Scientific Reports

https://doi.org/10.1038/s41598-025-34871-x

Article in Press

Understanding the impact of emotional engagement on learning outcomes in online education: an automated analysis approach

Received: 10 September 2025

Accepted: 31 December 2025

Published online: 07 January 2026

Cite this article as: Chen G., Han G., Niu J. *et al.* Understanding the impact of emotional engagement on learning outcomes in online education: an automated analysis approach. *Sci Rep* (2026). https://doi.org/10.1038/s41598-025-34871-x

Guanyu Chen, Guangxin Han, Juan Niu & Juhou He

Understanding the Impact of Emotional Engagement on Learning Outcomes in Online Education: An Automated Analysis Approach

Guanyu Chen1[1] , Guangxin Han[2] , Juan Niu[3], Juhou He1*

1.2.3.*Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi'an, Shaanxi, 710062,China.
*Corresponding author(s). E-mail(s): mtth@snnu.edu.cn;

Abstract:

Online education offers flexibility but often suffers from reduced learner engagement. This study developed an automated method to detect emotional engagement using an optimized Vision Transformer model with Transfer Learning. Facial data from 40 undergraduates produced a dataset of 71,185 labeled images across three engagement levels. The proposed model achieved 93.8% classification accuracy, surpassing conventional machine learning and deep learning baselines. Analysis showed engagement typically declined after six minutes of learning, with a modest rebound near session end. Pearson correlation revealed a significant positive relationship between engagement and learning outcomes, indicating that emotionally engaged learners achieved higher academic performance. These results demonstrate the feasibility of deep learning–based approaches for scalable monitoring of learner engagement and highlight the important role of emotional states in shaping online learning effectiveness. The findings provide practical insights for designing adaptive interventions to sustain attention and optimize digital learning environments.

Keywords: Emotional Engagement, Online Learning, Artificial Intelligence in Education, Educational Data Mining

Introduction

With the rapid progress and technological advancements in education, internet services have gained widespread adoption and implementation across major universities, as well as primary and secondary schools. Consequently, online education has undergone substantial growth. Online education offers the flexibility of learning at any time and from anywhere, breaking free from the constraints of traditional learning environments, and granting access to a vast array of educational resources. However, it also presents certain challenges. One notable challenge is the inherent separation between students and teachers in the virtual field of online learning.(2) This physical divide makes it arduous for teachers to gauge the level of

student engagement in the learning process, a difficulty that becomes increasingly pronounced as the number of learners rises.(7) Compared to face-to-face instruction, the spatial and temporal detachment in online learning hinders effective communication and interaction between learners and educators, giving rise to a recurring sense of emotional disconnection. This emotional disconnect significantly impacts learners' online educational experiences and their subsequent outcomes. Therefore, from a pedagogical standpoint, it becomes imperative for educators to automatically discern students' emotional engagement levels during online learning, furnish timely feedback, and proactively undertake necessary measures to actively involve students in the learning journey.

As per the theory of learner engagement, learner engagement stands as the most effective predictor of student development, the level of learner engagement and emotions share a close association with academic performance.(5) (6) Several studies have demonstrated that learner engagement correlates with the extent of psychological investment in activities and can serve as a reliable predictor of learning outcomes.(18) Presently, the widely accepted definition of learner engagement, proposed by Fredricks in 2004, encompasses three dimensions: emotional engagement, behavioral engagement, and cognitive engagement. Among these dimensions, emotional engagement pertains to the degree and nature of learners' positive or negative emotional responses to teachers, peers, school, and academics.(24) Learners who experience a sense of enjoyment tend to be more motivated in tackling challenging problems.(12) Behavioral engagement focuses on learners' active involvement in social, academic, and extracurricular activities throughout their educational journey, emphasizing quantity over quality in terms of engagement in learning activities.(24) Cognitive engagement relates to the level of knowledge construction during the learning process.(23) Notably, Pekrun et al.'s research suggests that emotional engagement serves as a prerequisite for both cognitive and behavioral engagement. In the context of online learning, the analysis and feedback regarding learners' emotional engagement assume a critical role. This is because learners' emotional engagement can serve as an indicator of their willingness to learn, their needs, and their motivation throughout the learning process.(8) Experienced educators can monitor students' engagement by observing their facial expressions during instruction and adapt their teaching strategies and content accordingly. Facial expressions serve as indicators of a person's emotional engagement state.

Considering the limited sustained attention span of typical students, the level of emotional engagement tends to fluctuate at different stages during a class. Attention span refers to the duration of time an individual can concentrate on a task.(28) Wilson and Korn's literature review highlighted that students' attention tends to decline after approximately 10-15 minutes.(1) Several studies have investigated attention span, exploring various aspects such as the relationship between note-taking quantity and attention span(2) (8) (15), the correlation between the amount of retained information in students' memory and lecture duration,(17) and the connection between attention span and heart rate per minute.(5) Guo's research indicated that students' engagement remains high for the first 6 minutes when watching online learning videos, but

subsequently declines rapidly.(7) Therefore, the implementation of a feedback system that automatically analyzes learners' emotional engagement at different time intervals can assist teachers in summarizing their teaching plans and promptly updating their instructional strategies.

From a methodological standpoint, researchers have traditionally relied on manual coding and conventional machine learning methods to identify learners' emotional engagement in online learning. However, manual coding of datasets is a time-consuming process and is often plagued by issues such as sample imbalance and limited sample size. Furthermore, traditional machine learning methods lack robustness, which has impeded both theoretical and practical advancements in this field. In recent years, the Vision Transformer-based network models have become the state-of-the-art technology in image processing technology, and have made revolutionary achievements in image classification. They address many of the limitations associated with traditional approaches. However, the application of Vision Transformer-based models for detecting learners' emotional engagement in online learning has not been fully optimized or extensively explored. Moreover, the development of Vision Transformer-based detection and feedback systems specifically tailored to the context of online learning is still needed. This presents challenges in the field of educational research and practice, as researchers and educators strive to leverage the potential of these state-of-the-art technologies.

Consequently, this study seeks to accomplish several objectives: (1) Assess the capability of an optimized Vision Transformer model to infer emotional engagement from facial images captured by a camera.(2) Investigate the notable variations in emotional engagement among learners at different stages of the online learning process. (3) Explore the relationship between emotional engagement and learning outcomes. These studies will offer educators and learners valuable methodological and theoretical insights, enhancing their understanding of the significance of emotional engagement in promoting effective learning.

Methods

Ethics Statement

This study, "Understanding the Impact of Emotional Engagement on Learning Outcomes in Online Education: An Automated Analysis Approach," has been approved by the Ethics Committee of the Key Laboratory of Modern Teaching Technology, Ministry of Education, under approval number L20250904-02, on September 4, 2025. This project is conducted strictly in accordance with relevant laws and regulations and complies with the ethical guidelines established by the Declaration of Helsinki. All participants signed written informed consent forms before participating in the study and were fully informed of the research objectives and the use of their facial images in academic publications. All facial images included in the study were taken with the explicit consent of the participants, who agreed to their use in scientific research and publication. For participants who did not consent to

the use of identifiable images, all facial data was anonymized to ensure their privacy.

Research background and participants

Participants were recruited at a university in western China, involving 40 junior undergraduate students (M-age = 20.9) from various majors, excluding psychology and Marxist philosophy. The participants, consisting of 20 males and 20 females, provided informed consent after a thorough explanation of the study. The recruitment for the experiment began on September 12, 2025 and ended on September 19, 2025. Informed consent forms were distributed to all 40 volunteers who participated in the experiment, and all 40 volunteers agreed and signed the informed consent forms. The use of data and facial information in this experiment has been agreed and approved by all volunteers.

A dedicated laboratory setting was prepared to ensure an uninterrupted environment for the participants during their involvement in the study. The lighting conditions in the laboratory were not manipulated and comprised natural light from both indoor fluorescent lamps and outdoor sunlight. To capture facial video data of the participants during their online learning sessions, a computer equipped with a high-definition camera was set up in the laboratory. The participants' online learning processes were recorded using the EV screen recording software, combined with the high-definition camera.

For the experimental phase, three approximately 10-minute instructional videos were selected from the Chinese University MOOC website. The videos were titled 'The Psychology of Love' 'Innovative Thinking Behind Open Minds' and 'Fundamental Principles of Marxism'. All three online courses were classified as national quality courses offered by the Chinese University MOOC. Corresponding test questions were designed for each course to evaluate the participants' learning outcomes. The test questions we utilized were carefully selected from the supplementary test materials provided after the MOOC courses. These test questions were evaluated by two experts in the respective field of the course, who confirmed that they accurately reflect students' learning outcomes. The test questions are scored out of 10 and consist of four multiple-choice questions, two fill-in-the-blank questions, and one short-answer question. Participants were required to complete the respective test questions after watching each video to obtain their final test scores. The overall data collection process is illustrated in Figure 1.

