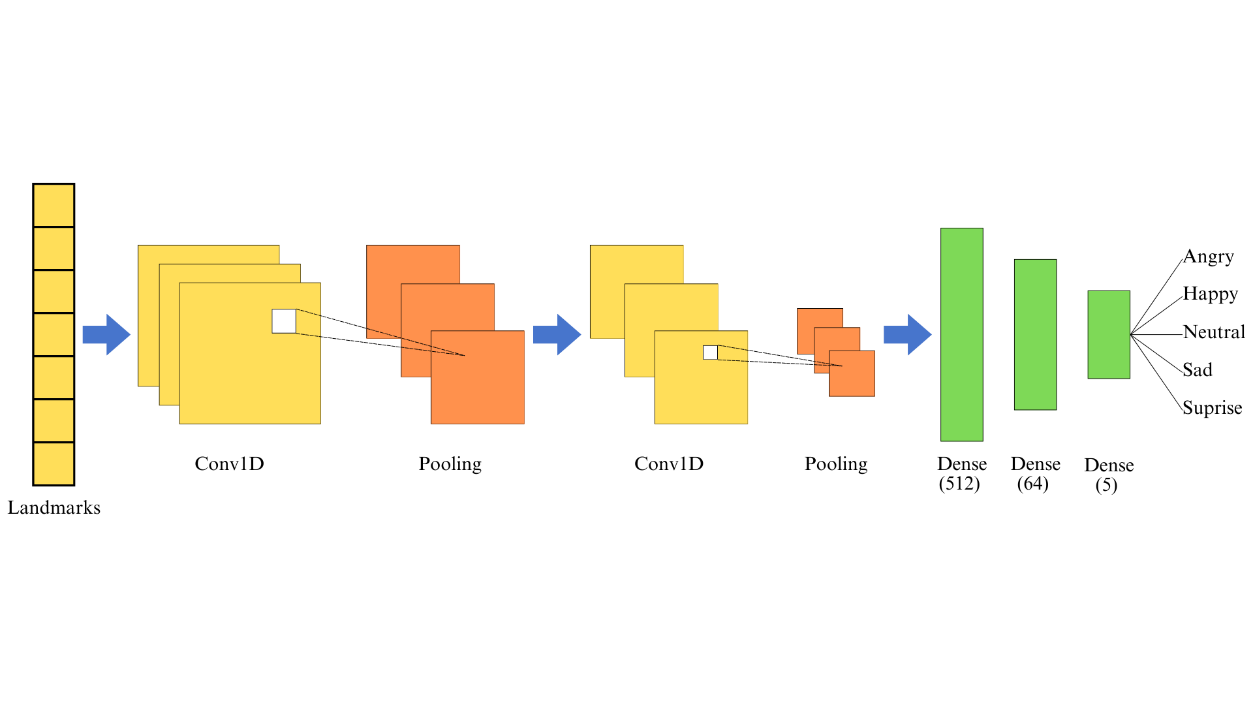


After completing the data normalization process, we transform the two-dimensional data, consisting of 202 landmark points into a 404-dimensional feature vector. This transformed vector is then fed into the training model for further processing.

**4.1.2 Landmark-based CNN for emotion recognition (LaCER)**

To recognize emotions from facial landmark features, we proposed a CNN-based model for facial emotion recognition, where preprocessed landmark coordinates are reshaped to (404, 1) as input. The overall architecture of the proposed LaCER model is illustrated in Figure 3.

The model begins with a Conv1D layer containing 64 filters, a 3×3 kernel, ReLU activation, and "same" padding to preserve the output dimensions. This is followed by a MaxPooling1D layer with a pool size of 2, which reduces the data dimensionality while extracting key features. A second Conv1D layer with 128 filters, a 3×3 kernel, ReLU activation, and "same" padding is added, followed by another MaxPooling1D layer to further enhance feature generalization. The output is then flattened using a Flatten layer and passed through two Dense layers with 128 and 64 neurons, respectively. Each Dense layer employs ReLU activation and is regularized with a Dropout rate of 0.5 to mitigate overfitting, ensuring the model does not become overly reliant on training data. Finally, the output layer consists of a Dense layer with five emotion classes (Angry, Happy, Neutral, Sad, Surprise), utilizing the softmax activation function to predict the probability distribution across the emotion categories.



