

Design and Implementation of a Deep Learning-Based “Smart Classroom Manager” System

Jing Chen¹, Junyi Liang²

¹ Guangxi Vocational and Technical Institute of Industry, Nanning, Guangxi, China

² Beihai Vocational College, Beihai, Guangxi, China

Abstract: Verifying whether learners are actually engaged with online course resources is challenging. Ineffective learning behaviors are common. These include dropping out midway and letting videos play without watching. The “Smart Classroom Manager” is a deep learning system with three layers: perception, analysis and application. The recognition module is a convolutional neural network for facial expressions. From its outputs, a learner psychological model is derived in the two-dimensional valence–arousal space, and the results are mapped to four learning emotion labels: focus, confusion, joy, and fatigue. These labels feed three functions: facial attendance tracking, early warning of abnormal learning states, and learning behavior profiling. Online classrooms impose strict latency limits, and two engineering choices address this constraint. Facial landmark prediction shrinks the region the detector has to search, and edge computing takes part of the inference load off the server. In practice, the system lets online teaching platforms watch learning states as they happen, and the data it collects can support ongoing reforms in teaching evaluation.

Keywords: Artificial Intelligence; Deep Learning; Facial Expression Recognition; Smart Classroom; System Design.

1. Introduction

Governments have moved AI in education from pilot projects to formal policy. In the United States, the Department of Education published *Designing for Education with Artificial Intelligence: An Essential Guide for Developers* in 2024, and an April 2025 executive order created the White House Task Force on Artificial Intelligence Education to coordinate federal efforts. China has moved in the same direction. The Outline of the Plan for Building a Strong Educational Nation (2024–2035), issued in 2025, calls for a national education digitisation strategy, AI-driven educational transformation, and evaluation and decision-making systems built on big data and AI. The “AI+Education” Action Plan,” released in 2026 by the Ministry of Education and four other departments, goes further: it sets out an integrated framework spanning talent development, integrated applications, infrastructure, and the development ecosystem, with the aim of embedding AI in every element, process, and scenario of education. Against this policy backdrop, using AI to monitor the learning process and strengthen teaching evaluation is no longer an experimental option—it is becoming a practical requirement of educational digital transformation.

In recent years, the scale of online course development—represented by first-class courses, “golden courses,” and massive open online courses (MOOCs)—has continued to expand, and online learning has become an integral part of higher education and vocational education. However, behavioral data recorded by online platforms, such as viewing duration and click counts, struggle to reflect learners’ actual learning status. Instances of ineffective learning, such as “course-hopping” and “course-farming,” occur frequently, and assessment methods that rely solely on platform log data carry the risk of evaluation distortion. How to obtain data on learners’ actual states during the course is an urgent issue that online teaching platforms must address.

Facial expressions serve as important outward cues to a learner’s state of engagement. Deep learning-based face

detection and facial expression recognition technologies can capture and analyze learners’ facial expressions in real time using standard webcams, providing online teaching platforms with process-based data beyond behavioral logs. This paper designs a deep learning-based “Smart Classroom Manager” system that integrates facial expression recognition models with psychological models of facial expressions to enable real-time monitoring of learning states in online classrooms, early warning of anomalies, and the generation of learning profiles, thereby offering a viable technical solution for monitoring the quality of online teaching.

2. Literature Review

2.1. Research on Learning State Recognition

Numerous studies have explored the use of artificial intelligence technologies to analyze learner states. Whitehill collected students’ facial images via standard webcams, used machine learning methods to automatically recognize student facial expressions, and classified learning states into four levels based on engagement, thereby validating the feasibility of using facial expressions for the automatic assessment of learning engagement [1]. D’Mello and Graesser [2] observed changes in learners’ emotions using an intelligent tutoring system and noted that emotional experiences during the learning process are context-specific; the classic set of basic emotions—does not fully cover learning scenarios.