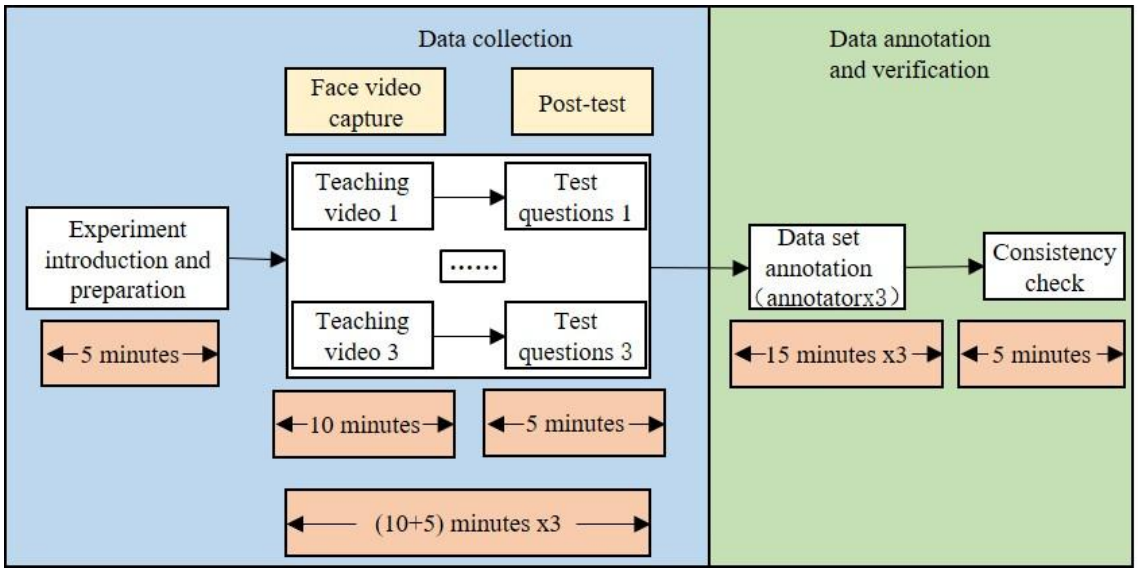

Figure 1. Flowchart for constructing a dataset of emotional engagement in online learning

In this data collection experiment, a total of 120 segments of online learning videos, each approximately 10 minutes in length, were collected. Building on the research conducted by Whitehill et al., which compared the usefulness of video-based sequences and image-based methods in recognizing engagement levels, this study found that image-based methods had relatively higher accuracy compared to video-based methods. This suggests that engagement is more of a spatial concept rather than a spatiotemporal one.(16) Based on these researches, we obtained a total of 71,185 images for further experimentation. Table 1 presents the number and proportion of images associated with each engagement level.

Learn emotional engagement	Highly engaged	Moderately engaged	Disengaged
Label	3	2	1
Number of pictures	16515	38468	16202
The proportion of the number of pictures	23.20%	54.04%	22.76%

Table 1. Distribution of the number of images for three levels of emotional engagement.

Research design

The research design consists of seven stages to address the objectives and research questions. Here is a detailed description of each stage.

Stage 1: Facial data is obtained from online learning environments using a webcam and stored in a database.

Stage 2: The collected data undergoes a cleaning process using Camtasia Studio video

editing software to remove any data that does not meet the experimental requirements. This step ensures that only valid and relevant data is retained for further analysis. (Camtasia is a software package produced by TechSmith in the United States that integrates computer screen recording and video editing. It also includes built-in features for Camtasia recorder, Camtasia Studio editor, Camtasia menu maker, Camtasia theater, Camtasia player, and Screencast).

Stage 3: Expert coders encode the emotional engagement data based on the theory of emotional engagement. The coders carefully analyze and label the collected data with the appropriate emotional engagement categories, applying their expertise and knowledge in emotional engagement research.

Stages 4 and 5: These stages involve the exploration of the first research question. The encoded emotional engagement data is utilized to train and evaluate optimized deep learning models. Through various iterations, the models are refined and adjusted to improve their performance in accurately identifying and classifying emotional engagement in the collected facial data.

Stage 6: The trained model with the best parameters is employed to identify and assign emotional engagement labels to unlabeled facial data. This allows for the automatic detection and classification of emotional engagement in previously unlabeled data.

Stage 7: Statistical analysis methods, such as Pearson correlation analysis, are utilized to address the second and third research questions. The collected data, including the labeled emotional engagement data and associated learning outcomes, are analyzed to examine the relationships between emotional engagement and learning outcomes. Statistical techniques are employed to determine the strength and significance of these relationships.

In the second stage, the cleaning process involves removing video data that does not meet the experimental requirements. Additionally, the videos are segmented into multiple video segments, ensuring that each segment contains only one category of emotional engagement. Camtasia Studio video editing software is employed for this purpose. Invalid video segments, where the learner's face is obscured or cannot be detected, are excluded during the segmentation process. Figure 2 illustrates the process of video segmentation using Camtasia Studio software, and Figure 3 displays the results of the video segmentation. In total, 1067 valid video segments were extracted from the initial 120 video segments for analysis in this study. We extracted one frame image every 5 frames from each video segment, excluding images that did not correspond to the engagement level of the video segment. The extracted images were then assigned the engagement level corresponding to their respective video segments. For instance, images extracted from highly engaged video segments were also assigned a highly engaged level. After data cleaning and annotation, a total of 71,185 images were obtained.

Figure 2. Video trimming using Camtasia Studio software.

Figure 3. Video segmentation results.

Coding scheme

In the field of online learning, learners' facial expressions generate a substantial volume of data, which poses challenges in terms of the time required for manual coding. To overcome this methodological challenge, we have developed an optimized

deep learning model. To effectively train and evaluate this model, we have devised an encoding scheme for emotional engagement levels in online learning, drawing upon the theory of learners' emotional engagement.(14) (26) The encoding scheme, presented in Table 2, categorizes learners' emotional engagement into three distinct levels: highly engaged, moderately engaged, and disengaged. Each emotional engagement category is thoroughly described in Table 2, providing detailed insights into the characteristics and attributes associated with each level of emotional engagement.

Class	Head features	Eye features	Facial expression features
Highly engaged	Head upright or inclined forward	Staring at the screen, eyes unconsciously widening, increased distance between upper and lower eyelids	Surprise, joy, focus, enthusiasm, and other positive expressions.
Moderately engaged	Head generally upright or slightly tilted to the left or right	Line of sight positioned within the screen area, eyes open normally, no change in the distance between upper and lower eyelids	Calm, neutral and other neutral expressions
Disengaged	Head not upright and significant tilt to the left or right	The line of sight is positioned at the edge of the screen area or outside the screen area, eyes slightly closed or even completely closed, and the distance between the upper and lower eyelids decreases	Bored, tired, indifferent, and other negative expressions

Table 2. Coding scheme for learning emotional engagement in online learning.

In the third phase, to ensure the quality and credibility of the dataset constructed for learners' engagement, a crowdsourcing approach was employed for data annotation. Three students with academic backgrounds in educational technology were recruited as data annotators, and they underwent training to familiarize themselves with the relevant definitions of learners' engagement states, the annotation tools, and the specific definitions of the three engagement labels. During the training, a portion of annotated data was provided for practice, and discussions and Q&A sessions were organized to address any issues or questions encountered by the annotators. Guidance and clarification were provided to resolve doubts or disagreements and to ensure a consensus among the annotators. Based on the performance of the annotators during training, they were confirmed as data annotators to participate in the annotation task.

To ensure data annotation validity and reliability, we adopted the consistency check method proposed by Kaur et al., using Kendall's coefficient of agreement.(17) The

results of Kendall's coefficient of agreement for the data annotation by all annotators revealed a high level of consistency, with a Kendall's coefficient of agreement of 0.889 ($p < 0.01$). This high reliability and accuracy of the data annotation confirm the validity and suitability of the annotated data for training and evaluating the online learning emotional engagement recognition model.

Automatic engagement detection based on Vision Transformer network and Transfer Learning

In the fourth and fifth stages of the study, an analysis of the encoding scheme for the emotional engagement data was conducted, revealing an issue of class imbalance within the collected dataset. Additionally, due to the smaller number of participants and a larger number of training samples per participant, there was limited diversity in the data, leading to a smaller intra-class distance and a larger inter-class distance. To address these challenges and improve the model's performance, robustness, and generalization capabilities, the study considered the possibility of pretraining the Vision Transformer network model using the DAiSEE dataset.(11) The pretrained model's weights would then be utilized as the initial weights for further training using the self-built emotional engagement dataset. By leveraging the pretrained model and incorporating it into the training process, it was anticipated that the model's performance and generalization abilities could be enhanced, leading to improved results in recognizing and classifying emotional engagement in the online learning context. This approach aimed to address the issue of limited data diversity and enhance the overall effectiveness of the model.

