EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

A Study on Real-time Early Warning and Adaptive Intervention for Online Learning Burnout Using Multimodal Temporal Feature Fusion

Kun Liu[1], Xiaoxiao Gu[1] , Jingxia Chen[1,*] and Wenjuan Shao[1]

[1]The University of Applied Science and Technology, Beijing Union University, Unit A1, Building 6, Beiyuan Residential Compound, Chaoyang District, Beijing

Abstract

INTRODUCTION: The lack of physical presence in online learning makes it difficult for teachers to perceive students' cognitive and emotional states in real time, with learning burnout being a particularly prominent issue. Existing research primarily relies on single-modality data or lagged learning analysis, making it difficult to achieve precise and timely burnout early warning and intervention.
OBJECTIVES: This paper proposes an online learning burnout early warning model based on multi-modal temporal feature fusion and designs a hierarchical adaptive intervention mechanism.
METHODS: First, utilizing lightweight convolutional neural networks and temporal encoders, facial expression features and body posture features are extracted in real-time from the video stream, respectively. Second, a multi-modal temporal fusion module based on attention mechanisms is designed to model the coordinated temporal evolution of facial expressions and body postures, enabling precise identification of fatigue states such as "confusion","fatigued" and "bored". Finally, a hybrid decision model combining rule-based and reinforcement learning approaches is developed. This model dynamically triggers personalized intervention strategies—such as content adjustment, suggested breaks, and interactive Q&A sessions—based on the type, intensity, and duration of fatigue.
RESULTS: Experiments conducted on a self-built authentic online learning dataset demonstrate that this system achieves an F1 score of 0.91 in burnout state recognition, which significantly outperforms unimodal approaches. Furthermore, user studies validate the effectiveness and high acceptance of its intervention mechanism.
CONCLUSION: This study offers a viable technical and practical approach for enabling precise, timely detection and adaptive intervention of burnout in online learning environments.

Keywords: Online Learning Burnout, Multi-modal Fusion, Affective Computing, Real-time Early Warning, Adaptive Intervention

Received on 23 October 2025, accepted on 06 May 2026, published on 26 May 2026

doi: 10.4108/eetsis.10698

*Corresponding author. Email: yykjtjingxia@buu.edu.cn

1. Introduction

The proliferation of online education has dismantled spatial and temporal barriers to learning[1,2]. However, it has also limited teachers' ability to observe students' nonverbal cues—such as facial expressions and body language—which are readily apparent in traditional classrooms. This disconnect can lead students into a state of learning burnout that goes undetected, ultimately resulting in decreased learning efficiency and increased dropout rates[3-5]. Consequently, there is an urgent need in educational technology to build intelligent learning systems that can

simulate teacher-like insight and provide timely, proactive support[6].

Scholars domestically and internationally have conducted preliminary explorations in learning state recognition. In single-modal analysis, Wang et al.[7] proposed an LD-identify network for action recognition based on passive RFID. By extracting phase and signal strength characteristics from two RFID tags, they combined a CNN model to recognize three head movements with over 95.5% accuracy. However, their method only models limited actions and omits common learning behaviors like writing. It also relies heavily on precise tag placement, making it susceptible to occlusion in real-world scenarios, which compromises its robustness. In multi-modal fusion, Cao et al.[8] constructed a model based on the ResNet architecture. By integrating facial images, EEG data, and learning logs, they achieved 87% accuracy in predicting engagement, confirming multi-modal superiority. However, their approach was designed solely for MOOC scenarios and does not generalize to other settings like live classrooms. Moreover, its reliance on specialized EEG equipment hinders seamless deployment. Jiang et al.[9] constructed a three-dimensional framework integrating audio, image, and text data. Their SMAT coding system analyzes behavioral frequency and temporal shifts in smart classrooms. However, their framework does not address fusion bias caused by inter-modal data quality differences, while also offering insufficient exploration of the correlation between learning behaviors and academic outcomes. Wang and Zheng[10] systematically reviewed multi-modal data types and integration strategies in intelligent education. However, the review does not propose actionable plans for optimizing model integration into specific educational scenarios. Furthermore, its discussion of ethical risks (e.g., privacy concerns) is brief and lacks systematic solutions. Wang et al.[11] constructed a multi-modal fusion model for online learning integrating behavioral, emotional, and cognitive dimensions. Utilizing IoT technology for automated data collection, their model achieved an accuracy of no less than 76%. However, its feature fusion does not fully leverage temporal correlations, which causes delayed responses to dynamic state changes. Furthermore, it neglects variations in modality importance across different task types, resulting in inadequate adaptability in scenarios like skill training.