2.2. Research on Deep Learning-Based Classroom Behavior Detection

Advances in deep learning-based object detection technology have driven research on classroom behavior recognition. Zhang Hao [3] constructed a dataset covering six typical classroom behaviors—writing, reading, looking up, turning one’s head, raising one’s hand, and standing—and proposed a pyramid-attention deformable detection model. By employing an adaptive hybrid backbone network and a context-guided feature pyramid, the model improved

detection accuracy in densely occluded scenes and achieved lightweight deployment through a knowledge distillation scheme. Gao Yawen's research addressed two challenges in classroom settings—occlusion in the back rows and small-object detection—with an improved YOLOv8 algorithm [4]. In the neck network, Gao combined clustered shuffled convolutions with a coordinate attention mechanism. To limit the loss of fine-grained information, spatial depth transformation convolutions were then introduced, and a small-object detection layer was added to the detection head to improve the accuracy and generalizability of classroom behavior detection.

2.3. Smart Classroom System Integration and Ethical Research

Research on system integration has made some headway. Wang Jianfei built a YOLOv11-based smart classroom management platform [5]. The platform identifies classroom behaviors—using cell phones, raising hands, sleeping—in real time; matches student identities through facial recognition; triggers voice and pop-up warnings when abnormal behaviors occur; and incorporates speech-to-text transcription and random roll call. Luo approached the issue from the ethics side, analyzing application cases of intelligent classroom feedback systems [6]. They found that algorithmic black boxes and algorithmic discrimination can lead to data distrust, privacy violations, and disordered power relations between teachers and students, and they proposed mitigating these risks through informed knowledge of algorithmic rules, data ethics protocols, and human-machine collaboration mechanisms. Taken together, these studies offer developers of such systems a twofold reference. On the engineering side, the combination of behavior recognition, identity matching, and instant warning shows that multi-model coordination works in practice. On the ethics side, the boundaries of data collection, algorithmic explainability, and the protection of

teacher and student rights need to be built into system design from the outset, not patched in afterward.

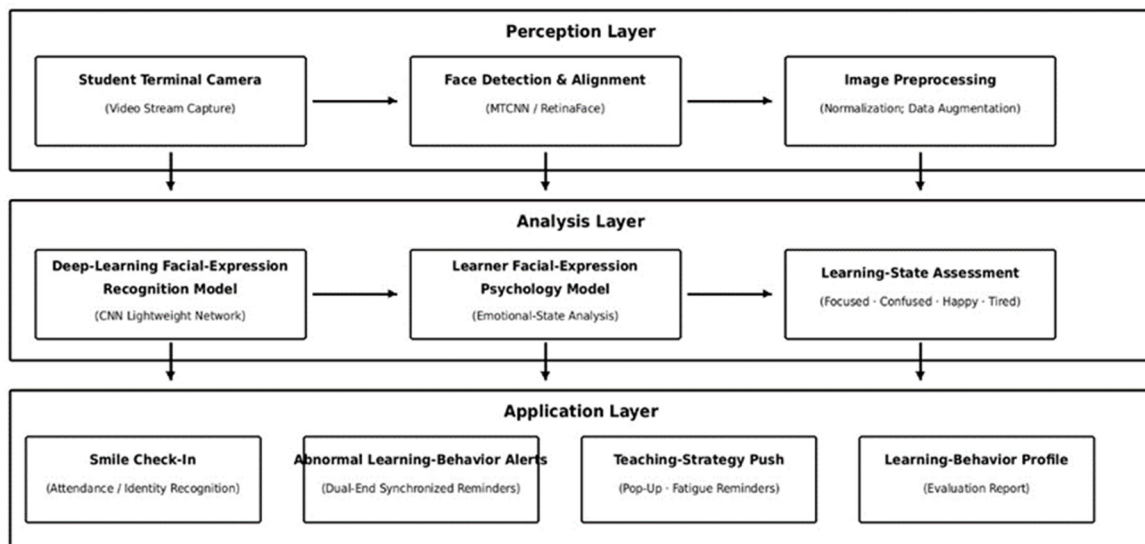
2.4. Summary

Existing research has concentrated on group behavior detection in physical classrooms [3,4,5]. Facial expression recognition in online classrooms—where terminal devices vary widely, lighting is complex, and students appear as individual close-ups—has received far less attention. Most current systems stop at state recognition [4,5,6]; how recognition results should feed into instructional interventions is still an open question [2]. On top of this, the collection of emotional data raises compliance issues that system design cannot ignore. The "Smart Classroom Manager" proposed in this paper responds to these gaps by combining deep learning-based facial expression recognition, a psychological model of facial expressions, and instructional strategy recommendations, so that online learning states can be perceived and acted on as they occur.