The Vision Transformer network model, introduced by Dosovitskiy et al. in 2020, is a notable innovation that adapts the Transformer architecture, originally designed for natural language processing tasks, to the field of computer vision.(21) This model represents a self-attention-based approach to image classification. In contrast to traditional convolutional neural networks, the Vision Transformer does not employ convolutional layers but instead relies exclusively on self-attention mechanisms to extract relevant features from images. The architecture of the Vision Transformer model is visualized in Figure 4, showcasing the arrangement of self-attention layers and feed-forward neural networks. Through the use of self-attention mechanisms, the model captures dependencies between different regions of an image, enabling it to effectively process and understand the visual information. This innovative approach has shown promising results in various computer vision tasks and has the potential to significantly impact the field of image classification.

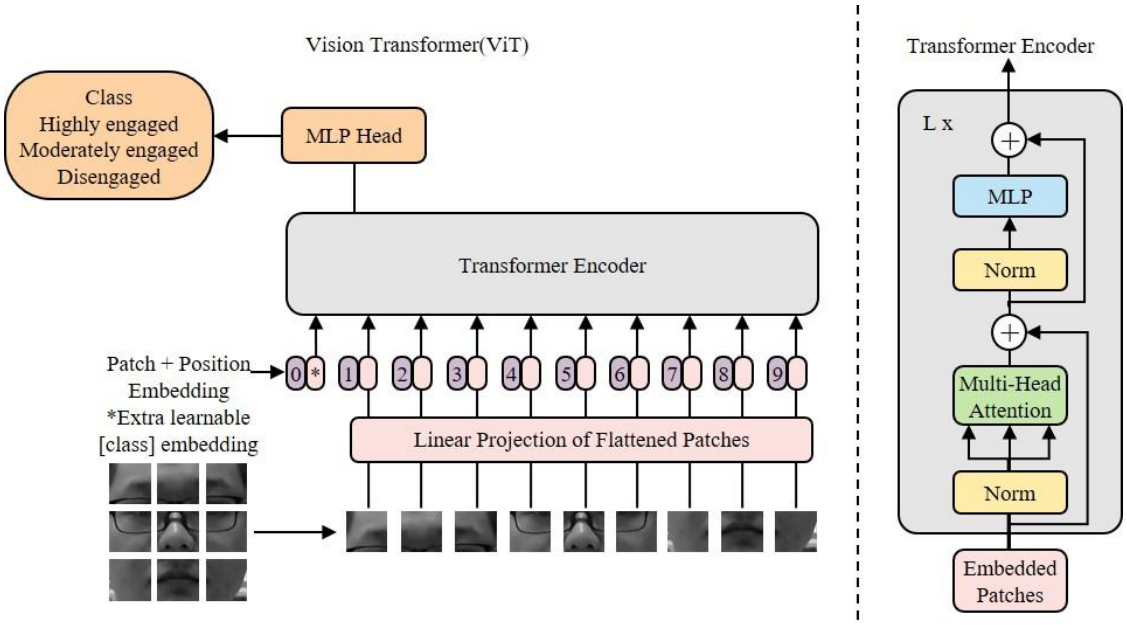

Figure 4. Vision Transformer model architecture diagram.

To capture more comprehensive and detailed feature information, the Vision Transformer model employs a multi-head self-attention mechanism. This mechanism involves running multiple self-attention mechanisms simultaneously, and then combining their outputs through concatenation and linear transformation to achieve the desired output dimensionality. The calculation formulas for the multi-head self-attention are provided in Equations (1) and (2) :

$$MultiHead(Q, K, V) = Concat(head_1, \cdots, head_h)W^O \tag{1}$$

where Q, K, and V represent the query vector matrix, key vector matrix, and value vector matrix, respectively. The MultiHead function concatenates the outputs of each individual self-attention head, denoted as $head_i$, for $i = 1 \ldots h$ (the total number of heads). W^O is the weight matrix used for linear transformation. Each self-attention head $head_i$ performs the following calculations:

$$head_i = Attention\left(Q \cdot W_i^Q, K \cdot W_i^K, V \cdot W_i^V\right) \tag{2}$$

where W_i^Q, W_i^K, and W_i^V are the learnable weight matrices for the query, key, and value projections of the ith self-attention head. The $Attention$ function computes the attention scores and applies them to the values to obtain the attended output. During the training process, we first pre-trained the Vision Transformer network model on the DAiSEE dataset, and then fine-tuned it on our self-built dataset of learners' learning engagement.

Data analysis and automated feedback model

To address the first research question, In the original methodology, we initially performed ‘10-fold cross-validation’ on the image-level data. However, this approach could lead to an overestimation of model performance because it is possible for data from the same participant to appear in both the training and test sets, violating the independence of the training and testing data.

To address this issue and provide a more accurate performance evaluation, we adopted a ‘subject-based k-fold cross-validation’ approach. In this method, the 40 participants were divided into 5 folds, with each fold containing 8 participants. For each fold, the data from 8 participants were used as the test set, while the data from the remaining 32 participants were used for training. This ensured that the data from each participant was either in the training set or the test set, but never in both, effectively eliminating the risk of data leakage and ensuring a fair evaluation of the model's performance.

This modification was made due to the potential difficulties in recalling and organizing all 40 participants for repeated testing in real-world settings. Therefore, the ‘K-fold cross-validation by subject’ method was chosen as a more practical solution. This method ensures that the performance metrics obtained from the cross-validation process reflect the model's true ability to generalize to new, unseen participants, rather than being inflated by repeated data from the same participants.

In this approach, the model was evaluated using performance metrics such as accuracy, precision, recall, and F1-score across all folds. This more rigorous evaluation ensures that the model's performance is both reliable and realistic, providing a true measure of its generalization capabilities.

To avoid overestimating the performance due to data leakage, we employed a ‘subject-based k-fold cross-validation’ method, where the 40 participants were divided into 5 folds, with each fold containing 8 participants. This method ensured that the data from the same participant were not used in both the training and test sets, thus providing a more accurate and unbiased estimate of the model’s performance.

To address the second research question, an automatic detection and feedback system was developed. The process flowchart of the system is illustrated in Figure 5. This system facilitated the analysis of learner engagement recognition and variations during online learning. The analysis results were communicated to teachers in a timely manner, enabling them to better understand and respond to learners’ engagement levels. In the seventh stage, Pearson correlation analysis was employed to investigate the relationship between learners’ emotional engagement and their learning outcomes, providing insights into the impact of emotional engagement on learning effectiveness.

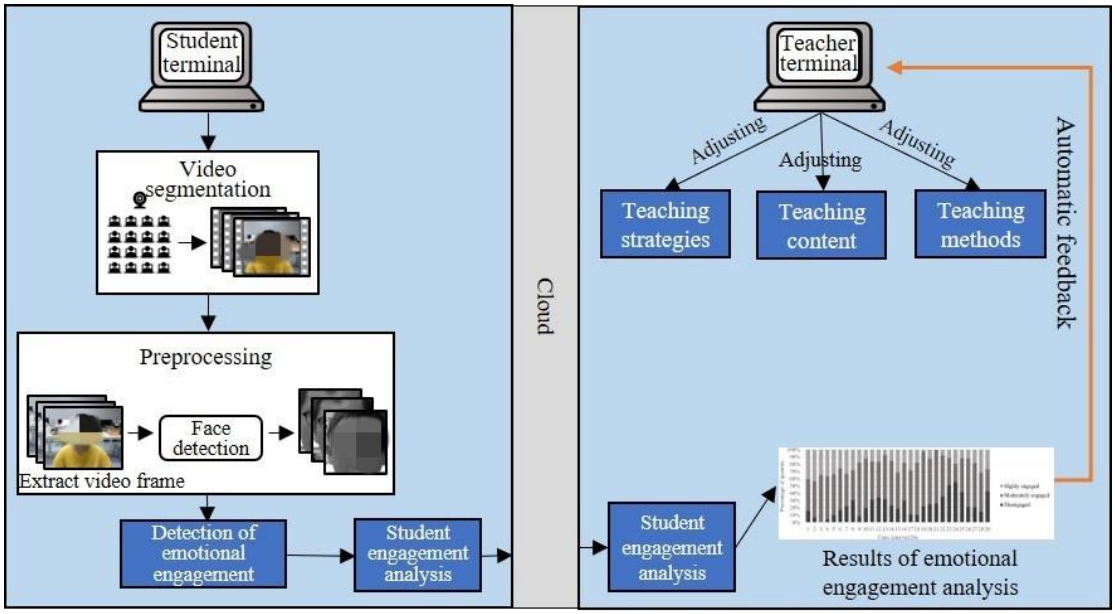

Figure 5. The automatic detection and feedback system.