Current research[12-14] exhibits three major shortcomings. The first is perceptual bias. Reliance on single-modal or discontinuous data hinders comprehensive and consistent capture of complex fatigue states. The second is static modeling. Most approaches adopt segmented classification, overlooking that burnout is a dynamic, cumulative, temporal process. The last is lack of closed-loop Intervention. Most studies stop at state recognition, failing to form a complete "perception-decision-intervention-evaluation" closed loop integrated with effective teaching strategies.

To address these limitations, this paper proposes an end-to-end online learning burnout early warning and intervention system. Our core innovation lies in a novel attention-based multi-modal temporal fusion model, which effectively captures the complementary and synergistic relationships between facial expressions and body postures over time, enabling fine-grained, dynamic fatigue recognition. We also design a hierarchical adaptive intervention mechanism that combines rule-based rapid responses with reinforcement learning-based long-term optimization, making interventions more intelligent and user-friendly. Finally, we have developed and validated a complete real-time system prototype in authentic learning scenarios, demonstrating its end-to-end effectiveness and positive user experience.

2. Introduction to Related Work and Technology

2.1 Related work

2.1.1 Research on Learning Burnout Detection

The assessment of learning burnout represents a significant research focus in educational technology[15]. Current methods primarily leverage physiological signals, behavioral analysis, and multi-modal fusion. Physiological approaches, such as electroencephalogram(EEG) and heart rate variability(HRV), offer objective measures of cognitive load but require specialized equipment and can be obtrusive. Behavioral analysis infers fatigue states from learning patterns like mouse click frequency and page dwell time; however, these metrics are often influenced by individual differences[16,17]. While multi-modal fusion techniques, which integrate diverse data sources, enable a more comprehensive assessment, they still face challenges in effective temporal modeling and fusion strategy design.

2.1.2 Multi-modal Temporal Fusion Techniques

Effective multi-modal temporal fusion [18,19] is crucial for modeling dynamic states like fatigue. Fusion strategies are commonly categorized into three types: The first is early fusion, which directly concatenates raw or low-level features. The second is late fusion, which aggregates decisions from models trained on separate modalities. The third is intermediate fusion, which integrates modalities at the feature representation level. Among these, attention mechanisms have shown remarkable performance by dynamically weighting the importance of different modalities. Nonetheless, a key limitation of many existing methods is their focus on static fusion, and they lack the sophisticated temporal modeling capabilities needed to capture the evolving patterns of fatigue states.

2.1.3 Adaptive Intervention Strategies

Adaptive intervention is a core function of intelligent educational systems. Current strategies predominantly utilize rule-based systems, machine learning, or reinforcement learning. Rule-based systems enable rapid responses but lack personalization. Machine learning approaches can learn personalized strategies but require large amounts of labeled data. Reinforcement learning

methods optimize strategies through environmental interaction, but their training complexity and convergence instability pose challenges. Consequently, balancing immediate responsiveness with long-term optimization through the integration of these diverse methods remains a key research focus.

2.2. Related Technologies

2.2.1 Facial Expression Recognition

This paper employs the lightweight MobileViT network as the backbone feature extractor. This model combines the local feature extraction capabilities of CNNs with the global context modeling capabilities of Vision Transformers. It maintains high accuracy while significantly reducing computational overhead, making it suitable for real-time operation on standard PCs or edge devices. The model outputs a continuous affective valence vector rather than discrete classification labels, thereby capturing the nuanced emotional shifts from positive to negative with greater precision[20,21].

2.2.2 Body Posture Estimation and Behavior Analysis

Real-time 2D human key-point detection using the Media Pipe Pose solution. This solution is lightweight and highly efficient. Based on detected key points (such as the nose, shoulders, and hips), it calculates a set of pose features describing the learned behavior, including:

- Head tilt angle: Indicates whether the head is “on the desk.”
- Trunk Stability: Calculated by analyzing the variance in movement of key shoulder points, this metric indicates whether the body exhibits excessive “swaying”.
- Arm Range of Motion: Indicates whether the student is engaged in activities unrelated to learning.

2.2.3 Deep Reinforcement Learning

This paper employs Deep Q-Networks to optimize long-term intervention strategies. The agent's state is the sequence of historical student burnout characteristics. The action space comprises various intervention strategies. The reward function comprehensively considers the degree of burnout state alleviation, progress in completing learning tasks, and student feedback on interventions—such as “turning off reminders” being regarded as negative feedback. Through continuous interaction with the environment (i.e., the student), DQN learns to select optimal intervention actions with a long-term perspective.