3. Overall System Design

To address the need for learning process monitoring in online classrooms, the system builds upon existing online teaching platforms based on learning behavior logs by adding learner identity verification, facial expression recognition, status analysis, and evaluation feedback functions. The design objectives include: First, to enable automatic learner facial check-in and identity verification using standard webcams; second, to perform real-time classification of learners' facial expressions and map them to emotional tags to identify abnormal learning states; third, to push alerts to both students and instructors based on the recognition results, and to generate learning process records that provide formative evidence for course assessment.

3.1. Overall Architecture



Teaching strategies are fed back to the student and teacher terminals in real time, forming a closed loop of "perception-analysis-intervention."

Figure 1. Overall System Architecture of the "Smart Classroom Manager"

The system adopts a three-tier architecture consisting of the "perception layer—analysis layer—application layer," as shown in Figure 1. The perception layer captures video streams via cameras on student terminals to perform face

detection, alignment, and image preprocessing; the analysis layer deploys a convolutional neural network (CNN) for facial expression recognition and a psychological model of learner expressions, outputting learning emotion labels and

learning status assessment results; The application layer provides four types of services—face-based attendance tracking, alerts for abnormal learning states, teaching strategy recommendations, and learning behavior profiles—and integrates with the online teaching platform to form a closed-loop “perception–analysis–intervention” processing cycle. The layers are decoupled via standard interfaces to facilitate functional expansion and platform migration.

Running deep learning inference demands substantial compute, and when dozens of classroom video streams arrive at once, a single server quickly becomes the bottleneck. Two design decisions keep the load manageable. The first concerns the detection pipeline itself: a per-learner facial model predicts where the landmarks will fall, so the detector only searches a small region of each frame instead of the full image, bringing down per-frame latency. The second concerns where computation happens. Inference is split between the cloud and the edge—edge devices take over part of the workload, while the central server handles just model training and data aggregation. With this division of labor, the system sustains real-time performance even in multi-stream online classrooms.

3.2. Technology Selection

The system is written in Python, with TensorFlow/Keras as the deep learning framework. The frontend is web-based, so any platform with a browser can access it. Face detection came down to a choice between MTCNN and RetinaFace. MTCNN's cascaded architecture is lightweight and runs fast on end-user devices. RetinaFace is more accurate, but the computational cost is considerably higher. In online classrooms the camera typically frames a single student at close range, and under such conditions responsiveness counts for more than a marginal gain in accuracy. MTCNN was therefore adopted for detection and alignment. For expression recognition, three networks (VGG, ResNet, and MobileNet) were benchmarked on accuracy and inference time. The backbone finally selected was a lightweight network that balances accuracy against model size, a compromise that suits deployment on edge devices.

3.3. Functional Modules Design

The system's functionality targets three pain points in online classroom management. Attendance still relies on manual roll call. Learning states cannot be sensed. Process data sits scattered across platforms and is hard to put to use. Three modules follow from this: facial attendance, early warning of abnormal learning states, and learning behavior profiling. They connect sequentially in the data flow—the attendance module supplies the identity baseline, the warning module generates state events, and the profile module aggregates both.

The facial attendance module answers a basic question: is the student actually there? When a learner joins a class, the client captures video frames and uploads them. The server runs face detection and feature matching. A successful match counts as valid check-in, and the arrival time goes into the attendance sheet. Identity is then re-verified by sampling at fixed intervals, and cumulative online duration is recorded. Peak-time check-ins could otherwise queue up, so verification requests go through an asynchronous queue. Results come back within seconds, and normal entry into the classroom is not affected.

The early-warning module is the core of the system. It

operates within the online periods established by facial attendance. It tracks each learner's expression sequence and maps expressions to learning states via the psychological model described in Section 3.4. Warning judgment and intervention execution are kept apart in the design. The judgment layer tallies the distribution of states over a sliding time window. Three conditions carry separate trigger thresholds: prolonged absence, sustained fatigue, and abnormal emotional changes. This way, a single misread frame does not set off repeated interruptions. Once a threshold is crossed, the execution layer pops up a reminder on the student side and pushes the warning event to the teacher's console. After confirmation, the teacher can launch a pop-up quiz or a fatigue reminder. Warning intensity is graded. Minor fluctuations are only logged, never pushed, so the teaching rhythm stays undisturbed.

