ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on College Students 'Behavior Analysis and Personalized Learning Support System Based on Multimodal Data Fusion

Linglan Gao[1,*]

[1] School of Education, Fuzhou University of International Studies and Trade, Fuzhou, 350202, Fujian, China

SUMMARY: *Under the background of digital teaching transformation in colleges and universities, students 'learning behavior presents multi-source, dynamic and implicit characteristics, and a single platform log is difficult to support accurate analysis and personalized support. This paper constructs a college students 'behavior analysis and personalized learning support system based on multimodal data fusion. It collects learning platform logs, classroom visual behaviors, discussion texts, evaluation records, resource access and terminal signals, and constructs a unified learning behavior sequence through timestamp alignment, missing repair, anomaly suppression, robust normalization, text semantic coding and sliding time Windows. In the model layer, CNN local behavior feature extraction, BiLSTM bidirectional temporal dependency modeling and Attention key modality weighting mechanism are fused to realize the recognition of concentration, participation status and learning risk. Personalized feedback is generated by combining state prediction, risk scoring, resource matching and profile update. Based on the 16-week learning data of 1280 students, 24860 sequence samples were formed. The results show that the Accuracy of the model reaches 95.8%, the F1 is 95.1%, the AUC is 0.972, the recommendation accuracy is 92.6%, and the average response delay is 149 ms. The research provides technical reference and practical significance for college students 'learning status perception, risk early warning and personalized support.*

KEYWORDS: *Multi-modal data fusion; Student behavior analysis; Personalized learning support; Deep learning*

1 Introduction

1.1 Research Background

The teaching scene in colleges and universities is shifting from traditional classroom management to data-driven intelligent learning support. Students continue to generate multi-source behavior data in the process of classroom learning, online platform access, resource browsing, assignment submission, interactive discussion and terminal operation. These data not only contain explicit learning records such as clicks, pauses, answers and viewing time, but also include implicit behavioral cues such as expression, posture, attention changes, learning rhythm and interaction frequency [1-3]. It is difficult for a single data source to completely depict students' learning status, and it is easy to cause problems such as lagging behavior judgment, extensive support strategies, and untimely recognition of learning

*gaolinglann@163.com
https://doi.org/10.65102/is2026908 Published: 30 April, 2026

risks [4]. With the development of computer vision, natural language processing, time series modeling and deep learning technologies, multimodal data fusion provides a new technical basis for college student behavior analysis [5, 6]. Through the unified modeling of learning platform logs, classroom video features, text interaction content and academic performance data, students 'concentration, participation level, knowledge mastery status and potential learning difficulties can be identified in a more detailed way. On this basis, the construction of personalized learning support system is helpful to realize the dynamic perception of learning status, accurate push of learning resources and automatic generation of learning intervention strategies, so as to improve the intelligent level of teaching management in colleges and universities [7].

1.2 Technical Route

Focusing on the needs of college students 'behavior analysis and personalized learning support, this paper constructs an intelligent analysis framework based on multimodal data fusion. In the data layer, the system collects learning management platform logs, classroom image behaviors, online discussion texts, assignment assessment results and learning resource access records, and forms a unified data view through timestamp alignment, student identity mapping and course unit association. In the preprocessing stage, missing values, outliers and noise data are cleaned, and normalization, text vectorization, image feature extraction and sliding time window segmentation methods are used to transform heterogeneous data into sequence representations that can be input to the model. In the model layer, CNN is introduced to extract the local features of visual behavior, BiLSTM is used to capture the dependence of learning behavior over time, and the attention mechanism is combined to allocate the feature weights of different modalities and different periods of time to improve the accuracy and interpretability of learning state recognition. The output layer of the system completed the concentration recognition, participation classification, learning risk prediction and personalized resource recommendation, and fed back the results to the teacher end and the student end to realize the closed-loop operation of behavior analysis, state prediction, resource matching and intervention feedback.

2 Related work

2.1 Analysis and Research of College Students 'Behavior Based on multimodal Data

Cerezo et al. systematically sorted out the differences between learning analytics and educational data mining, and pointed out that the two are developing towards the integration of educational data science, which can support learning process modeling, behavior interpretation and teaching decision-making, but there are still problems such as scattered data sources, insufficient explanation depth and inadequate adaptation of educational scenarios [8]. Khosravi et al. built an intelligent learning analysis dashboard to help teachers explore students' learning data through automatic dry-down recommendation, indicating that visual analysis can enhance teachers' ability to identify abnormal behaviors and learning risks, but the fusion of multi-modal behavior features is still limited [9]. Rets et al. put forward practical suggestions around the ethical use of predictive learning analytics, emphasizing that data collection, model prediction and learning intervention must balance privacy protection and educational equity [10]. Yusuf et al. used multimodal learning analysis to model students' learning behaviors in the programming classroom, and proved that information such as

expression, posture and interaction frequency could improve the accuracy of behavior recognition [11]. Corza-Vargas et al. analyzed the visualization of cognitive emotional states from the perspective of students, suggesting that the behavior analysis system in colleges and universities should pay attention to privacy, cultural differences and interface acceptability [12].