3. Framework Introduction

This paper introduces a closed-loop teaching assistance system that automates the management of student learning burnout by integrating real-time perception, intelligent early warning, and adaptive intervention. Built upon the "End-Edge-Cloud" collaborative computing paradigm, our design achieves holistic process management while maintaining real-time operation and protecting user privacy. The system architecture is shown in Figure 1, which consists of three core modules. Seamless integration and secure data flow between these modules are achieved through well-defined interfaces.

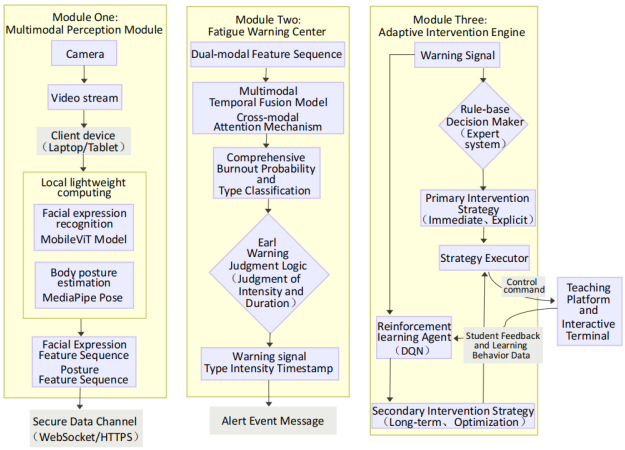

Figure 1. System Architecture

3.1 Multi-modal Perception Module

The Multi-modal Perception Module is deployed on student client devices, such as personal laptops and tablets. It is responsible for non-invasive data collection and lightweight front-end intelligence. The module captures video streams and concurrently executes facial expression recognition and pose estimation models to generate real-time temporal sequences of facial expression and body pose features. The workflow is as follows:

(i) Data Capture

The device camera is activated to continuously capture video streams at a configurable frame rate (e.g., 15 fps).

(ii)Parallel Model Inference

Facial Expression Analysis: A lightweight MobileViT model performs real-time inference on the facial region of each frame. Instead of a simple classification label, the model outputs a high-dimensional feature vector that encapsulates both global and local semantic information of the expression, providing a rich basis for subsequent temporal modeling.

Body Pose Analysis: The MediaPipe Pose solution runs concurrently to detect and extract 2D coordinates of body keypoints in real-time. Based on these coordinates, a set of postural features—such as head tilt angle, trunk stability, and arm movement frequency—is calculated locally.

(iii)Data Security and Upload

To ensure privacy, original video frames are immediately discarded after processing and never leave the client. Only the anonymized feature sequences and corresponding timestamps are uploaded in real-time to the Burnout Early Warning Center via a secure communication protocol.

3.2 Burnout Early Warning Module

This module is deployed on edge servers or regional cloud servers. Its primary function is to receive the dual-modality feature sequences from client devices and perform joint analysis using the core multi-modal temporal fusion model proposed in this paper. The model outputs a comprehensive burnout probability score along with a specific burnout type classification. An early warning is triggered when the score exceeds a predefined threshold and persists for a specified duration. This edge/cloud deployment strategy reduces network latency, enables rapid responses for regional students, and further minimizes the transmission of sensitive data across public networks. The workflow is as follows:

(i)Feature Reception and Caching

The module receives feature streams from multiple clients and maintains a sliding time-window feature sequence for each student.

(ii)Multi-modal Sequence Fusion

The cached bimodal sequences are fed into the multi-modal temporal fusion model. First, two independent LSTM networks encode the facial expression and pose feature sequences separately, capturing the temporal dynamics within each modality. Then, through cross-modal attention layers, the model learns the interactive relationships between them. For instance, it can learn to assign greater importance to the co-occurrence of a "frown" expression and a "leaning forward" posture when identifying a state of "confusion." Finally, the fused high-level representation is passed to a classification layer, which outputs a probability distribution over four student states: focused, confused, fatigued, and bored.

(iii)Early Warning Decision-Making

The early warning decision-making module continuously monitors this probability distribution. It employs a decision logic based on threshold and duration. For example, a structured fatigue alert is generated only if the combined probability of the fatigued and bored states exceeds a threshold of 0.7 for a continuous one-minute period. This alert signal contains key information such as fatigue type, intensity, and onset time, and is immediately transmitted to the adaptive intervention engine.

3.3 Adaptive Intervention Engine Module

This module is co-deployed with the early warning center on edge servers to ensure low-latency intervention. Its core function is to receive early warning signals and generate the most appropriate personalized intervention strategy through a hybrid decision-making mechanism, thereby closing the "perception-decision-action" loop. Upon receiving a signal, the rule-based decision-maker first triggers a primary intervention based on predefined mappings. Simultaneously, the reinforcement learning (RL) agent evaluates the context and may recommend a more strategic, secondary intervention based on the student's current state and historical interaction data, aiming to maximize long-term learning gains. The workflow is as follows:

(i) Signal Reception and Analysis

Parses incoming signals from the early warning center to determine the specific context and type of fatigue.