The learning behavior profile module serves the assessment stage. Emotional sequences and warning events are grouped by learner. They are then linked with access logs and assignment records exported from the online teaching platform to form an individual learning-process profile. At the class level, attendance rates, state distributions, and warning frequencies are aggregated weekly into an analysis report. Access to profile data is restricted to the course instructor and teaching administrators. Raw facial images are never stored. Only features and state labels are retained, which keeps privacy risk under control.

4. Implementation of Key Technologies

4.1. Deep Learning-Based Facial Expression Recognition

The recognition model is a convolutional neural network built from stacked convolutional, pooling, and fully connected layers. The convolution kernels pick up local texture and contour cues from the face, and as these features pass through successive layers they are abstracted into higher-level representations; a Softmax layer finally outputs the probability for each emotion category. Training proceeds in two phases. The model is first pre-trained on public datasets such as FER2013, which gives it general feature extraction ability for facial expressions, and it is then fine-tuned through transfer learning on expression samples collected and annotated in actual online classrooms, so that it adapts to real teaching conditions rather than laboratory data. Real classrooms are in fact messier than benchmark datasets: learners use different devices, camera angles vary, and lighting is often poor. To cope with this, the training set is expanded through data augmentation—histogram equalization, normalization, random cropping, and rotation—making the model more tolerant of lighting changes, pose deviations, and partial occlusions. Real-time inference imposes a further constraint. Instead of scanning the whole frame, the system uses predicted facial landmarks to mark out the expression-bearing regions, namely the eyebrows, eyes, and mouth. Recognition then runs only on these regions of interest, which cuts the processing time per frame substantially.

4.2. Facial Expression–Emotion Mapping Model

The basic emotion categories output by the facial expression recognition model cannot be used directly for learning status determination; a psychological mapping

model tailored to the learning context must be established. The model developed here is based on the Circumplex Model of Affect [7], which represents emotions as a two-dimensional continuous space composed of valence and arousal: valence describes the degree of pleasantness of an emotion, while arousal describes the level of physiological and psychological activation. Pekrun's Control-Value Theory of Academic Emotions [8] provides a further basis for label selection. This theory holds that academic emotions are shaped by learners' sense of control over learning tasks and by the value they attach to those tasks. It also clarifies the dynamic relationship among academic emotions, learning engagement, and academic achievement, giving the choice of learning emotion labels a grounding in educational psychology.

Based on the above theories, the system maps facial expression recognition results onto the valence-arousal two-

dimensional space. It then combines the duration of the expression to locate the learning emotion within this space and ultimately determines the learning state. The mapping process is illustrated in Figure 2. Specifically, high-arousal, positive-valence expressions such as happiness and surprise are mapped to positive learning emotions such as pleasure and concentration; neutral expressions are mapped to either focus or calmness based on their duration; negative-valence expressions such as disgust and sadness are mapped to confusion or anxiety; and facial features such as drooping eyelids and yawning are mapped to a state of fatigue in the low-arousal, negative-valence region. The system sets a duration threshold such that an alert is triggered only when a particular emotion persists beyond the threshold time, thereby avoiding misjudgments caused by fleeting expressions.