Quizzes Validation, Reliability, and Counterbalancing

Quizzes Validation

To ensure the validity of the tests, we performed item analysis on each test. The purpose of item analysis was to assess whether each test item could effectively differentiate between participants with high and low performance, thereby ensuring that the tests were indeed effective in measuring the intended affective engagement and academic performance. Each item was designed according to specific learning objectives to ensure that the tests comprehensively reflected participants' affective engagement and learning outcomes.

Reliability

The internal consistency of the quizzes was assessed using Cronbach's α. The Psychology of Love quiz achieved $\alpha = 0.89$, the Innovative Thinking quiz $\alpha = 0.82$, and the Marxism quiz $\alpha = 0.77$, indicating acceptable reliability across all three courses.

Resolution of Annotation Disagreement and Item Review

To ensure annotation quality, all video segments were independently coded by two trained expert annotators. After the initial round of coding, disagreements between annotators were identified through item-level comparison. When inconsistencies occurred, the annotators first engaged in a structured discussion to review the specific video segments and justify their coding decisions. If consensus could not be reached

through discussion, a third senior annotator acted as an adjudicator and made the final determination.
In addition, item-level reliability was examined during this process. For quiz items or annotation codes that demonstrated low inter-annotator agreement, the items were reviewed and clarified to remove ambiguities. No items were removed because all items achieved acceptable reliability after revision and consensus adjudication. This multi-step procedure ensured that the final annotation set met the required standard of inter-rater reliability.

Counterbalancing

To eliminate the influence of order and topic effects on the results, we implemented counterbalancing and randomization on the presentation order of the tests. Specifically, we used a counterbalancing method to randomize the order in which each participant received the test. This ensured that each participant received the test items in a different order, thus avoiding the interference of order effects on the results. Simultaneously, we also randomized the different topics involved in the test to ensure that the order of the topics did not affect the participants' performance.
Through this randomization and counterbalancing design, we effectively controlled the confounding effects that might arise from the test order and topic order, ensuring the reliability of the results and ensuring that the relationship between affective engagement and learning outcomes is not interfered with by external order effects.

Results

Repeat analyses on subject-split data

Given the challenges of recalling all 40 participants and organizing repeated experiments in real-world settings, we opted for K-fold Cross-Validation by Subject, a more practical solution. In this approach, the 40 participants were divided into 5 folds, with each fold containing data from 8 participants. For each fold, the data from 8 participants were used as the test set, while the data from the remaining 32 participants were used for training. This method ensured that the model was trained on data from one set of participants and evaluated on a completely separate set, providing a more realistic and generalizable evaluation of the model's performance.
This subject-based cross-validation approach also provides a more efficient alternative to leave-one-out cross-validation, as it allows for fewer iterations while still maintaining data independence between the training and testing phases. It ensures that the model is not biased by repeated data from the same subjects, making the evaluation more reliable and representative of the model’s real-world applicability.
In this study, the original dataset contained 71,185 images. These images were categorized into three groups based on emotional engagement: Highly Engaged (16,515 images, approximately 23.20%) , Moderately Engaged (38,468 images, approximately 54.04%), and Disengaged (16,202 images, approximately 22.76%). We

conducted 5-fold Cross-Validation by Subject (each time with 8 participants as the test set) to acquire and analyze accuracy, precision, and recall data across the five folds. The results for each fold include Accuracy, Precision, and Recall, and the final valueswill be the mean and standard deviation. All quiz items and annotation codes met the acceptable reliability threshold after discussion-based reconciliation, and therefore no items were removed from the analysis. The simplified data is in Table 3.

Fold Number	Accuracy(%)	Precision(%)	Recall(%)
Fold 1	88.75	86.56	85.21
Fold 2	94.51	85.58	94.70
Fold 3	92.32	93.66	93.32
Fold 4	90.99	91.01	87.12
Fold 5	86.56	92.08	86.82
Mean	90.62	89.78	89.43
Std Dev	3.09	3.53	4.27

Table 3. K-fold Cross-Validation by Subject Fold Results Summary(simplified data)

The results indicate that the model's performance was lower than the initial image-based cross-validation, as expected, due to the more stringent evaluation process. The mean accuracy across all 5 folds was '90.62%', with a standard deviation of '3.09%'. The precision and recall metrics were also consistent, with average values of '89.78%' and '89.43%', respectively. These results indicate that the model maintains a stable performance when evaluated using independent subject data, ensuring that the reported performance metrics are not inflated.

In Table 4 , To ensure full transparency, the Cronbach's α values for each quiz are reported as follows: the Psychology of Love quiz demonstrated an internal consistency of $\alpha = 0.89$ (naturally higher), the Innovative Thinking quiz showed $\alpha = 0.82$ (moderate internal consistency) , and the Marxism quiz yielded $\alpha = 0.77$ (indicating acceptable internal consistency) . All three values fall within the acceptable reliability range, indicating that the quiz items consistently measured the intended learning constructs.

Course	Cronbach's α
Psychology of Love	0.89
Innovative Thinking	0.82
Marxism	0.77

Table 4. The Cronbach's α values for each quiz

By utilizing 'K-fold cross-validation by subject', we were able to present a more reliable evaluation of the model's true generalization ability, which is essential for its practical application in real-world scenarios.

Correlation analysis

Performance Metrics

The results of the analysis showed that there was a significant positive correlation between emotional engagement and learning outcomes (measured by quiz scores). The mean accuracy of the model across all 5 folds was 90.62%, with a standard deviation of 3.09%. The precision and recall values were 89.78% and 89.43%**,** respectively. These results suggest that the model was able to effectively predict emotional engagement based on the quiz data.

Correlation Between Engagement and Scores

The relationship between emotional engagement and learning outcomes was analyzed using Pearson's correlation. The correlation coefficient was 0.68, indicating a moderate to strong positive relationship between engagement levels and quiz performance. This suggests that students who were more emotionally engaged in the learning process tended to perform better in the quizzes. However, it is important to emphasize that this is a correlational analysis and not a causal one, so no claims can be made about the direction of influence. The observed relationship should not be interpreted as one variable causing the other.

Controlling for Order and Topic Effects

To ensure the validity of the results, potential order effects and topic effects were controlled through the use of randomization and counterbalancing. The randomization of quiz order ensured that no specific sequence of questions influenced the participants' responses. The counterbalancing technique further minimized any bias from the sequence of topics covered in the quizzes. As a result, the observed correlations between engagement and quiz performance are not confounded by these extraneous factors.

Optimized Vision Transformer model identify learners' emotional engagement in online learning

The purpose of this experiment was to evaluate the performance of the optimized Vision Transformer model and Transfer Learning model in detecting emotional engagement. Firstly, the Vision Transformer network model was pre-trained using the DAiSEE dataset to enhance its feature representation and generalization ability. To assess the impact of Transfer Learning on model performance, comparative experiments were conducted to evaluate the accuracy of models with and without pre-training. The term 'without pre-training' refers to the Vision Transformer network model trained directly on the self-built learning engagement dataset without prior pre-training on the DAiSEE dataset. The results of the comparative experiments are

presented in Table 5.

	Accuracy	Macro-Recall	Macro-Precision	Macro-F1
Vision Transformer	91.79 ± 1.54	91.48 ± 1.72	93.04 ± 1.46	91.99 ± 1.55
Vision Transformer +Transfer Learning	93.82 ± 1.20	93.78 ± 1.04	94.26 ± 0.80	93.95 ± 1.12

Table 5. Comparison of experimental results (%) before and after Transfer Learning.

Table 5 presents a comparison of model performance before and after the application of Transfer Learning. As shown in the table, incorporating Transfer Learning leads to consistent improvements across all evaluated metrics, including accuracy, precision, recall, and F1-score. These results indicate that fine-tuning the pre-trained model on the target dataset effectively enhances classification performance and improves the model's ability to generalize across participants.

Table 6 compares the performance of the proposed Vision Transformer–based approach with several baseline models. The results demonstrate that the Vision Transformer consistently outperforms the baseline architectures across all performance metrics. This comparison highlights the effectiveness of the Vision Transformer architecture for modeling emotional engagement in online learning scenarios.