(ii)Tiered Decision-Making

Rule-Based Decision Maker (Level 1 Response): This is an expert system based on generative rules.It incorporates explicit IF-THEN rules to handle common, typical fatigue scenarios, achieving millisecond-level response times.

Reinforcement Learning Agent (Secondary Optimization): It takes the current state (fused features, historical intervention effects, user profiles) as input and outputs a long-term value assessment for all possible intervention actions.Its objective is to explore long-term optimal strategies that transcend fixed rules and are more personalized.

(iii)Decision Integration and Execution

The system will compare the outcomes of rule-based decisions with those of RL decisions. Typically, rule-based decisions serve as the default execution option. However, if the RL agent assigns a significantly higher and credible Q-value to an action differing from the rules, the system will adopt this more strategic secondary intervention plan. The strategy executor ultimately translates the selected intervention actions into specific instructions, which are transmitted to the teaching platform and interactive terminals via API calls.

(iv)Feedback Loop and Online Learning

Following intervention implementation, the system continuously monitors students' subsequent responses, converting this data into reward signals. Reward signals are used to update RL agents online, enabling their strategies to continuously evolve and increasingly align with the genuine needs of both the entire student body and individual students.

Through the seamless collaboration of three modules, this framework not only enables real-time, precise detection of learning burnout but also integrates intelligent care into the online learning process via a layered, adaptive intervention

mechanism. Ultimately, it achieves the core objective of enhancing students' learning experience and efficiency.

4. Algorithm Analysis

This section will rigorously define the two core components of this system in terms of mathematical formalism and algorithmic procedures: the learning burnout recognition model and the hierarchical intervention decision mechanism.

4.1 A Time-Series Fusion Early Warning Model Based on Cross-Modal Attention Mechanism

This model maps online learning video streams to a probability distribution of fatigue states, defining it as a temporal classification problem.

4.1.1 Formal Definition of the Problem

Input a video clip of length V, $V=\{I_1,I_2,...,I_T\}$, where I_t represents the t-th frame image. The goal is to learn a function F as: $F(V)=\{\hat{Y}_1,\hat{Y}_2,...,\hat{Y}_T\}$.

Given an input video clip $V=\{I_1,I_2,...,I_T\}$ of length V, where I_t denotes the t-th frame, the goal is to learn a function F that maps the video to a sequence of output states: $F(V)=\{\hat{Y}_1,\hat{Y}_2,...,\hat{Y}_T\}$.

Here, $\hat{y}_t=\left[p_t^{(1)},p_t^{(2)},p_t^{(3)},p_t^{(4)}\right]\in R^4$ is a probability vector, whose elements $p_t^{(1)},p_t^{(2)},p_t^{(3)},p_t^{(4)}$ represent the predicted probabilities for the four states—focused, confused, fatigued, and bored, respectively. It satisfies the condition: $\sum_{i=1}^{4}p_t^{(i)}=1$.

4.1.2 Model Architecture and Mathematical Formulation

(i)Modality-specific Feature Encoding

Facial Expression Feature Extraction: For each frame I_t, we employ a lightweight Mobile ViT model Φ_{face}, pre-trained on the Aff-Wild2 dataset, to extract facial features. Specifically, we take the feature map from the layer preceding the global average pooling layer, flatten it, and obtain a d_f-dimensional feature vector f_t.

$$f_t=Flatten\left(GAP\left(\Phi_{face}(I_t)\right)\right)\in R^{d_f} \quad (1)$$

Among these, GAP denotes Global Average Pooling. Input sequence $\{f_1,f_2,...f_T\}$ into a unidirectional LSTM network to learn the temporal context of facial expressions. At each time step t, the LSTM outputs a hidden state $h_T^f\in R^{d_h}$.

$$h_t^f=LSTM_{face}\left(f_t,h_{t-1}^f\right),\quad H_{face}=\{h_1^f,h_2^f,...,h_T^f\},\ h_t^f\in R^{d_h} \quad (2)$$

Body Pose Feature Extraction: For each I_t frame, the Media Pipe Pose function Ψ extracts N key-point coordinates $k_t\in R^{2N}$, forming a $2N$-dimensional raw pose vector p_t. Calculate a set of engineering feature vectors $e_t\in R^{d_e}$, including head pitch angle, torso stability, etc. Concatenate these engineering features with the original coordinates to form a d_p-dimensional augmented posture feature vector $\tilde{p}_t$, yielding the augmented posture features after concatenation.