2.2 Deep learning methods for learning behavior time series recognition

Fazil et al. proposed a deep learning student performance prediction model based on engagement data, indicating that sequential learning behaviors can effectively reflect the changes in students' states [13]. Khenkar et al. conducted a deep analysis of students' physical activities in online learning scenarios and realized the detection of learning engagement states, but its input mode still favored a single behavioral dimension [14]. Trakunphutthirak and Lee used progressive time data to predict students' academic performance, indicating that learning records in different periods had differentiated contributions to the prediction results [15]. Alamri and Alharbi systematically summarized interpretable student performance prediction models, and pointed out that the deep model still needs to enhance the explanation ability of feature contribution while improving accuracy [16]. Albreiki et al. reviewed the application of machine learning in student performance prediction and argued that classification, regression and ensemble learning methods have been relatively mature, but the description of complex temporal relationships is still insufficient [17]. Rodriguez-Hernandez et al. analyzed the implementation of artificial neural network in academic prediction and emphasized that the choice of predictor variables would directly affect the stability of the model [18]. Bhardwaj et al. applied deep learning to online learning engagement analysis and verified the value of deep feature extraction for complex behavior recognition [19]. Gupta et al. built a real-time learning engagement detection system based on facial expression recognition, which provided a reference for visual modal behavior recognition [20]. Alnasyan et al. further summarized the research on deep learning prediction in virtual learning environment, indicating that deep models are suitable for processing high-dimensional, multi-source and dynamic learning data [21].

2.3 Research on Personalized Learning Support System and Predictive Learning Analysis

Kaouni et al. designed an adaptive online learning model based on artificial intelligence to improve the adaptability of online teaching through student status perception and resource adjustment, but the system feedback mechanism still relies on rule setting [22]. Al-Zahrani and Alasmari studied the role of learning analytics in data-driven decision making in online higher education and pointed out that learning data can support teaching personalization and student engagement improvement [23]. Saleem and Aslam proposed a multi-dimensional deep learning method, which combined student participation insight with adaptive content recommendation, and proved that deep feature fusion could improve the effect of personalized learning support [24]. Janiesch et al. discussed the basic mechanism of machine learning and deep learning, indicating that deep learning has the advantages of automatic feature learning and complex pattern recognition [25]. Sarker systematically summarizes machine learning algorithms, real-world applications and research directions, and provides a general method basis for predictive modeling in educational scenarios [26]. Prabowo et al. used time series data and tabular data aggregation for GPA prediction of college students, and verified the effectiveness of cross-type data fusion for learning outcome prediction [27]. These studies

provide a technical basis for this paper to build a personalized learning support system integrating behavior recognition, state prediction and resource recommendation.

3 Multimodal data fusion of college students' behavior analysis and personalized learning support system construction

3.1 Data collection and unified representation of multimodal learning behavior of college students

College students' learning behavior has cross-scenario, cross-terminal and cross-modal characteristics. A single platform log can only reflect students' "whether to visit" and "how long to visit", which is difficult to explain their real participation status, knowledge absorption process and potential learning risks. In order to improve the integrity of behavior analysis, this paper constructs a multi-modal learning behavior data collection framework for university teaching scenarios, which integrates learning management platform logs, classroom visual behaviors, online discussion texts, evaluation scores, resource access trajectories and terminal perception signals into a unified data space. The system takes student identity, course unit and timestamp as the core index, transforms discrete learning events into continuous behavior sequences, and provides data basis for subsequent student status recognition and personalized learning support.

Let the set of multimodal learning observations of student i at time t be as follows:

$$X_i^t = \{x_{i.1}^t, x_{i.2}^t, \cdots, x_{i.M}^t\} \tag{1}$$

where, M represents the number of data modalities, $x_{i.M}^t$ represents the learning behavior characteristics collected by the M-th modality at time t, such as click log, classroom posture, text interaction, score record and resource access path. This expression can integrate data from different sources into the modeling process of student behavior, avoiding the subsequent analysis only relying on a single learning record.

In the actual acquisition process, the sampling frequency of different modal data is not consistent. Classroom images are usually collected at the second level or frame level, platform logs are recorded in an event-triggered manner, and evaluation scores show low-frequency update characteristics. In order to ensure that multi-source data can be processed by the model at the same time scale, this paper uses the timestamp alignment mechanism to map each modal data into a fixed learning window:

$$\tilde{x}_{i,m}^k = \frac{1}{|T_k|} \sum_{t \in T_k} x_{i,m}^t \tag{2}$$

where, T_k represents the KTH learning time window, $|T_k|$ represents the number of effective sampling points within this window, and $x_{i,m}^t$ represent the MTH class of modal features after window aggregation. Through this process, the system can compress the scattered learning clicks, stage evaluation results and continuous visual behavior into a unified time slice, so as to retain the stage characteristics of student behavior changes.