**Fig. 3. LaCERs’ architecture**

**4.1.3 Facial Expression Recognition based on Pixel feature**

Besides, we also conduct experiments with three pixel-based models ResNet50, VGG19, and MobileNetV1 on the proposed dataset. All three models are pre-trained on the ImageNet dataset and take input images of size (224, 224, 3).

ResNet50 employs a bottleneck design, where each convolutional block consists of three layers, reducing computational costs while maintaining deep learning capability. To reduce the dimensionality before feeding data into fully connected layers, we apply a GlobalAveragePooling2D layer, followed by two Dense layers with 2048 and 512 neurons, using the ReLU activation function.

VGG19 consists of 19 convolutional layers. The output of VGG19 is passed through a GlobalAveragePooling2D layer to reduce dimensionality, followed by a Dense layer with an output size of 1024 neurons and a ReLU activation function to capture higher-level features.

MobileNetV1 is a lightweight CNN model family optimized for mobile devices. The 4D tensor output is transformed into a 1D vector using a GlobalAveragePooling2D layer, reducing the number of parameters and mitigating overfitting. This is followed by two Dense layers with 2048 and 1024 neurons, both utilizing the ReLU activation function for feature extraction.

After extracting features from each model, we obtain corresponding feature vectors. These vectors are then passed into a Dense output layer with 5 neurons, utilizing the softmax activation function to classify emotions into 5 classes.

**4.2 Behavior Prediction Model**

**4.2.1 Calculation of Eye Aspect Ratio (EAR)**

Eye openness is quantified based on the frequency of eye opening, blinking, and closing in each video frame. The assessment of alertness is conducted using the Eye Aspect Ratio (EAR), with key factors including blink frequency and prolonged eye closure duration. After extracting facial landmark points, the system focuses on the key landmark points of the left and right eyes in the real-time video stream. Each eye is modeled using six coordinate points , as illustrated in Figure 4, resulting in a total of 12 landmark points for both eyes. Based on this set of points, the system computes the EAR index to detect blinking behavior and accurately evaluate the learner's alertness.

A red line on a white surface

AI-generated content may be incorrect.

**Fig. 4. Illustration of six eye landmarks**

Each eye is represented by six landmark points, where each point is defined by coordinates . The ratio of these points is then computed, referred to as the Eye Aspect Ratio (EAR) [20], using the following formula:





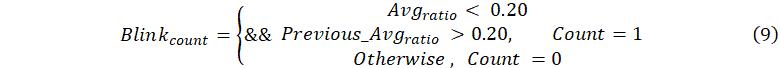
Where and represent the eye aspect ratio of the left and right eye, respectively. denotes the -th specific landmark on the eye, where each landmark is defined by coordinates . After computing the EAR values for both eyes, the eye state is determined based on the following thresholds:

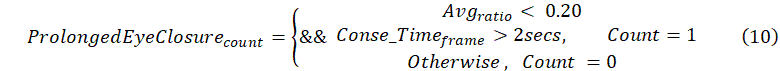


The value of EAR is referenced as . If  is greater than 0.20, it indicates that the eyes are open, whereas it approaches zero when the eyes are closed. To detect blinks and prolonged eye closures, this study utilizes the method proposed in [21] using EAR values for calculation. First, the average eye aspect ratio  is calculated based on the EAR of the left eye ( ) and the EAR of the right eye ( ) using the following formula:



Then, the number of blinks () and the number of prolonged eye closures () are computed using the following formulas:



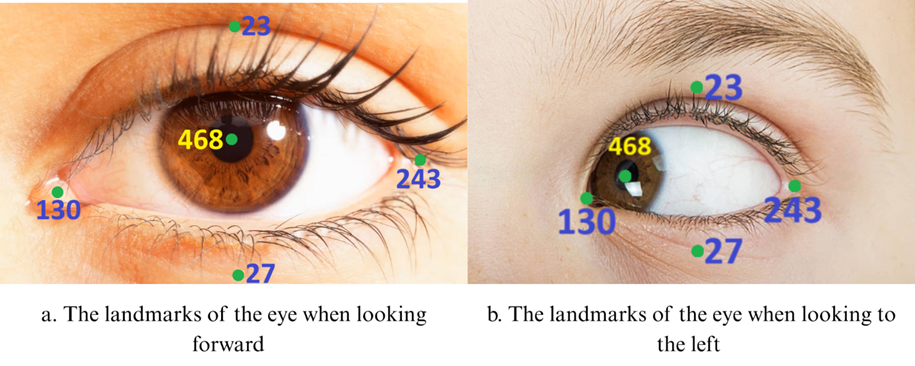


Where  represents the average eye aspect ratio  from the previous frame, and denotes the number of consecutive frames in which  is below 0.2, converted into seconds.