$$\tilde{p}_t=[k_t;e_t]\in R^{d_p} \quad (3)$$

Similarly, input the sequence $\{\tilde{p}_1,\tilde{p}_2,...,\tilde{p}_T\}$ into another unidirectional LSTM for encoding.

$$h_t^P=LSTM_{pose}\left(\tilde{p}_t,h_{t-1}^P\right),\quad H_{pose}=\{h_1^P,...,h_T^P\},\ h_t^P=R^{d_h} \quad (4)$$

(ii)Cross-Modal Attention Fusion Mechanism

This represents the core innovation of our model. Rather than simply concatenating H_{face} and H_{pose}, we introduce a cross-modal attention mechanism. This allows the model to dynamically determine when and which elements of the pose stream to attend to, as the expression stream actively queries relevant information from it using a scaled dot-product attention layer.

Attention Weight Calculation: To prioritize the facial modality while incorporating supplementary information from posture, we designate H_{face} as the Query and H_{pose} as both the Key and Value. This configuration enables the facial features to actively retrieve the most relevant contextual cues from the pose sequence.

$$Q=H_{face}W^Q,\quad K=H_{pose}W^K,\quad V=H_{pose}W^V \quad (5)$$

Where $W^Q,W^K,W^V\in R^{d_h\times d_k}$ is the learnable parameter matrix.

Attention Output Calculation: Compute the attention from the expression stream to the pose stream using a scaled dot-product attention mechanism. This operation generates a new sequence $H_{pose2face}$, which represents the most critical portion of the pose information for understanding facial expressions. Computing the attention weight matrix $A\in R^{T\times T}$ and the weighted pose representation $H_{pose2face}$.

$$A=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right),\quad H_{pose2face}=AV \quad (6)$$

Here, $A_{i,j}$ denotes the degree of attention paid by the facial feature to the pose feature at time step j during time step i.

Modal Fusion: Concatenate the raw facial expression representation with the attention-weighted pose representation to obtain the fused temporal features:

$$H_{fused} = \left[H_{face}; H_{pose2face} \right] \in R^{T \times 2d_h} \quad (7)$$

(iii)Status Classification and Early Warning Triggering

Classification: The fused temporal feature H_{fused} is mapped to a four-state classification probability distribution via a fully connected layer $W_c \in R^{2d_h \times 4}$ and a $Softmax$ function.

$$\hat{y}_t = Softmax\left(W_c^T h_{fused,t} + b_c \right) \quad (8)$$

Alert Trigger Conditions: Define an alert function $Alert(.)$. Let the sliding window length be L, the intensity threshold be θ, and the duration threshold be τ. $p_t^{(3)}$ and $p_t^{(4)}$ represent the probabilities of fatigue and boredom, respectively. When the moving averages of both probabilities within the window simultaneously exceed the threshold θ and persist for longer than τ, the system triggers a fatigue alert and sends the signal to the intervention engine.

For the current time T, if:

$$\frac{1}{L} \sum_{t=T-L+1}^{T} \left(p_t^{(3)} + p_t^{(4)} \right) > \theta \quad \text{且}$$

$$Count\left(\left\{ t \in [T-L+1, T] \mid p_t^{(3)} + p_t^{(4)} > \theta \right\} \right) > \tau \quad (9)$$

Then trigger an alert.

4.2 Hierarchical Adaptive Intervention Decision Mechanism

Upon receiving an early warning signal, the mechanism generates the optimal intervention strategy $a^* \in A$, where A is the set of intervention actions. It employs a hierarchical structure to ensure both immediate responsiveness and long-term optimization.

4.2.1 Formal Definition of the Problem

Model intervention decisions as a sequential decision problem. At each decision point k (triggered by an alert), the agent observes the current state $s_k \in S$, selects an action $a_k \in A$, and subsequently transitions the environment to the new state S_{k+1}, receiving a reward r_k. The objective is to learn a policy $\pi : S \to A$ that maximizes the cumulative discounted reward $E\left[\sum_{k=0}^{\infty} \gamma^k r_k \right]$, where $\gamma \in [0,1]$ is the discount factor.