In the unified representation stage, different modalities cannot be simply concatenated, otherwise it is prone to problems such as high-dimensional redundancy, noise amplification and unbalanced modal contribution. In this paper, an independent encoding function is set for

each type of modality to map the raw data into low-dimensional semantic vectors:

$$z_{i,m}^{k} = f_m(\tilde{x}_{i,m}^{k}; \theta_m) \tag{3}$$

where, $f_m(\cdot)$ represents the feature encoder corresponding to the m-th mode, θ_m represents the encoder parameters, $z_{i,m}^{k}$ represent the semantic features of the mode under the KTH time window. For the platform logs, the system extracted the characteristics of access frequency, stay time and learning path jump. For visual behavior, the system extracts features of expression, gaze direction and posture change. For text interaction, the system obtains the discussion topic and sentiment orientation through word embedding and semantic coding. For the evaluation records, the accuracy rate, submission delay and knowledge mastery degree are extracted.

Considering that the contribution of different modes in different learning stages is not fixed, this paper introduces the adaptive weight fusion method to generate the unified representation vector of students 'learning behavior:

$$h_i^k = \sum_{m=1}^{M} \alpha_m^k z_{i,m}^k \tag{4}$$

where, h_i^k represents the unified behavior representation of the ith student in the KTH time window, α_m^k represents the fusion weight of the m-th modality in the current learning window, and satisfies $\sum_{m=1}^{M} \alpha_m^k = 1$. When students were in the classroom learning stage, the weight of visual posture and interactive behavior was relatively high. When students are in the stage of autonomous learning after class, the weight of resource access, assessment performance and text feedback will increase. The dynamic fusion method can enhance the adaptability of the system to different learning scenarios. Figure 1 shows the architecture of multimodal student behavior data collection and fusion.

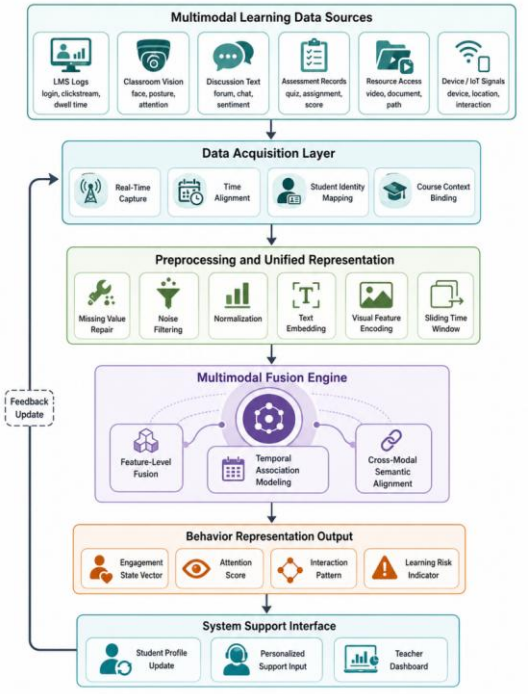

Figure 1: Architecture diagram of multimodal student behavior data collection and fusion

As shown in Figure 1, the multimodal student behavior data acquisition and fusion architecture consists of a data source layer, a collection layer, a preprocessing layer, a fusion representation layer, and a system support layer. The data source layer is responsible for accessing platform logs, classroom vision, discussion texts, evaluation records, resource access and terminal signals. The acquisition layer completed real-time capture, timestamp alignment, student identity mapping and course context binding. The fusion representation layer outputs engagement status, attention score, interaction mode, and learning risk indicators. Through the above design, student behavior data are transformed from scattered records into computable, traceable, and interpretable unified representations, which lays a foundation for the subsequent CNN-BiLSTM-Attention model to identify learning states and generate personalized learning support strategies.

3.2 Data preprocessing and feature engineering for multi-source heterogeneous learning

Multi-source heterogeneous learning data usually have problems such as inconsistent time granularity, uneven distribution of missing values, more visual noise in the classroom, sparse text semantics and large differences in numerical scales after collection. In order to avoid feature deviation and recognition error caused by the original data directly entering the model, a preprocessing link of "quality repair, anomaly suppression, scale unification, semantic coding, sequence construction" is constructed at the front end of the system, so that platform logs, classroom images, discussion texts, evaluation records and terminal signals can participate in the subsequent learning state recognition in the unified computing space.

Aiming at the problems of platform log breakpoint, terminal signal packet loss and visual frame missing, this paper uses the joint estimation method of time neighborhood and student similar neighborhood to complete missing value imputation:

$$\hat{x}_{i,m}^{t} = \gamma \cdot \frac{1}{|\Omega_t|} \sum_{\tau \in \Omega_t} x_{i,m}^{\tau} + (1-\gamma) \cdot \sum_{j \in \mathcal{N}_i} \omega_{ij}\, x_{j,m}^{t} \tag{5}$$

where, $\hat{x}_{i,m}^{t}$ represents the MTH modal feature after repair, Ω_t represents the effective temporal neighborhood near time t, N_i represents the set of students with similar learning behavior with student i, ω_{ij} represents the contribution weight of similar students, γ represents the adjustment coefficient between temporal neighborhood information and group similarity information. This method not only retains the continuity of students' own behavior, but also compensates for short-term data missing by using similar learning groups.