Additionally, we compared the classification performance of our proposed optimized Vision Transformer + Transfer Learning model with other models in the task of emotional engagement detection. The comparison methods were as follows: Gabor+SVM. (42) Decision tree, (41) ResNet+TCN. (26) LDP-KPCA-DBN. (17) The algorithm is described in Section 2.2. Table 6 presents the comparison of the performance results on the self-built dataset between our algorithm and the baseline models.

Class	Gabor+SVM	Decision Tree	ResNet+TCN	LDP-KPCA-DBN	Vision Transformer +Transfer Learning
Highly engaged	77.52 ± 1.61	89.22 ± 1.13	89.93 ± 1.72	95.52 ± 1.68	95.97 ± 1.01
Moderately engaged	73.65 ± 1.82	73.17 ± 1.46	82.24 ± 1.16	85.83 ± 1.41	92.46 ± 0.79
Disengaged	82.19 ± 1.81	83.15 ± 1.72	84.02 ± 1.19	83.39 ± 1.02	93.03 ± 1.29
Mean	77.78 ± 1.52	81.84 ± 0.94	85.40 ± 1.29	88.25 ± 1.13	93.82 ± 1.20

Table 6. Classification results (%) of emotional engagement.

From Table 6, it can be observed that the Vision Transformer–based approach consistently outperforms other baseline models across all engagement categories.

Students' engagement vary during online learning

Figure 6 illustrates the overall automatic recognition results of learners' engagement. The results of the self-built dataset of emotional engagement reveal interesting patterns throughout the online course. It can be observed from the figure that the majority of learners' engagement is categorized as 'moderate engagement' throughout the entire duration of the course. However, as the learning session progresses, there is a gradual decline in learners' engagement. The proportion of 'high engagement' and 'moderate engagement' decreases, while the proportion of 'low engagement' increases. Notably, around the 6-minute mark, there is a more pronounced increase in the proportion of 'low engagement', indicating a decline in learners' overall engagement during that period. Interestingly, a slight improvement in learners' engagement was observed towards the end of the course, despite the course nearing its conclusion. This suggests a potential rebound or revitalization of engagement towards the end of the learning session. These insights provide essential information about the fluctuations in learners' engagement levels during online learning and highlight specific time periods when learners may experience changes in their level of engagement. Understanding these patterns can be valuable for educators and instructional designers to optimize learning experiences and implement interventions to sustain and enhance learner engagement throughout the course.

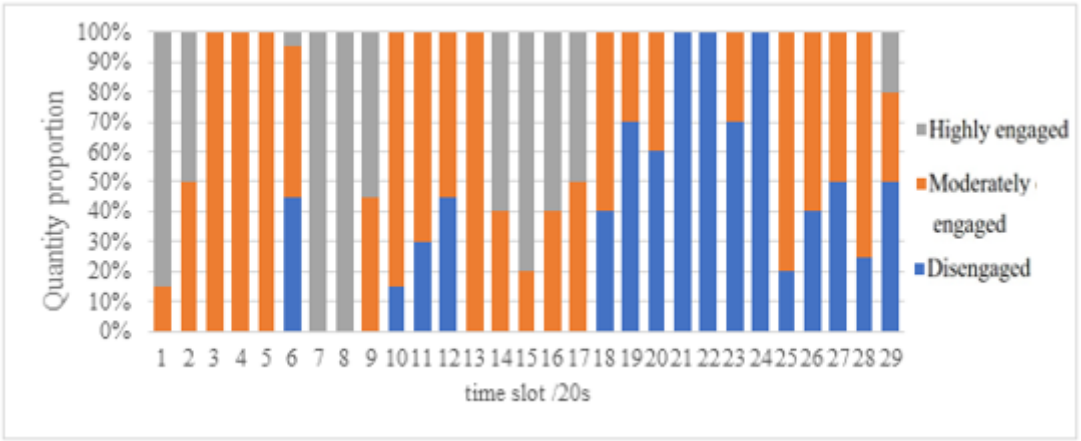

Figure 6. The automatic detection results of all learners.

Figure 7 presents the overall manual annotation results of learners' engagement during online learning. The experimental results demonstrate the effectiveness of our proposed online learning sentiment recognition model in accurately identifying and categorizing learners' engagement levels. The manual annotation results are generally consistent with the automated detection results, indicating that our algorithm is effective in capturing the overall patterns of learners' changes throughout the online

learning process. The alignment between manual annotations and automated detection results allows for a deeper analysis of consistent patterns in learners' engagement. This highlights the robustness and reliability of our algorithm in providing timely and accurate understanding of learners' engagement states. Moreover, the effectiveness of our algorithm becomes particularly evident when comparing it to the automated detection and feedback processes. Our algorithm provides a comprehensive and nuanced understanding of learners' engagement, surpassing the limitations of automated detection alone. It enables us to gain valuable insights into the dynamics of learners' engagement, facilitating timely and effective interventions to enhance the learning experience. Overall, the experimental results from Figure 7 affirm the effectiveness of our online learning sentiment recognition model in identifying learners' engagement during online teaching. This supports the notion that our algorithm is a valuable tool for gaining a deeper understanding of student states and optimizing the online learning environment.

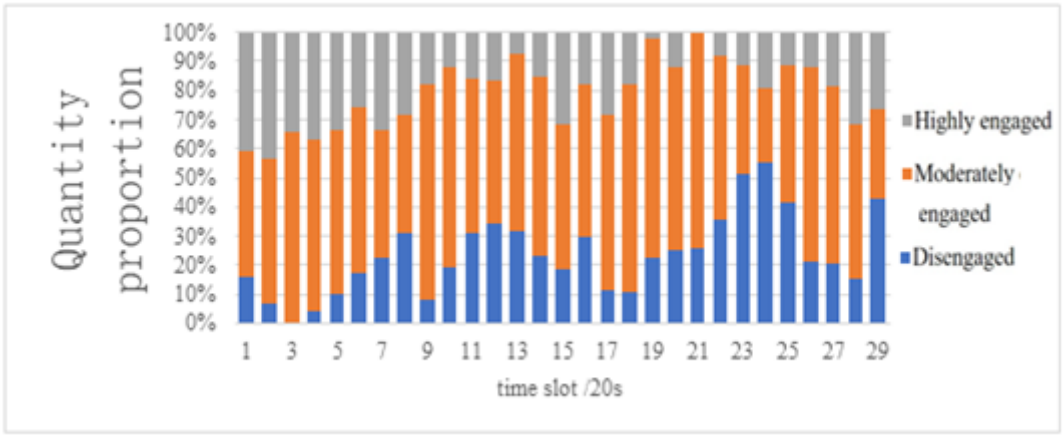

Figure 7. The manual annotation results of all learners.

The relationship between emotional engagement and learning outcomes

In Table 7, the statistical analysis results are presented for the automatic detection results, manual annotations, average of manual annotations, and test results. The metrics included in the table are the mean and standard deviation (SD). For the automatic recognition results, the mean values on three online courses are 2.470 ,1.931 ,1.651 , respectively, indicating the average emotional engagement level of all learners as determined by the online learning emotion recognition model. The SD represents the variability in the recognition results. The manual annotations by three annotators provide an additional perspective on learners' emotional engagement. The average of manual annotations combines the individual annotations from the three annotators, providing a more comprehensive assessment of learners' emotional engagement. Lastly, the test scores represent the learners' performance on the quiz questions designed to assess their learning effectiveness. The mean and SD values of the test results are provided. These statistical analysis results offer valuable insights into the agreement between the automatic recognition results and manual annotations,

as well as the learners' learning outcomes. They provide a quantitative assessment of the emotional engagement levels, helping to validate the effectiveness of the online learning emotion recognition model and evaluate the learners' understanding and retention of the video content.

Course ID		Annotator1	Annotator2	Annotator3	Average of manually annotated results	Automatic detection	Test score
1	Mean	2.782	2.224	2.565	2.523	2.470	8.100
	SD	0.323	0.329	0.410	0.377	0.323	1.370
	N	40	40	40	120	40	40
2	Mean	2.10	1.98	1.89	1.986	1.931	6.300
	SD	0.201	0.216	0.260	0.226	0.245	0.823
	N	40	40	40	120	40	40
3	Mean	1.73	1.80	1.63	1.716	1.651	4.400
	SD	0.122	0.174	0.158	0.175	0.121	0.516
	N	40	40	40	120	40	40

Table 7. Statistical analysis of emotional engagement and learning outcomes.