4.2.2 Mathematical Formulation of Hierarchical Decision Models

(i) Rule-Based Rapid Responder

This component is a deterministic function: $\pi_{rule}(s)$.

$$\pi_{rule}(s) = \begin{cases} a_1, & if\ s \in S_1 \\ a_2, & if\ s \in S_2 \\ \ldots & \\ a_{default}, & otherwise \end{cases} \quad (10)$$

In this framework, the state space is partitioned into multiple sub-regions S_i. For instance,

$S1 = \{ s \mid s.type = "Fatigue" \wedge\ s.intensity > 0.8 \wedge s.duration > 5min \}$,

and the corresponding intervention action a_1 is "Mandatory Rest". The decision logic $Rule(S\_\{type\}, S\text{-}\{intensity\}, T\_\{duration\})$ is summarized in the table below. This rule-based system guarantees millisecond-level response times for frequently encountered scenarios. The results are shown in Table 1. The "Focused" state in Table 1 refers to situations where prolonged high concentration may potentially lead to fatigue.

Table 1. Decision Logic Table

Fatigue Type	Strength Condition	Duration Condition	Intervention Action
Focused	S_intensity>0.8	T_duration>5min	Pop up a gentle reminder of "appropriate rest to maintain focus" which is non-intrusive and no screen locking, and provide an optional "5-minute rest timer" that students can activate voluntarily; simultaneously adjust the learning interface to reduce eye strain.
Fatigue	S_intensity>0.6	T_duration>3min	Pop up a reminder to take a break and lock the screen for 30 seconds.
Boredom	S_intensity>0.7	T_duration>3min	Push a fun short video or a small piece of knowledge related to the current topic.
Confusion	S_intensity>0.75	T_duration>2min	Provide a 'Need a hint?' option, or bring up the relevant knowledge graph card.
Any Category	S_intensity>0.9	Any	Highest priority alert, it is recommended to contact a

Fatigue Type	Strength Condition	Duration Condition	Intervention Action
			teacher or counselor.

(ii)Reinforcement Learning-Based Policy Optimizer

To achieve personalized, long-term adaptation for different students, a Deep Q-Network (DQN) agent was introduced.

State Space: The state S_k is a vector defined as the aggregation of the fused fatigue feature vectors H_{fused} from the past m time steps, augmented with an embedded representation of the current learning task context.

$$s_k = \left[MeanPool\left(H_{fused}\right); e_{task}; acc \right] \tag{11}$$

Here, e_{task} represents the embedding vector of the current learning task/knowledge point, while acc denotes the student's historical acceptance rate vector for various interventions.

Action Space: Action A is a discrete set comprising all available intervention strategies.

A = {Do nothing, Recommend rest, Push engaging content, Adjust content difficulty, Provide hints}.

Reward Function $r(s,a)$: The design of the r_k reward is crucial, as it guides the agent in learning "good" policies.

$$r_k = w_1 \cdot \Delta F + w_2 \cdot \Delta B + w_3 \cdot II_{complete} - w_4 \cdot II_{reject} - w_5 \cdot II_{intrude} \tag{12}$$

Among them, ΔF and ΔB represent the decrease in fatigue and boredom probability over the Δt time period following the intervention. $II_{complete}$ is a task completion indicator. If students complete learning tasks after the intervention, it serves as a positive reinforcement. II_{reject} is a punishment where students explicitly reject interventions. If a student manually closes or negatively evaluates an intervention, it counts as a negative reward. $II_{intrude}$ is a minor penalty for overly frequent interference .

Network Architecture and Training: Use a neural network $Q(s,a;\theta)$ to approximate Q^* , and learns parameters θ by minimizing the mean squared Bellman error, where the input is the state s and the output is the Q-value corresponding to each action a.

$$Q^*(s,a) = E_{s'}\left[r + \gamma \max_{a'} Q^*(s',a') \mid s,a \right] \tag{13}$$

$$L(\theta) = E_{(s,a,r,s')\sim D}\left[\left(r + \gamma \max_{a'} Q(s',a',\theta^-) - Q(s,a;\theta) \right)^2 \right] \tag{14}$$

Where D is the experience replay cache, θ^- are the parameters of the target network, periodically copied from θ. Employing an experience replay mechanism, interaction data $(st, at, rt, st+1)$ is stored in the memory bank and randomly sampled for training to break data correlations. Target Q-networks and periodic update strategies are employed to stabilize the training process.

When an early warning is triggered, the rapid responder first implements primary intervention. Simultaneously, the alert event is fed into the policy optimizer, where the agent selects an action a_t^{RL} based on the current state s_k --which may differ from the rule. If the confidence level (Q-value) of a_t^{RL} significantly exceeds the estimated value of the rule-based action, the system may override the rule and execute this more strategic secondary intervention. This design ensures the system maintains foundational performance while possessing the capability for continuous learning and optimization

5. Experiment

5.1 Experimental Setup

The experimental dataset is the self-built "Online-Learning-Fatigue" dataset, comprising 120 hours of online learning video data from 50 volunteers, with fatigue states meticulously annotated by educational psychology experts. The proposed method is compared with four models: an LSTM model based solely on facial expressions, an LSTM model based solely on body posture, an early fusion model, and a late fusion model. Data analysis encompasses classification accuracy, precision, recall, F1 score, intervention effectiveness (measured by changes in pre-intervention and post-intervention burnout scores), and user satisfaction surveys.