In order to reduce the interference of abnormal clicks, invalid pauses, illumination occlusion and extreme terminal signals on model training, this paper introduces an abnormal score based on median deviation:

$$r_{i,m}^{t} = \frac{\left| x_{i,m}^{t} - \mathrm{Med}(x_m) \right|}{\mathrm{MAD}(x_m) + \varepsilon} \tag{6}$$

where $r_{i,m}^{t}$ denote the intensity of anomalies, $\mathrm{Med}(x_m)$ denotes the median of the m-th class features, $\mathrm{MAD}(x_m)$ denotes the median absolute deviation, and ε is used to avoid the denominator being zero. When the abnormal intensity exceeds the threshold, the system truncates, reduces the weight or labels this sample, so as to reduce the non-learning behavior noise into the feature space.

The dimensions of different data modalities are quite different, for example, dwell time,

click count, achievement score, and visual confidence are not in the same numerical range. In order to improve the convergence stability of the model, this paper uses the robust normalization method with quantile constraints:

$$x_{i,m}^{t*} = \mathrm{clip}\left(\frac{x_{i,m}^{t} - Q_{50}(x_m)}{Q_{75}(x_m) - Q_{25}(x_m) + \varepsilon}, -c, c\right) \tag{7}$$

where $x_{i,m}^{t*}$ denote the normalized eigenvalues, $Q_{25}(x_m)$, $Q_{50}(x_m)$ and $Q_{75}(x_m)$ denote the lower, median and upper quartiles, respectively, and c denotes the cut-off boundary. Compared with ordinary standardization, this method is not sensitive to abnormal learning records, and is more suitable for imbalanced behavior data in real teaching scenarios in colleges and universities.

In terms of text feature processing, online discussions, course questions and answers, and learning feedback usually contain short sentences, ellipsis expressions, and emotional tendencies. In this paper, we input the segmented text into a lightweight semantic encoder, and fuse the sentiment score and topic distribution to form text features:

$$e_{i,\mathrm{text}}^{k} = \mathrm{Pool}(\mathrm{Encoder}(w_1, w_2, \cdots, w_L)) \oplus s_i^k \oplus p_i^k \tag{8}$$

where, $e_{i,\mathrm{text}}^{k}$ represent the text semantic vector under the KTH learning window, $w_1, w_2, \cdots, w_L$ represent the text word sequence, Encoder($\cdot$) represent the semantic encoder, Pool($\cdot$) represent the pooling operation, s_i^k represent the sentiment tendency feature, p_i^k represent the topic distribution feature, and $\oplus$ represent the vector concatenation. This processing can convert discussion quality, sentiment change and learning perplexity into computable features together.

After completing numerical inpainting, exception handling, scale normalization and semantic encoding, the system organizes multi-class features into a masked learning behavior sequence:

$$S_i^k = [v_{i,\log}^k \oplus v_{i,\mathrm{vis}}^k \oplus e_{i,\mathrm{text}}^k \oplus v_{i,\mathrm{test}}^k \oplus v_{i,\mathrm{res}}^k] \odot M_i^k \tag{9}$$

where, S_i^k represents the feature sequence of the ith student in the KTH window; $v_{i,\log}^k$,$v_{i,\mathrm{vis}}^k$,$v_{i,\mathrm{test}}^k$ and $v_{i,\mathrm{res}}^k$ represents the log, visual, evaluation and resource access features respectively; M_i^k represents the modal validity mask; $\odot$ represents the per-element constraint. The mask mechanism can distinguish between true low activity and missing data, and avoid the model misjudging the missing modality as insufficient learning input. After the above feature engineering processing, the originally scattered and noisy learning data is transformed into an input sequence with stable structure, clear semantics, and can be used for deep temporal modeling, which provides a reliable basis for subsequent student behavior recognition models.

3.3 Construction of student behavior recognition Model based on CNN-BiLSTM-Attention

After the multi-source heterogeneous learning data preprocessing, the student learning behavior sequence has been composed of platform log, visual posture, text interaction, evaluation performance and resource access characteristics. In order to improve the accuracy of college students 'behavior state recognition, this paper constructs a student behavior

recognition model fused with CNN-BiLSTM-Attention. The model takes the unified feature sequence in a sliding time window as input, extracts the local behavior combination features through CNN, captures the sequential dependence of students 'learning states by BiLSTM, and then highlights the key modes and key time slices through the attention mechanism, and finally outputs the concentration, participation and learning risk status. Compared with the single classification model, this structure can deal with short-term behavior fluctuations and long-term learning trends at the same time, which is more suitable for the continuous monitoring of college students 'learning process.