Table 8 presents the results of the Pearson correlation analysis conducted to examine the relationship between learners' emotional engagement and their learning outcomes. The variables analyzed include emotional engagement through manual annotation and automatic detection, as well as learning outcomes. Whether it is automatic emotion recognition or manual annotation results, there is a significant relationship with the learning outcomes. Specifically, the coefficient correlations between automatic detection and test score were r=0.860**,0.664*, and 0.707*. The coefficient correlations between manual annotation and test score were r=0.799**,0.657*, and 0.636*. This suggests that the correlation between emotional engagement and learning outcomes is statistically significant. The results of the Pearson correlation analysis support the hypothesis that learners' emotional engagement has a positive influence on their learning outcomes. This experimental result emphasizes the importance of emotional engagement in online learning and suggests that learners who are more emotionally engaged tend to achieve better learning outcomes. Overall, the results of the Pearson correlation analysis provide evidence of a significant and positive correlation between learners' emotional engagement and their learning outcomes, highlighting the relevance of emotional engagement in the context of online learning.

Course ID		Manual annotation	Automatic detection	Test score
1	Manual annotation	1		
	Automatic detection	0.869**	1	
	Test score	0.799**	0.860**	1
2	Manual annotation	1		
	Automatic detection	0.876**	1	
	Test score	0.657*	0.664*	1
3	Manual annotation	1		
	Automatic detection	0.887**	1	
	Test score	0.636*	0.707*	1

Note. N = 40, *$p < 0.05$, **$p < 0.01$, **$p < 0.001$

Table 8. The Pearson correlation analysis between test score, automatic detection and manual annotation.

Discussion

This study investigated the relationship between emotional engagement and learning outcomes in online courses using automated facial-expression analysis supported by Vision Transformer models. The findings demonstrate a clear positive association between learners' emotional engagement and their quiz performance, suggesting that emotional engagement provides meaningful insight into how learners interact with online content. By applying Transfer Learning, the model achieved stable performance across subject-based validation, indicating that the proposed approach can generalize across different individuals despite limited data quantities.

Importantly, the temporal analysis revealed a general decline in emotional engagement over the duration of the courses, with a slight recovery toward the end. This pattern aligns with prior studies suggesting that sustained participation in online environments often leads to reduced attentional and emotional investment. However, the present work extends this understanding by quantifying engagement using automated, fine-grained behavioral indicators rather than self-report measures.

At the same time, several limitations must be acknowledged. The three MOOCs included in the study cover heterogeneous content domains, which may influence how learners emotionally respond to different topics. This content heterogeneity introduces potential confounding effects, especially when interpreting cross-course trends, and limits generalization. Although subject-based cross-validation addressed participant-level variance, topic-level variance was not fully controlled. Future work should therefore incorporate more homogeneous course materials or conduct course-level stratified analyses.

Despite these limitations, the study provides evidence that automated engagement detection, combined with advanced deep learning models, can offer reliable indicators

for understanding online learning behaviors. These findings support the broader use of affective analytics to enhance adaptive learning systems and improve instructional design in digital education environments.

Conclusions

This study demonstrates that emotional engagement, detected through automated facial-expression analysis, is positively associated with learning performance in online courses. The Vision Transformer model, enhanced through Transfer Learning, provided reliable recognition under subject-based validation and supported the analysis of engagement trends.

The results highlight the potential of integrating affective analytics into online learning platforms to better monitor learners' engagement and support personalized interventions. Nonetheless, the heterogeneity of course topics and the limited number of participants impose constraints on generalizability. Future research should employ more standardized course content, larger samples, and longitudinal designs to further validate the proposed approach.

Funding

Funding: National Natural Science Foundation of China (No. 62177032), “Research on the Autonomous Training and Evaluation Model for Pre-service Teachers’ Classroom Teaching Expression Competence.”

References

1. Means, Barbara and Toyama, Yuki and Murphy, Robert and Bakia, Marianne and Jones, Karla. Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. Project Report. Centre for Learning Technology[M]. Association for learning technology, 2009(8).
2. Venton B J, Pompano R R. Strategies for enhancing remote student engagement through active learning[J]. Analytical and Bioanalytical Chemistry, 2021, 413(6): 1507-1512.
3. Sümer Ö, Goldberg P, D'Mello S, et al. Multimodal engagement analysis from facial videos in the classroom[J]. IEEE Transactions on Affective Computing, 2021.
4. Baker R S, D'Mello S K, Rodrigo M M T, et al. Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments[J]. International Journal of Human-Computer Studies, 2010, 68(4): 223-241.
5. D’Mello S, Lehman B, Pekrun R, et al. Confusion can be beneficial for learning[J]. Learning and Instruction, 2014, 29: 153-170.
6. Jagers R J, Rivas-Drake D, Williams B. Transformative social and emotional learning (SEL): Toward SEL in service of educational equity and excellence[J]. Educational Psychologist, 2019, 54(3): 162-184.
7. Pabba C, Kumar P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition[J]. Expert Systems, 2022, 39(1):e12839.1-e12839.28.
8. Bosch N, Chen Y, D’Mello S. It’s written on your face: detecting affective states from facial expressions while learning computer programming[C]. Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014 Proceedings 12. Springer, 2014:39-44.
9. Dewan M, Murshed M, Lin F. Engagement detection in online learning: a review[J]. Smart Learning Environments, 2019, 6(1): 1-20.
10. D'Mello S K, Craig S D, Graesser A C. Multimethod assessment of affective experience and expression during deep learning[J]. International Journal of Learning Technology, 2009, 4(3-4): 165-187.
11. D’Mello S K, Graesser A. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features[J]. User Modeling and User-Adapted Interaction, 2010, 20: 147-187.
12. Whitehill J, Serpell Z, Lin Y-C, et al. The faces of engagement: Automatic recognition of student engagementfrom facial expressions[J]. IEEE Transactions on Affective Computing, 2014, 5(1): 86-98.
13. Kamath A, Biswas A, Balasubramanian V. A crowdsourced approach to student engagement recognition in e-learning environments[C]. 2016 IEEE Winter

Conference on Applications of Computer Vision (WACV). IEEE, 2016:1-9.
14. Gupta A, D'Cunha A, Awasthi K, et al. Daisee: Towards user engagement recognition in the wild[J]. arXiv preprint arXiv:160901885, 2016.
15. Kaur A, Mustafa A, Mehta L, et al. Prediction and localization of student engagement in the wild[C]. 2018 Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2018:1-8.
16. Mukhopadhyay M, Pal S, Nayyar A, et al. Facial emotion detection to assess Learner's State of mind in an online learning system[C]. Proceedings of the 2020 5th international conference on intelligent information technology. 2020:107-115.
17. Wei Q, Sun B, He J, et al. BNU-LSVED 2.0: Spontaneous multimodal student affect database with multi-dimensional labels[J]. Signal Processing: Image Communication, 2017, 59: 168-181.
18. Bian C, Zhang Y, Yang F, et al. Spontaneous facial expression database for academic emotion inference in online learning[J]. IET Computer Vision, 2019, 13(3): 329-337.
19. Schmieder A. A glossary of educational reform[J]. Journal of Teacher Education, 1973, 24(1): 55-62.
20. Fisher C W, Berliner D C, Filby N N, et al. Teaching behaviors, academic learning time, and student achievement: An overview[J]. The Journal of classroom interaction, 1981, 17(1): 2-15.
21. Skinner E A, Belmont M J. Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year[J]. Journal of educational psychology, 1993, 85(4): 571-581.
22. Connell J P, Wellborn J G. Competence, Autonomy, and Relatedness: A motivational analysis of self-system processes[J]. Journal of Personality and Social Psychology, 1991(65):43-77.
23. Fredricks J A, Blumenfeld P C, Paris A H. School engagement: Potential of the concept, state of the evidence[J]. Review of educational research, 2004, 74(1): 59-109.
24. Pekrun R, Lichtenfeld S, Marsh H W, et al. Achievement emotions and academic performance: Longitudinal models of reciprocal effects[J]. Child development, 2017, 88(5): 1653-1670.
25. Kahu E R, Nelson K. Student engagement in the educational interface: Understanding the mechanisms of student success[J]. Higher education research & development, 2018, 37(1): 58-71.
26. Alyuz N, Okur E, Oktay E, et al. Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1:1 learning scenario?[C]. 24th ACM Conference on User Modeling, Adaptation and Personalization (UMAP). 2016.
27. Aslan S, Mete S E, Okur E, et al. Human expert labeling process (HELP): towards a reliable higher-order user state labeling process and tool to assess student engagement[J]. Educational Technology, 2017: 53-59.
28. Sidney K D, Craig S D, Gholson B, et al. Integrating affect sensors in an intelligent tutoring system[C]. Affective Interactions: The Computer in the Affective Loop Workshop at. 2005:7-13.
29. Kuh G D . The national survey of student engagement: Conceptual and empirical foundations[J]. New Directions for Institutional Research, 2010, 2009(141):5-20.
30. Whitehill J, Serpell Z, Lin Y C, et al. The faces of engagement: Automatic recognition of student engagement from facial expressions[J]. IEEE Transactions on Affective Computing, 2014,(1):86-98.