Since the model generates engagement level assessments every 10 seconds, we set the threshold for detecting a prolonged eye closure at 2 seconds. Both values are used to predict the learner's alertness level. If the blink rate and the number of prolonged eye closures persist over a certain period, it can be inferred that the learner is fatigued, showing signs of drowsiness, or even sleeping when the eyes remain closed for an extended duration. Conversely, when the learner returns to an alert state, this is indicated by the EAR value remaining above 0.20 for more than 5 seconds. At that point, the system predicts the learner as Alert.

**4.2.2 Gaze Direction Recognition**

Inspired by the study of Matthew Hullstrung [22], this research utilizes 10 key landmark points on the eyes, extracted from the data provided by MediaPipe’s Face Mesh. Specifically, Face Mesh assigns landmarks to key points around the eyes, including the inner and outer corners, the pupils, as well as the upper and lower parts of each eye. Figure 5 illustrates the eye landmarks across different gaze directions.



**Fig. 5. Visualization of Eye Landmarks for Gaze Direction Analysis**

Each eye landmark is represented by a pair of coordinates . Based on this set of points, the study defines a bounding region surrounding both eyes to determine the range of pupil movement, thereby analyzing the learner's gaze direction. The position of the pupil is identified along two primary coordinate axes: the horizontal axis  is measured from the inner to the outer corner of the eye, while the vertical axis  is determined from the upper to the lower eye landmark.

To determine gaze direction along two coordinate axes, we calculate the pupil position index based on the coordinates of key eye landmarks. On the horizontal axis, the relative position of the pupil is determined by the ratio of the distance from the pupil center to the inner eye corner to the total distance between the inner and outer eye corners. The formulas for computing the horizontal position index for the left eye  and the right eye ) are defined using equations (11) and (12):





Where, and  represent the horizontal position indices of the left and right eyes, respectively, and  denotes the x-coordinate of the -th landmark.

Similarly, along the vertical axis, the pupil position indices ( and ) reflect the relative position of the pupil concerning the upper and lower eye boundaries. These indices are computed as the ratio of the distance from the pupil to the lower landmark to the total distance between the upper and lower landmarks. The formula is expressed as follows (13) and (14):





Where,  and  represent the vertical position indices of the left and right eyes, respectively, and denotes the y-coordinate of the i-th landmark point.

The values of the four gaze indices  represent the relative position of the pupil within the constrained space of the eye. Based on predefined thresholds, the subject's gaze direction is inferred as follows:

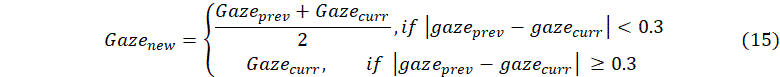
Leftward gaze: 

Rightward gaze: 

Upward gaze: 

Downward gaze: 

If the computed value falls outside the predefined thresholds, the system classifies the learner's gaze direction as straight ahead. Additionally, rapid eye movements, head oscillations, and variations in lighting conditions can introduce noise or blur in the captured images, leading to inaccuracies in gaze estimation. To mitigate these abrupt variations, we implement a signal filtering mechanism. Specifically, if the change in gaze direction between two consecutive frames is below a certain threshold, the system applies an averaging technique as defined in Equation (15) to smooth the data, thereby reducing noise interference and enhancing the stability of gaze tracking results.



Where represents the updated gaze value. The notation refers to the gaze value from the previous frame, while  denotes the gaze value in the current frame.

**4.3 EngageNet**

The proposed overall model integrates two main branches: the emotion recognition branch and the behavior prediction branch, aiming to assess learner engagement in online learning environments. The architecture is illustrated in Figure 2.

The input data consists of video frames captured from a camera, and processed frame-by-frame. The emotion recognition branch employs the LaCER model to classify them into five emotion states: "angry," "happy," "neutral," "sad," and "surprise". This branch outputs the predicted emotion probability (DEP) and the emotion weight (EW).