5.2 Experimental Results and Analysis

Burnout identification results are shown in Table 2. The findings indicate that all evaluation metrics of the proposed method achieve optimal performance, with an F1 score 3.3% higher than the best baseline method. Multi-modal approaches generally outperform unimodal methods. The cross-modal attention mechanism significantly enhances fusion effectiveness, underscoring the critical importance of temporal modeling. While maintaining high performance, the proposed method keeps parameter counts and inference time within reasonable bounds

Table 2. Burnout Identification Results

Model	Accuracy	Precision	Recall	F1 Score	Inference Time (ms)
Only facial expressions	0.814	0.798	0.805	0.801	15.2
Just posture	0.763	0.752	0.741	0.746	12.8
Early Fusion	0.841	0.829	0.832	0.830	18.5
Late-stage fusion	0.857	0.843	0.848	0.845	20.3
This paper's model	0.923	0.911	0.909	0.910	21.9

The classification performance for different burnout states is shown in Table 3. The focused state achieved the highest recognition accuracy due to its most distinct features. Confusion and fatigue states exhibited similar recognition performance with some overlap. Although the boredom state had fewer samples, its recognition performance remained acceptable.

A user satisfaction survey revealed that 92% of students found the intervention "timely and helpful," and reported "feeling cared for by the system."

Table 3. Experimental Data for Different Burnout States

Status Type	Precision	Recall	F1 Score	Number of Support Samples
Focus	0.945	0.932	0.938	2,160
Confusion	0.889	0.901	0.895	1,200
Fatigue	0.876	0.864	0.870	960
Boredom	0.834	0.821	0.827	480

Real-time performance analysis is shown in Table 4, demonstrates that the system can process data in real-time at a frame rate of 35 fps, meeting the real-time requirements for online learning scenarios.

Table 4. Performance Data

Module	Processing Time (ms)	Memory Usage (MB)	CPU Usage (%)
Facial Expression Recognition	12.3	156	15.2
Pose Estimation	8.7	89	8.9
Sequence Fusion	5.2	234	12.1
Intervention Decision	2.1	45	3.2
Total	28.3	524	39.4

Evaluation of Intervention Effectiveness: Analysis of 100 intervention-triggering events revealed that 85% of interventions occurred within two minutes of triggering, students' overall burnout scores decreased by more than 50%.

6. Conclusion

This paper designs and implements an online learning burnout real-time early warning and intervention system based on multi-modal sequence fusion.Through the innovative use of cross-modal attention mechanisms, the system can accurately discern students' emotional and behavioral states. A hybrid intervention strategy combining rule-based and reinforcement learning enables the system to respond rapidly while achieving long-term optimization, thereby delivering adaptive interactions with a human touch. Experiments have demonstrated the superiority of this system in terms of recognition accuracy and intervention effectiveness. This study provides a feasible technical pathway and practical model for constructing next-generation online learning environments equipped with emotional intelligence. Future work will explore integrating additional multi-modal data such as eye tracking and heart rate, and conduct in-depth research into the differentiated preferences for intervention strategies among students from diverse cultural backgrounds and age groups.

Acknowledgments

This work was supported by the education reform project of Beijing Union University "Construction and Practice of Diversified Computer Science and Technology Major Course Group for Students"(JJ2024Z004) and the R&D Program of Beijing Municipal Education Commission(KM202111417002).

References

[1] Fida A, Umer M, Saidani O, Hamdi M, Alnowaiseret K, Bisogni C, et al. Real time emotions recognition through facial expressions. Multimed Tools Appl. 2025;84(29):34753-34780.

[2] Dubbaka A, Gopalan A. Detecting learner engagement in MOOCs using automatic facial expression recognition. In: IEEE Global Engineering Education Conference; 2020; virtual. Piscataway: IEEE; 2020. p.447-456

[3] Mukhopadhyay M, Pal S, Nayyar A, Pramanik PKD, Dasgupta N, Choudhury P. Facial emotion detection to assess learner's state of mind in an online learning system. In: 5th International Conference on Intelligent Information Technology; 2020; Hanoi, Vietnam. New York: ACM; 2020. p.107-115.

[4] Gupta S, Kumar P, Tekchandani RK. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimed Tools Appl. 2023;82(8):11365-11394.