Author contributions statement

Guanyu Chen is responsible for article writing and revision, as well as communication work.
Guangxin Han is responsible for data calculation and processing, as well as icon creation and modification.
Juan Niu has provided academic research and theoretical support for this study.
Juhou He is the administrator of this project and provides guidance throughout the research process.

Data availability statement

The datasets generated and analyzed during the current study are not publicly available due to ethical restrictions related to identifiable facial data, but are available from the corresponding author upon reasonable request and subject to approval by the institutional ethics committee.

Additional information

Competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

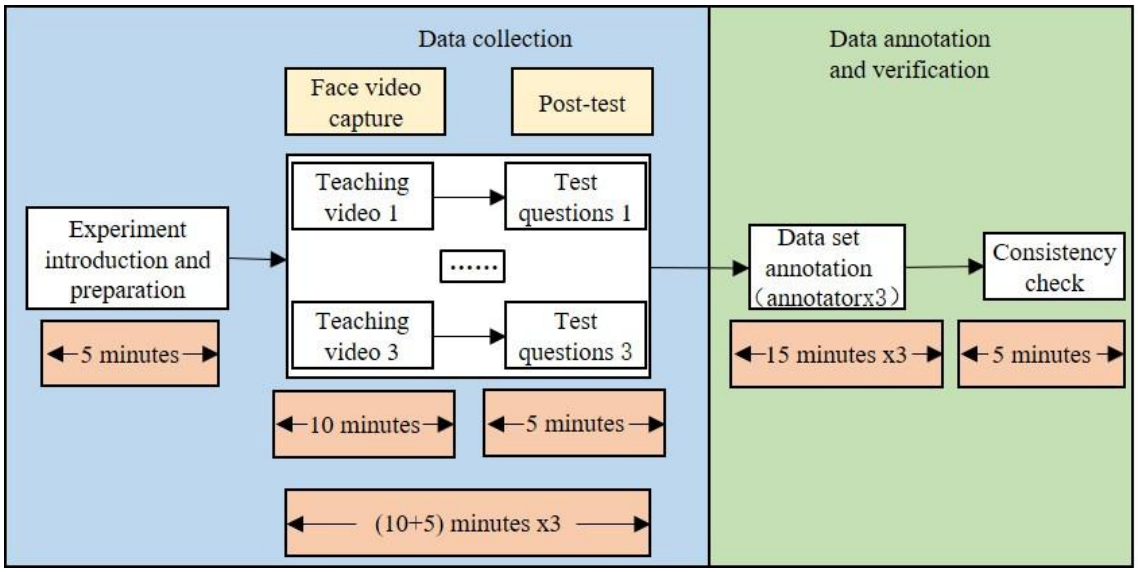

Figure 1. Flowchart for constructing a dataset of emotional engagement in online learning

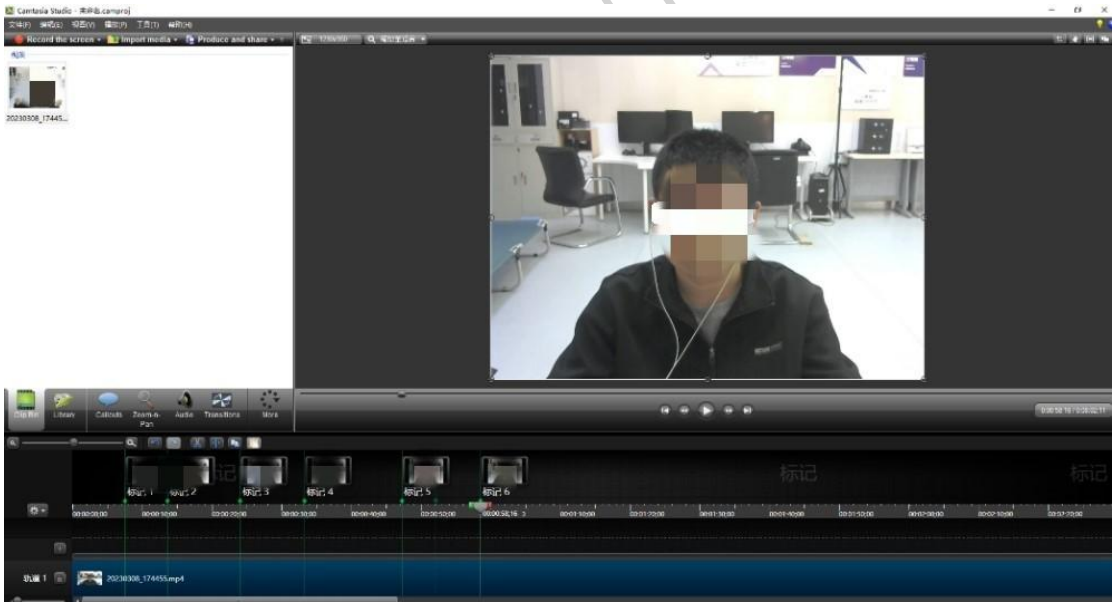

Figure 2. Video trimming using Camtasia Studio software.

Figure 3. Video segmentation results.

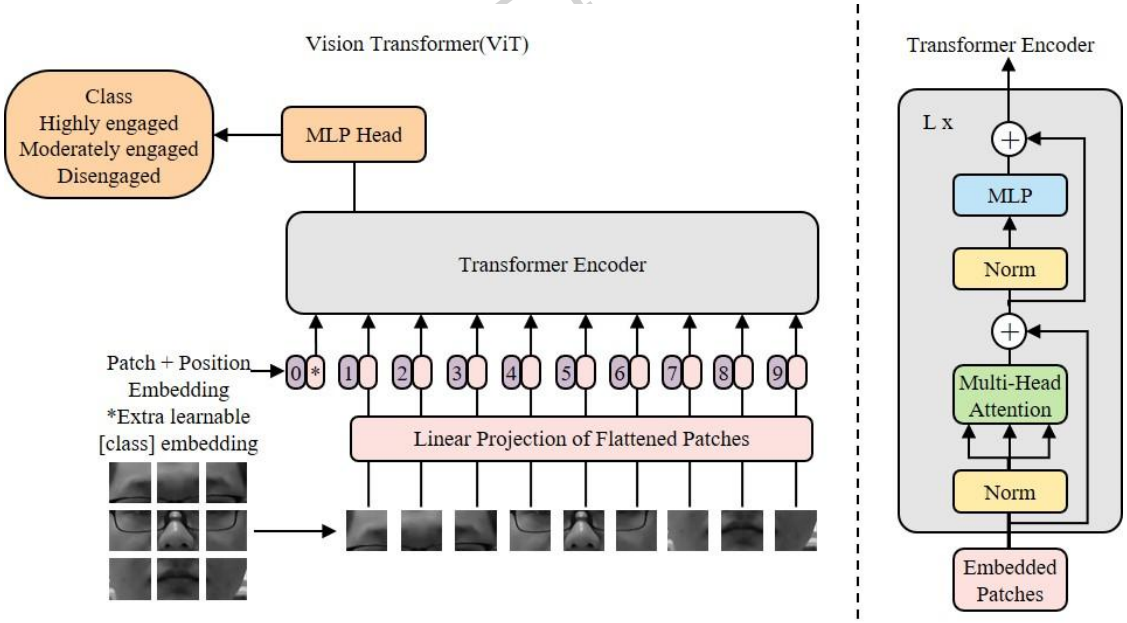

Figure 4. Vision Transformer model architecture diagram.