Let the input sequence of the ith student in the KTH time window be S_i^k. The model first extracts the local learning behavior pattern through a one-dimensional convolution layer:

$$C_i^k = \sigma\left(W_c * S_i^k + b_c\right) \tag{10}$$

where, C_i^k represents the local behavior feature after convolution extraction, W_c represents the convolution kernel parameter, $*$ represents the convolution operation, b_c represents the bias term, and $\sigma(\cdot)$ represents the nonlinear activation function. This layer mainly identifies the behavior combination relationship in a short time, such as the local patterns of "frequent switching of resources - decrease of stay time - decrease of discussion activity", which provides compact features for subsequent time series modeling.

After the convolution output enters the BiLSTM layer, the model learns the change law of student behavior from forward and reverse respectively:

$$H_i^k = \left[\overrightarrow{LSTM}(C_i^k), \overleftarrow{LSTM}(C_i^k)\right] \tag{11}$$

where, H_i^k represents bidirectional temporal hidden state, $\overrightarrow{LSTM}$ is used to capture the evolution process from the early learning behavior to the current state, and $\overleftarrow{LSTM}$ is used to supplement the reverse association of subsequent behaviors to the current state. The structure is able to identify continuous change features such as a gradual decline in student engagement, enhanced resource access after evaluation fluctuation, and increased risk after discussion reduction.

In order to avoid all time slices and all modalities being equally weighted, this paper introduces an attention mechanism after BiLSTM to perform weighted aggregation of key behavior segments:

$$\beta_t = \frac{\exp(q^\top \tanh(W_a h_t + b_a))}{\sum_{\tau=1}^{T} \exp(q^\top \tanh(W_a h_\tau + b_a))} \tag{12}$$

where β_t represents the attention weight of the TTH time slice, h_t represents the hidden state output by BiLSTM in this time slice, W_a, b_a and q are trainable parameters, and t represents the number of time slices in the window. Attention weights are able to highlight segments with higher contribution to action recognition, such as signals such as abnormal visits before and after assignment submission, persistent low interaction in class, and increased assessment error rate in a short period of time.

The model inputs the weighted temporal context vector into the multi-task recognition layer to obtain the output of student behavior state:

$$\hat{y}_i^k = \mathrm{Softmax}\left(W_o \sum_{t=1}^{T} \beta_t h_t + b_o\right) \tag{13}$$

where, $\hat{y}_i^k$ represents the student behavior recognition results, W_o and b_o represent the output layer parameters. The output categories included high focus, medium focus, low focus, high engagement, average engagement, low engagement, and potential learning risk ratings. In the model training phase, cross-entropy loss and weight decay are jointly optimized, and Dropout and early stopping strategies are used to reduce the risk of overfitting. The structure combines local behavior perception, bidirectional temporal reasoning and key segment focusing, which makes student behavior recognition shift from static label judgment to dynamic process analysis.

Figure 2 shows the structure of the student behavior recognition model fused with CNN-BiLSTM-Attention. The model consists of input feature sequence, convolutional feature extraction, bidirectional temporal modeling, attention weight allocation and multi-task output modules, which can transform multi-modal learning data into interpretable action recognition results layer by layer.