[5] Long DT. A facial expressions recognition method using residual network architecture for online learning evaluation. J Adv Comput Intell Intell Inform. 2021;25(6):953-962.

[6] Mohamad Nezami O, Dras M, Hamey L, Richards D, Wan S, Paris C. Automatic recognition of student engagement using deep learning and facial expression. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2019; Cham, Switzerland. Cham: Springer International Publishing; 2019. p.273-289

[7] Wang TC, Qiu Q, Wang CT, Chen FL. LD-identify: network learning state recognition based on passive RFID. Control Decis. 2024;39(1):219-226.

[8] Cao XM, Zhang YH, Pan M, Zhu S, Yan HL. Research on learning engagement recognition method from the perspective of artificial intelligence: deep learning experimental analysis based on multimodal data fusion. Mod Educ Technol. 2025;35(3):5-12.

[9] Jiang J, Yu WT, Wang HY. Research on students' learning behaviors in smart classroom based on multimodal data. China Educ Inf. 2024;30(4):107-117.

[10] Wang YY, Zheng YH. Multimodal data fusion: core driving force to solve key problems in intelligent education. Mod Dist Educ Res. 2022;34(2):93-102.

[11] Wang LY, He YF, Tian JH. Construction and empirical study of multimodal data fusion model for online learning behavior. Chin J Dist Educ. 2020;6:23-51.

[12] Xue YF, Chen Z, Qiu YS, Zhu FQ. Intelligent recognition of cognitive styles in online learning based on multimodal data. Open Educ Res. 2024;30(5):112-120.

[13] Zhang S, Yin CY. Sequential multimodal sentiment analysis model based on multi-task learning. J Comput Appl. 2021;41(6):1631-1639.

[14] Jagadeesh M, Baranidharan B. Facial expression recognition of online learners from real-time videos using a novel deep learning model. Multimed Syst. 2022;28(6):2285-2305.

[15] Savchenko AV, Savchenko LV, Makarov I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Trans Affect Comput. 2022;13(4):2132-2143.

[16] Olivetti EC, Violante MG, Vezzetti E, Marcolin F, Eynard B. Engagement evaluation in a virtual learning environment via facial expression recognition and self-reports: A preliminary approach. Appl Sci. 2019;10(1):314.

[17] Aly M. Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model. Multimed Tools Appl. 2025;84(13):12575-12614.

[18] Ngo D, Nguyen A, Dang B, Ngo H. Facial expression recognition for examining emotional regulation in synchronous online collaborative learning. Int J Artif Intell Educ. 2024;34(3):650-669.

[19] Pabba C, Kumar P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert Syst. 2022;39(1):e12839.

[20] Sumalakshmi CH, Vasuki P. Fused deep learning based facial expression recognition of students in online learning mode. Concurr Comput Pract Exp. 2022;34(21):e7137.

[21] Gupta S, Kumar P, Tekchandani RK. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimed Tools Appl. 2023;82(8):11365-11394.

Author Contributions

Kun Liu (First Author): Conceptualized the research framework and core objectives of online learning burnout early warning and adaptive intervention; led the design and mathematical derivation of the multi-modal temporal feature fusion algorithm (including cross-modal attention mechanism and LSTM-based sequence encoding); responsible for the overall implementation of experiments, including model training, performance testing, and result analysis; drafted the initial manuscript and revised key technical sections.

Xiaoxiao Gu (Second Author): Responsible for experimental testing and validation; executed the real-time performance testing of the system (e.g., inference time, CPU/memory usage measurement); verified the effectiveness of intervention strategies through user feedback collection; assisted in organizing and cleaning experimental result data; and participated in the preparation of experimental tables and figures.

Jingxia Chen (Third Author & Corresponding Author): Contributed to the design and development of the core algorithm, particularly participating in the optimization of the hierarchical adaptive intervention mechanism (rule-based decision logic and DQN-based reinforcement learning policy design); provided critical guidance on research methodology and experimental design; supervised the entire research process, including data validation and algorithm robustness verification; revised the manuscript for academic rigor and language accuracy; and coordinated the submission process. Corresponding author email:yykjtjingxia@buu.edu.cn.

Wenjuan Shao (Fourth Author): Undertaken the preparation of the experimental dataset ("Online-Learning-Fatigue" dataset), including collecting 120 hours

of online learning video data from 50 volunteers, collaborating with educational psychology experts for fatigue state annotation, and preprocessing multi-modal data (video frame extraction, key-point coordinate normalization); responsible for the maintenance of experimental equipment and software environments (e.g., configuring MediaPipe Pose and MobileViT inference environments); and supported the statistical analysis of intervention effectiveness data.

All authors have read and approved the final manuscript, and agree to the submission of this work.