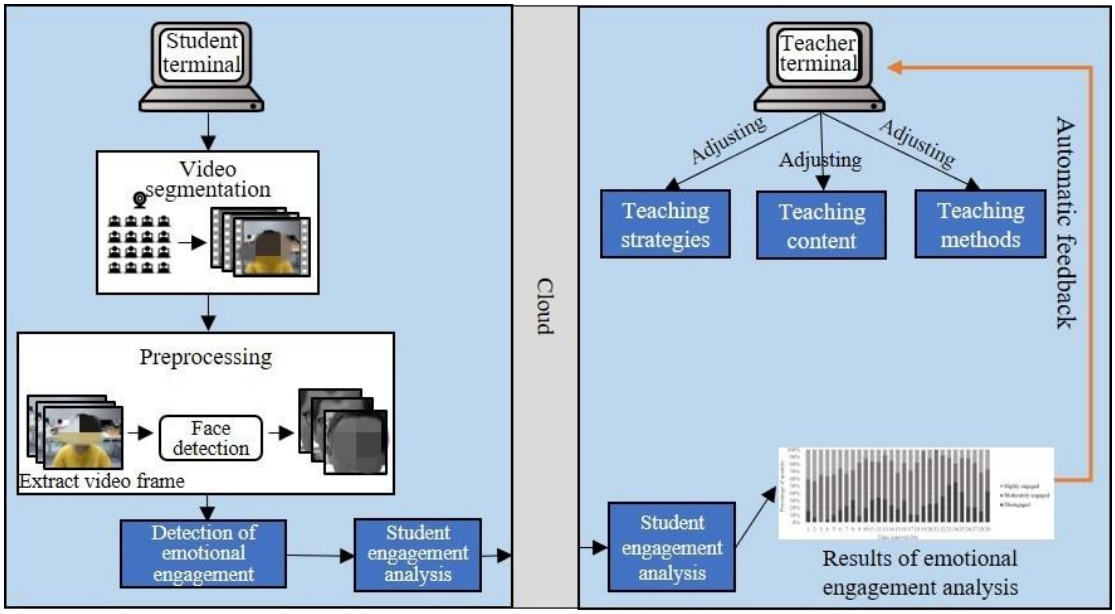

Figure 5. The automatic detection and feedback system.

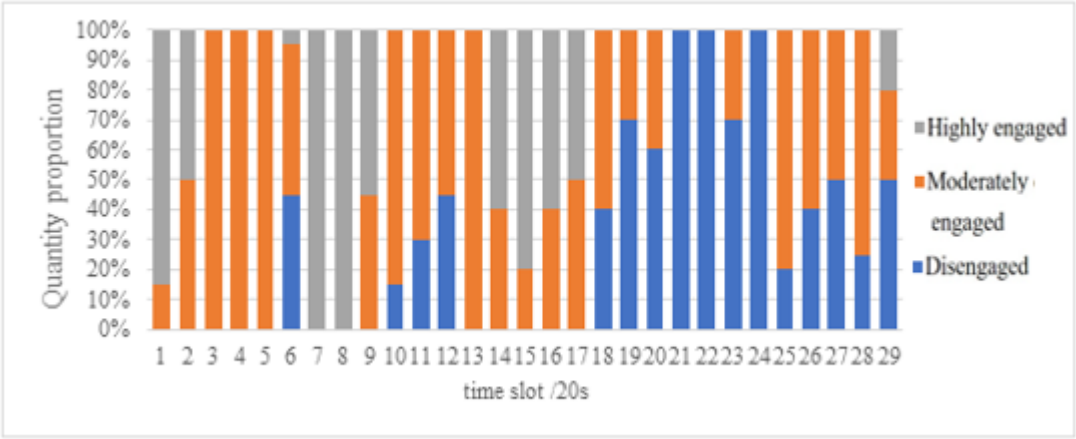

Figure 6. The automatic detection results of all learners.

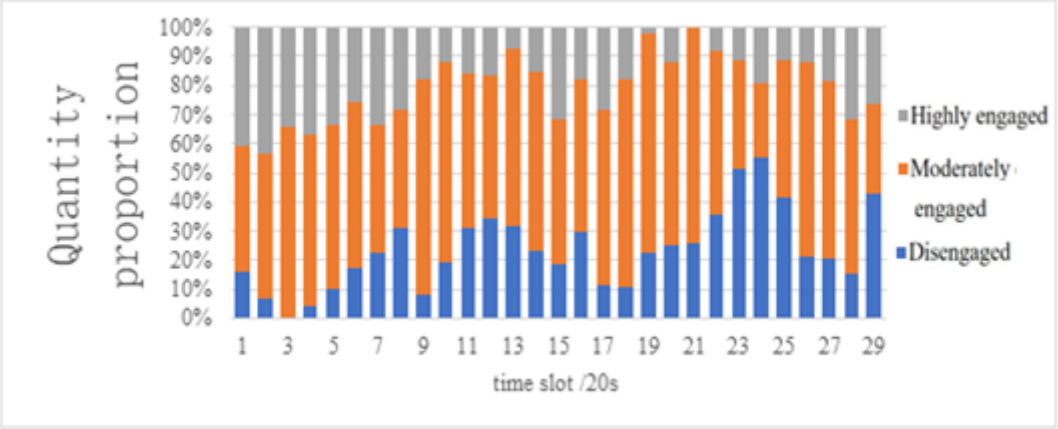

Figure 7. The manual annotation results of all learners.

Learn emotional engagement	Highly engaged	Moderately engaged	Disen gaged
Label	3	2	1
Number of pictures	16515	38468	16202
The proportion of the number of pictures	23.20%	54.04%	22.76 %

Table 1. Distribution of the number of images for three levels of emotional engagement.

Class	Head features	Eye features	Facial expression features
Highly engaged	Head upright or inclined forward	Staring at the screen, eyes unconsciously widening, increased distance between upper and lower eyelids	Surprise, joy, focus, enthusiasm, and other positive expressions.
Moderately engaged	Head generally upright or slightly tilted to the left or right	Line of sight positioned within the screen area, eyes open normally, no change in the distance between upper and lower eyelids	Calm, neutral and other neutral expressions
Disengaged	Head not upright and significant tilt to the left or right	The line of sight is positioned at the edge of the screen area or outside the screen area, eyes slightly closed or even completely closed, and the distance between the upper and lower eyelids decreases	Bored, tired, indifferent, and other negative expressions

Table 2. Coding scheme for learning emotional engagement in online learning.

Fold Number	Accuracy(%)	Precision(%)	Recall(%)
Fold 1	88.75	86.56	85.21
Fold 2	94.51	85.58	94.70
Fold 3	92.32	93.66	93.32
Fold 4	90.99	91.01	87.12
Fold 5	86.56	92.08	86.82
Mean	90.62	89.78	89.43
Std Dev	3.09	3.53	4.27

Table 3. K-fold Cross-Validation by Subject Fold Results Summary(simplified data)

Course	Cronbach's α
Psychology of Love	0.89
Innovative Thinking	0.82
Marxism	0.77

Table 4. The Cronbach's α values for each quiz

	Accuracy	Macro-Recall	Macro-Precision	Macro-F1
Vision Transformer	91.79 ± 1.54	91.48 ± 1.72	93.04 ± 1.46	91.99 ± 1.55
Vision Transformer + Transfer Learning	93.82 ± 1.20	93.78 ± 1.04	94.26 ± 0.80	93.95 ± 1.12

Table 5. Comparison of experimental results (%) before and after Transfer Learning.

Class	Gabor+SVM	Decision Tree	ResNet+TCN	LDP-KPCA-DBN	Vision Transformer +Transfer Learning
Highly engaged	77.52 ± 1.61	89.22 ± 1.13	89.93 ± 1.72	95.52 ± 1.68	95.97 ± 1.01
Moderately engaged	73.65 ± 1.82	73.17 ± 1.46	82.24 ± 1.16	85.83 ± 1.41	92.46 ± 0.79
Disengaged	82.19 ± 1.81	83.15 ± 1.72	84.02 ± 1.19	83.39 ± 1.02	93.03 ± 1.29
Mean	77.78 ± 1.52	81.84 ± 0.94	85.40 ± 1.29	88.25 ± 1.13	93.82 ± 1.20

Table 6. Classification results (%) of emotional engagement.

Course ID		Annotator1	Annotator2	Annotator3	Average of manually annotated results	Automatic detection	Test score
1	Mean	2.782	2.224	2.565	2.523	2.470	8.100
	SD	0.323	0.329	0.410	0.377	0.323	1.370
	N	40	40	40	120	40	40
2	Mean	2.10	1.98	1.89	1.986	1.931	6.300
	SD	0.201	0.216	0.260	0.226	0.245	0.823
	N	40	40	40	120	40	40
3	Mean	1.73	1.80	1.63	1.716	1.651	4.400
	SD	0.122	0.174	0.158	0.175	0.121	0.516
	N	40	40	40	120	40	40

Table 7. Statistical analysis of emotional engagement and learning outcomes.

Course ID		Manual annotation	Automatic detection	Test score
1	Manual annotation	1		
	Automatic detection	0.869**	1	
	Test score	0.799**	0.860**	1
2	Manual annotation	1		
	Automatic detection	0.876**	1	
	Test score	0.657*	0.664*	1
3	Manual annotation	1		
	Automatic detection	0.887**	1	
	Test score	0.636*	0.707*	1

Note. N = 40, *$p < 0.05$, **$p < 0.01$, **$p < 0.001$

Table 8. The Pearson correlation analysis between test score, automatic detection and manual annotation.