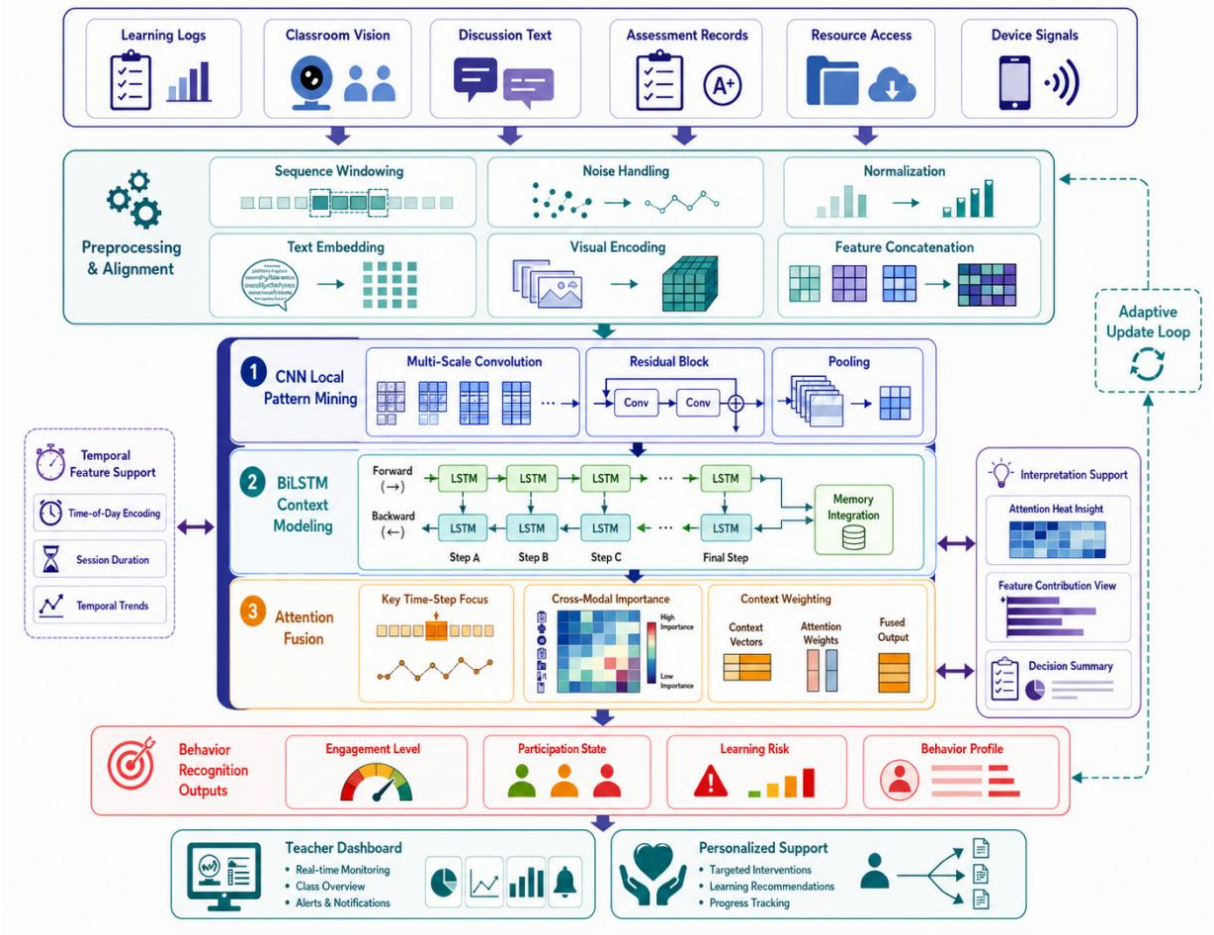

Figure 2: CNN-BiLSTM-Attention student behavior recognition model structure diagram

Figure 2 shows that the CNN module is mainly responsible for extracting the local behavior association in a short time window, the BiLSTM module is responsible for capturing the pre and post evolution relationship of the learning state, and the Attention module weights the key time slices and key behavior signals. After the collaboration of the three, the system can more accurately distinguish between transient behavior fluctuations and continuous learning risks, which provides a reliable basis for the generation of personalized learning support strategies. Table 1 shows the main parameters of the model and the training configuration, which is used to ensure the reproducibility of the model training process and

the comparability of the experimental results.

Table 1: Main model parameters and training configuration table

Parameter Category	Configuration Item	Parameter Value	Technical Function
Input Window	Sliding time window length	30 min	Preserve stage-specific changes in learning behavior
Input Features	Multimodal fused feature dimension	128	Uniformly represent log, visual, text, and assessment features
CNN Kernel	One-dimensional convolution kernel size	3	Extract short-term local behavior combination patterns
CNN Channels	Convolution output channels	64	Enhance local behavior feature representation
BiLSTM Layers	Number of bidirectional recurrent network layers	2	Capture forward and backward temporal dependencies
BiLSTM Hidden Units	Number of hidden layer units	128	Represent the evolution features of students' learning states
Attention Dimension	Attention mapping dimension	64	Assign weights to key time slices
Dropout Ratio	Random dropout rate	0.30	Reduce the risk of model overfitting
Optimizer	Parameter update method	Adam	Improve model convergence efficiency
Learning Rate	Initial learning rate	0.001	Control the gradient update magnitude
Batch Size	Number of samples per training batch	32	Balance training efficiency and GPU memory usage
Output Task	Behavior recognition categories	Focus, participation, and risk level	Support personalized learning feedback generation

The parameter configuration in Table 1 takes into account the recognition accuracy and computational overhead, which is suitable for deployment in the online learning platform of colleges and universities or in the teacher-side analysis system. By unifying the input dimension, controlling the network scale and introducing the attention mechanism, the model can improve the stability of action recognition while ensuring the inference efficiency.

3.4 Learning state prediction and feedback generation mechanism for personalized learning Support

Based on the results of student behavior recognition, the system further constructs the state prediction and feedback generation mechanism for personalized learning support, which integrates the concentration, participation status, resource access, evaluation performance and historical behavior portrait into the decision-making process. The core goal of the mechanism

is to transform the "behavior recognition results" into "executable learning support strategies", so that the system can automatically judge the support needs according to the students' current learning status, push resources to the students and generate intervention tips to the teachers. In this paper, students 'learning status is divided into five categories: stable investment, mild fluctuation, weak knowledge, low participation and high risk, and a closed loop is formed through state prediction, risk score, resource matching and feedback update.

Let the behavior recognition output of the ith student in the KTH learning window be $\hat{y}_i^k$, the unified behavior representation be h_i^k, and the learning state prediction result be expressed as follows:

$$P_i^k = \mathrm{Softmax}(W_s \hat{y}_i^k + U_s h_i^k + b_s) \tag{14}$$

where, P_i^k represents the probability distribution of students in different learning states, W_s and U_s represent the state mapping weights, and b_s represents the bias term. This prediction method uses both model output labels and deep behavior representations, which can reduce the misjudgment of state caused by a single classification result.