The behavior prediction branch computes the eye aspect ratio (EAR) from 12 key landmarks around both eyes to analyze eye blinking and closure duration. A prolonged eye closure is detected if it lasts more than 2 seconds, while an alert state is identified if EAR > 0.20 for over 5 seconds. Additionally, the model determines gaze direction using 10 eye landmarks, estimating the pupil position along the X and Y axes (left, right, up, down, straight). A noise filter (threshold < 0.3) is applied, and a behavioral weight (BE) is assigned to each detected gaze state.

The computed values DEP, EW, and BE are summarized in Tables 3 and 4. Finally, the model integrates data from both branches to compute the engagement level (EL) based on a formula (16) [23].



Where DEP represents the predicted emotion probability, with a minimum threshold of 50% if the classification is uncertain. EW and BE correspond to the emotion and behavior weights, respectively, as detailed in Tables 4 and 5.

The engagement level (EL) is classified into four categories: High Engagement (75% < EL ≤ 100%), Normal Engagement (50% < EL ≤ 75%), Low Engagement (25% < EL ≤ 50%), and Disengagement (EL ≤ 25%). The computation is performed every 10 seconds based on dominant features from emotion recognition, gaze direction, and alertness. The system provides real-time engagement classification, enabling instructors to monitor learning effectiveness. By leveraging CNN for emotion recognition and behavioral analysis, the model is optimized through temporal threshold adjustments and noise filtering, ensuring both accuracy and feasibility in online learning environments.

**Table 3.** **Emotion weight**

|  |  |
| --- | --- |
| **Domination Expression** | **Expression Weight** |
| Surprise  Happy  Angry  Neutral  Sad | 0.9  0.82  0.82  0.675  0.55 |

**Table 4.** **Behavior weight**

|  |  |  |  |
| --- | --- | --- | --- |
| Direction/Alert Status | Alert | Drowsy | Sleeping |
| Forward | 1 | 0.5 | 0 |
| Right | 0.25 | 0.125 | 0 |
| Left | 0.25 | 0.125 | 0 |
| Down | 0.25 | 0.125 | 0 |
| Up | 0.25 | 0.125 | 0 |

**5. RESULTS AND DISCUSSION**

**5.1 Experimental setup**

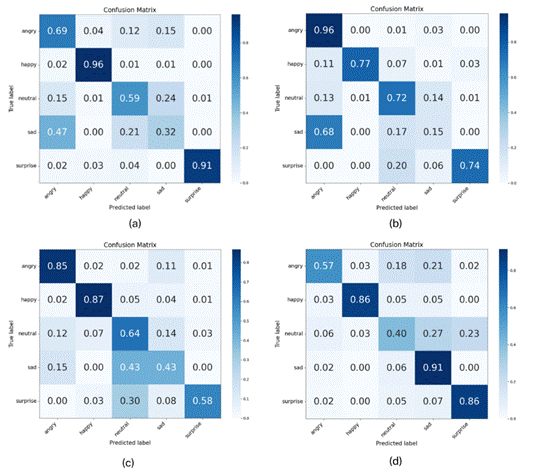
The training process for the emotion recognition model was conducted on a P100 GPU with 12GB of VRAM. The model utilizes the AdamW optimizer [24], with hyperparameters set to a learning rate of 0.001 and a weight decay of 0.0001. Training was performed over 100 epochs with a batch size of 32. Additionally, to evaluate the model's accuracy, we employed performance metrics such as Accuracy, Precision, Recall, and F1-score.

**5.2 Emotion Recognition Results**

The emotion recognition model was trained on a custom-built dataset, with the training results presented in Table 5. The LaCER model achieved an accuracy of 73.33%, while ResNet50, VGG19, and MobileNetV1 reached 72%, 74%, and 68%, respectively. These results reflect the emotion recognition capabilities of the models; however, performance variations may be attributed to the limited number of images and imbalanced data distribution. Although LaCER's accuracy is slightly lower than VGG19, its stability in terms of Precision, Recall, and F1-score, along with a significantly lower parameter count (1,688,453 compared to 20,554,821 for VGG19), makes it a more lightweight and efficient CNN model. Consequently, LaCER was selected for the emotion recognition branch. Figure 6 illustrates the model's prediction performance across different emotional expressions, highlighting accuracy discrepancies among emotion states.