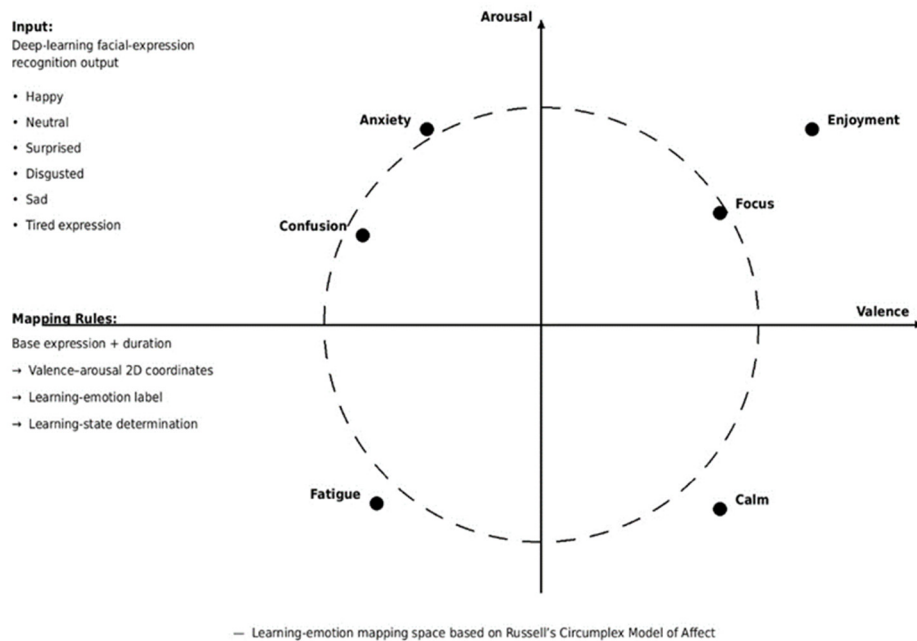


Figure 2. Schematic of the Expression-Emotion Mapping Model

4.3. Learning State Analysis and Profile Generation

Based on the sequence of emotion labels, the system calculates the proportion of each emotion type using a sliding time window, computes an attention index, and identifies sudden changes and sustained abnormal intervals in learning states; it aligns the sequence of emotional changes with the course timeline to analyze emotional response characteristics during different instructional segments. The attention index is calculated by weighting the duration of the two positive states—attention and calmness—within the window, while the durations of confusion and fatigue are subtracted; a lower score indicates a poorer learning state. Individual profiles include emotional distribution, attention curves, and records of early-warning events; class reports summarize overall emotional trends and common issues, providing teachers with guidance for adjusting the pace of instruction. Profile data retains only feature vectors and statistical results; raw video frames are deleted immediately after analysis to reduce the risk of privacy breaches.

5. Conclusion

Monitoring what students are actually doing in an online classroom is hard to get right. This paper has presented the "Smart Classroom Manager," a deep learning system built to tackle that problem. The system follows a three-layer architecture. At its center are a CNN-based facial expression recognition model and a psychological model of expressions grounded in the valence-arousal space. Together they support three functions: facial attendance tracking, early warning of abnormal learning states, and learning behavior profiling. Online teaching platforms thus gain a way to observe learning status in real time, which helps correct the assessment distortions caused by dropouts and course-crashing.

The system remains a work in progress. Technically, recognition accuracy still degrades under difficult lighting and occlusion, and real-time processing of many concurrent video channels needs to be faster. On the application side, the rules around student facial and emotional data—who authorizes collection, how it is stored, where its use ends—must be strict. System outputs should serve as teaching references, not as grounds for assessment. Follow-up work

will proceed along three lines: fusing multimodal information, lightening the model, and tightening data compliance mechanisms.

Acknowledgments

This study was supported by a research project from Guangxi Vocational and Technical Institute of Industry: Research and Implementation of a Deep Learning-Based “Smart Classroom Manager” System (Grant No.: GYKY2020002A).

References

- [1] Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1), 86–98.
- [2] D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145–157. <https://doi.org/10.1016/j.learninstruc.2011.10.001>
- [3] Zhang Hao. (2026). A Study on Deep Learning-Based Methods for Student Behavior Recognition in Smart Classrooms (Master's Thesis, Qilu University of Technology).
- [4] Gao Yawen, Huang Xianming, Liang Ainan, Yu Hongxiang, & Si Pengju. (2026). A Study on Algorithm Optimization for Monitoring Student Classroom Behavior. *Journal of Hunan University of Science and Technology*, 40(6), 56–63. <https://gfgfy7789d7bbe5f747dbsbpxvpbc5699k69f0ficg.res.gxlib.org.cn/10.20271/j.cnki.1673-9833.2026.6008>.
- [5] Wang Jianfei, Gao Guiyuan, & Wang Haiyan. (2025). Design and Application Research of a Deep Learning-Based Smart Classroom Management System. *Journal of Changchun University*, 35(12), 73–77.
- [6] Luo Jianghua, Wang Lin, & Liu Lu. (2022). Ethical Dilemmas and Risk Mitigation in AI-Empowered Classroom Feedback. *Research on Modern Distance Education*, 34(2), 29–36.
- [7] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.
- [8] Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4), 315–341.