To identify students who need timely support, the system constructs a comprehensive learning risk score:

$$R_i^k = \rho_1(1 - p_{\mathrm{focus}}) + \rho_2 p_{\mathrm{low}} + \rho_3(1 - g_i^k) + \rho_4 d_i^k \tag{15}$$

where, R_i^k represents learning risk score, pfocus represents concentration probability, plow represents low participation probability, g_i^k represents knowledge mastery degree, d_i^k represents deviation degree of learning behavior, and ρ_1 to ρ_4 are weight coefficients. This score is able to unify behavioral status, knowledge performance, and anomalous deviation into the same risk scale.

In the resource recommendation stage, the system generates recommendation scores based on the matching degree between student profiles and resource features:

$$M_{i,r}^k = \mathrm{sim}(u_i^k, q_r) + \eta a_r - \mu c_{i,r} \tag{16}$$

where, $M_{i,r}^k$ represent the matching score between student i and learning resource r, u_i^k represent the student profile vector, q_r represent the resource feature vector, a_r represent the resource adaptation quality, $c_{i,r}$ represent the repetition cost between the student's learned content and the resource, η and μ are the adjustment parameters. The design can avoid resource recommendation ranking only according to popularity, and improve the degree of correspondence between resources and individual weaknesses.

In order to realize the continuous optimization of learning support strategies, the system updates the student profile according to the feedback results:

$$u_i^{k+1} = \lambda u_i^k + (1 - \lambda)\Phi(F_i^k, E_i^k, A_i^k) \tag{17}$$

Here, u_i^{k+1} represents the updated student profile, F_i^k represents the system feedback content, E_i^k represents the student completion effect, A_i^k represents the subsequent behavior response, $\Phi(\cdot)$ represents the feedback encoding function, and λ represents the historical portrait retention coefficient. This update mechanism enables the system to adjust the subsequent support strategies based on the true responses of students to the recommended resources and intervention information. The personalized learning support strategy matching rules are shown in Table 2.

Table 2: Personalized learning support strategy matching rule table

Learning State Type	Determination Basis	Student-Side Support Strategy	Teacher-Side Feedback Content
Stable Engagement State	High focus, stable participation, and high assessment accuracy	Push extended reading, advanced exercises, and project cases	Display learning strengths and sustained performance
Mild Fluctuation State	Short-term decline in focus and reduced resource dwell time	Push micro-course clips, knowledge prompts, and staged reminders	Mark students with short-term fluctuations
Knowledge Weakness State	Concentrated assessment errors and low knowledge mastery	Push error-cause analysis, basic exercises, and similar question sets	Provide weak knowledge point distribution
Low Participation State	Limited discussion, low interaction frequency, and single access path	Push interactive tasks, discussion topics, and collaborative resources	Indicate insufficient interaction and participation trends
High-Risk State	Low focus, low participation, and low mastery appearing simultaneously	Push personalized remedial paths and learning plans	Generate key intervention alerts

As can be seen from Table 2, the system feedback does not stop at the risk reminder level, but connects the learning status, resource type, teacher intervention and optimization goal. Through the collaborative operation of state prediction, risk scoring, resource matching and portrait update, the personalized learning support system can form a closed loop of "identification-recommendation-feedback-update", and provide more sophisticated computer-aided decision-making capabilities for the management of college students 'learning process.

4 Experimental design and system implementation

4.1 Source of experimental data and explanation of sample characteristics

In order to verify the effectiveness of multimodal data fusion method in college students 'behavior analysis and personalized learning support, this paper constructs an experimental data set oriented to the real teaching process. The data sources included six types of information, including learning management platform logs, classroom visual behavior records, online discussion texts, course evaluation scores, resource access trajectories and terminal interaction signals. The experimental subjects were students in 5 general courses and professional basic courses in a university, and a total of 1280 valid samples were collected, covering a complete teaching cycle of 16 weeks. The original data included 986432 platform access logs, 18400 classroom behavior image segments, 32860 discussion texts, 76,420 evaluation records and 214,560 resource access records. After data cleaning, anomaly removal, missing value repair and time window segmentation, 24860 groups of multi-modal learning behavior sequence samples were finally formed.

The sample characteristics mainly include login frequency, page stay time, resource switching path, classroom posture change, expression state, discussion emotional tendency, assignment submission delay, evaluation accuracy and knowledge mastery. In order to ensure the stability of model training and testing, this paper divides the training set, validation set and test set according to the ratio of 7:1.5:1.5, and keeps the distribution of samples from different courses, different learning stages and different participation levels relatively balanced. The dataset contains both explicit learning records and implicit behavior changes, which can completely reflect the evolution process of college students 'learning status and provide a data basis for subsequent behavior recognition, risk prediction and personalized support effect analysis.