**Table 5.** **Experimental Results on proposed dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | Feature | Accuracy | Precision | Recall | F1-scores | Params |
| LaCER | Landmarks | 73.33% | 69.95% | 69.41% | 69.09% | 1,688,453 |
| ResNet50 | RGB | 72% | 69% | 67% | 67% | 28,835,717 |
| VGG19 | RGB | 74% | 68% | 67% | 66% | 20,554,821 |
| MobileNetV1 | RGB | 68% | 67% | 72% | 66% | 7,431,365 |



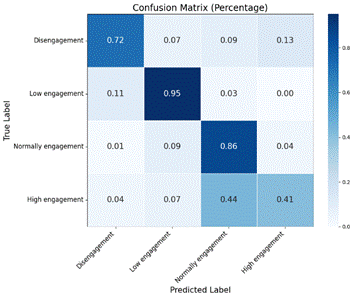
**Fig. 6. Confusion Matrix: (a) LaCER model, (b) VGG19 model, (c) ResNet50 model, (d) MobileNetV1 model**

**5.3 Results of the Engagement Evaluation Model**

The proposed model was tested on a self-constructed dataset comprising 1,294 video samples. Based on the Confusion Matrix (Figure 7) and the performance metrics outlined in Table 6, the model achieves an Accuracy of 79.76% and an F1-score of 73.41%, indicating reliable classification across four engagement levels: High Engagement, Normal Engagement, Low Engagement, and Disengagement. Furthermore, the confusion matrix highlights the greatest discrepancy between Normal Engagement and High Engagement, largely attributed to the limited number of High Engagement samples and the subtle boundary separating these two states. In contrast, the model effectively distinguishes between Disengagement and Low Engagement by integrating data such as blink frequency (EAR), gaze deviation from the screen, and relevant emotional states. Environmental factors, including head tilting, poor lighting, and particularly cases where learners display minimal facial expressions, may also present challenges to the model’s performance. Nevertheless, the results suggest that combining multiple factors, such as facial expressions via LaCER, gaze direction, and eye openness, enhances recognition stability compared to single-factor approaches.

**Table 6. Results of the Engagement Evaluation Model**

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1-score |
| 79.76% | 73.23% | 79.74% | 73.41% |



**Fig. 7. Confusion Matrix of Engagement Levels**

To further clarify how the model evaluates engagement levels across different stages, the analysis results from two learners throughout an online learning session are illustrated in Figure 8. Specifically, Dis and Normal Engagement are prevalent during the problem-solving (0s-900s), reflecting frequent transitions between the states of 'distraction' and 'normal focus'. In contrast, High Engagement only appears between 1000s–1400s, when learners participate in the minigame, where engagement peaks. This suggests that learners are influenced by the attractiveness of teaching activities; the problem-solving phase feels monotonous, leading to distraction, while minigame phase stimulates high interaction.A graph of a graph of a graph

AI-generated content may be incorrect.

**Fig. 8. Engagement Over Time of Two Learners Throughout the Online Class**

In this work, we significantly improved the performance of engagement assessment by integrating multiple emotional and behavioral factors. Specifically, combining facial emotion data with gaze direction and eye openness information not only increased accuracy but also provided a more comprehensive evaluation method compared to studies that relied on a single factor, such as facial expression analysis alone. Consequently, our model addressed limitations in distinguishing between closely related states, such as low and high engagement levels, a challenge that some prior studies, like that of Ma and colleagues [25], encountered due to overlapping states. However, this study also has certain limitations. The data collected in the online learning environment was somewhat limited, with an uneven quantity and quality of video samples and emotional instances. This could impact the model’s performance, particularly in accurately classifying emotional states and engagement levels across diverse scenarios. Additionally, environmental factors such as lighting and camera quality could significantly affect the accuracy of facial emotion recognition.

**6. CONCLUSION**

In this study, we propose EngageNet - a model for assessing learner engagement in online learning environments by integrating facial expression analysis, gaze direction, and eye openness. This approach enhances the accuracy of recognizing learners' focus and participation levels. Experimental results on a self-constructed dataset demonstrate that the model achieves an accuracy of 79.76%, proving its effectiveness and stability in evaluating learner engagement in virtual classrooms. Additionally, we introduce the LaCER model, which utilizes facial landmarks for emotion recognition. With an F1-score of 69.09%, LaCER exhibits greater stability and efficiency compared to other models such as ResNet50 (67%) and VGG19 (66%). As part of this research, we have also developed two datasets: an emotion dataset with labels including *angry, happy, neutral, sad,* and *surprise*, and a learner engagement dataset with categories such as *High Engagement, Normal Engagement, Low Engagement,* and *Disengagement*. These datasets enhance the model's practical applicability and real-world usability. This study makes a significant contribution by integrating emotional and behavioral information to improve the accuracy of learner engagement assessment. In future work, we aim to extend our research by incorporating advanced models such as Vision-LLM and knowledge graphs to further enhance accuracy and facilitate real-world deployment.

**DISCLAIMER (ARTIFICIAL INTELLIGENCE)**

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

**REFERENCES**

[1] Yeh, C. Y., & Tsai, C. C. (2022). Massive distance education: barriers and challenges in shifting to a complete online learning environment. *Frontiers in Psychology*, *13*, 928717.

[2] Muksin, S. N. B., & Makhsin, M. B. (2021). A level of student self-discipline in e-learning during pandemic Covid-19. *Procedia of Social Sciences and Humanities*, *1*, 278-283.

[3] AKPEN, Catherine Nabiem, et al. Impact of online learning on student's performance and engagement: a systematic review. *Discover Education*, 2024, 3.1: 1-15.

[4] Lasri, I., Riadsolh, A., & Elbelkacemi, M. (2023). Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning. *Education and Information Technologies*, *28*(4), 4069-4092.

[5] Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, *15*(9), 517.

[6] Xiang, A., Qi, Z., Wang, H., Yang, Q., & Ma, D. (2024, August). A multimodal fusion network for student emotion recognition based on transformer and tensor product. In *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)* (pp. 1-4). IEEE.

[7] Bergdahl, N. (2022). Engagement and disengagement in online learning. *Computers & Education*, *188*, 104561

[8] Elbawab, M., & Henriques, R. (2023). Machine Learning applied to student attentiveness detection: Using emotional and non-emotional measures. *Education and information technologies*, *28*(12), 15717-15737.

[9] Mutawa, A. M., & Sruthi, S. (2024). Enhancing human–computer interaction in online education: A machine learning approach to predicting student emotion and satisfaction. *International Journal of Human–Computer Interaction*, *40*(24), 8827-8843.

[10] Alruwais, N., & Zakariah, M. (2023). Student-engagement detection in classroom using machine learning algorithm. *Electronics*, *12*(3), 731.

[11] Ayouni, S., Hajjej, F., Maddeh, M., & Al-Otaibi, S. (2021). A new ML-based approach to enhance student engagement in online environment. *Plos one*, *16*(11), e0258788.

[12] Akpen, C. N., Asaolu, S., Atobatele, S., Okagbue, H., & Sampson, S. (2024). Impact of online learning on student's performance and engagement: a systematic review. *Discover Education*, *3*(1), 1-15.

[13] Dewan, M. A. A., Lin, F., Wen, D., Murshed, M., & Uddin, Z. (2018, October). A deep learning approach to detecting engagement of online learners. In *2018 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computing, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (pp. 1895-1902). IEEE.

[14] Santoni, M. M., Basaruddin, T., Junus, K., & Lawanto, O. (2024). Automatic detection of students’ engagement during online learning: a bagging ensemble deep learning approach. *IEEE Access*.

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

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

[17] Li, B., & Lima, D. (2021). Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering*, *2*, 57-64.

[18] Vignesh, S., Savithadevi, M., Sridevi, M., & Sridhar, R. (2023). A novel facial emotion recognition model using segmentation VGG-19 architecture. *International Journal of Information Technology*, *15*(4), 1777-1787.

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

[20] Thiha, S., & Rajasekera, J. (2023). Efficient online engagement analytics algorithm toolkit that can run on edge. *Algorithms*, *16*(2), 86.

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

[22]Hullstrung, M. (n.d.). *Gaze estimation* [GitHub repository]. GitHub. <https://github.com/matthullstrung/gaze-estimation/tree/main>

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

[24] Loshchilov, I., & Hutter, F. (2017). Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, *5*, 5.

[25] Ma, Y., Wei, Y., Shi, Y., Li, X., Tian, Y., & Zhao, Z. (2022). Online learning engagement recognition using bidirectional Long-Term recurrent convolutional networks. *Sustainability*, *15*(1), 198.