4.2 System implementation environment and experimental parameter setting

The experimental system was implemented with a three-layer architecture of "data processing layer, model training layer and feedback service layer". The data processing layer was built based on Python 3.10, and mainly used Pandas, NumPy and Scikit-learn to complete multi-modal data cleaning, normalization, time window segmentation and feature encoding. OpenCV and a lightweight convolutional network were used for visual behavior feature extraction, and the text interaction features were used to generate semantic vectors through a Transformer encoder. The model training layer uses PyTorch 2.1 to implement CNN-BiLSTM-Attention network, and the hardware environment is configured as Intel Core i7 processor, 32 GB memory, and NVIDIA RTX 3060 GPU. The operating system is Windows 11 and Ubuntu 22.04, which ensures that the experimental results have good reproducibility.

In terms of experimental parameter Settings, the length of the sliding time window is set to 30 min, the window step is 10 min, and the feature dimension of multi-modal fusion is set to 128. The CNN module adopted a one-dimensional convolution structure, the convolution kernel size was set to 3 and 5, and the number of output channels was 64. BiLSTM was set to 2 layers and the number of hidden units was 128. The Attention map has a dimension of 64. Adam optimizer was used in the training process, the initial learning rate was set to 0.001, the batch size was 32, and the maximum training rounds were 100. To suppress overfitting, a Dropout of 0.30 was added to the model, and an early stopping strategy was used to stop training when the validation set loss did not decrease for 10 consecutive rounds. The evaluation indicators include Accuracy, Precision, Recall, F1, AUC, recommendation hit rate and average response delay, which are used to verify the effectiveness of the method from three aspects of recognition performance, recommendation effect and system deployment efficiency.

5 Performance analysis

5.1 Performance analysis of student learning behavior state recognition

In order to verify the classification ability of the model in the recognition of students 'learning behavior states, CNN, BiLSTM, CNN-bilSTM and CNN-bilSTM-Attention are selected as the comparison models, and the stable investment, mild fluctuation, low participation, weak knowledge and high-risk states are taken as the recognition objects. Figure 3 shows the ROC curves of student behavior recognition for different models.

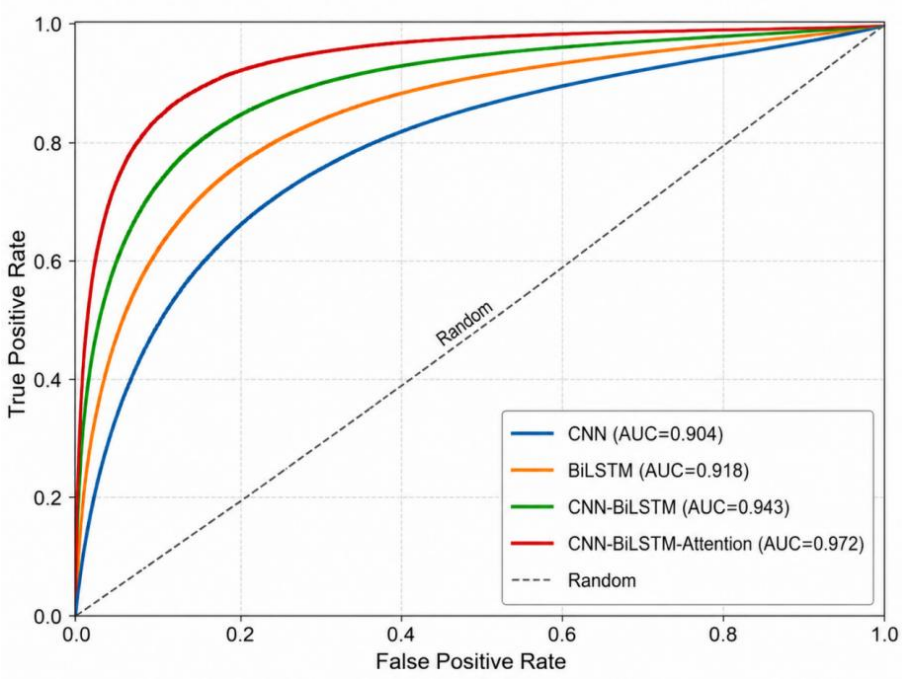

Figure 3: ROC curves of student behavior recognition for different models

As can be seen from Figure 3, the ROC curve of the proposed CNN-BiLSTM-Attention model is generally closer to the upper left corner, indicating that it can still maintain a high recognition rate at a low false alarm rate. The AUC of the CNN model is 0.904, the AUC of the BiLSTM model is 0.918, the CNN-bilSTM model is improved to 0.943, and the AUC of the proposed model reaches 0.972. Compared with CNN-BiLSTM, the AUC of the model in this paper is increased by 0.029, indicating that the attention mechanism can further strengthen the recognition ability of key learning behavior segments, and has a better distinguishing effect on low participation and high risk student states.

5.2 Evaluation of learning state transfer effect after personalized learning support

The effectiveness of personalized learning support not only depends on whether the resource recommendation results match the current state of students, but also depends on whether the system can adjust the support strategy in time according to the change of learning state. Based on the results of student behavior recognition, this paper divides the learning state into five categories: stable investment, mild fluctuation, weak knowledge, low participation and high risk. The transfer between different states after the recommendation feedback is counted to evaluate the promotion effect of the system on the improvement of learning state. The student learning state transition probability is shown in Figure 4.

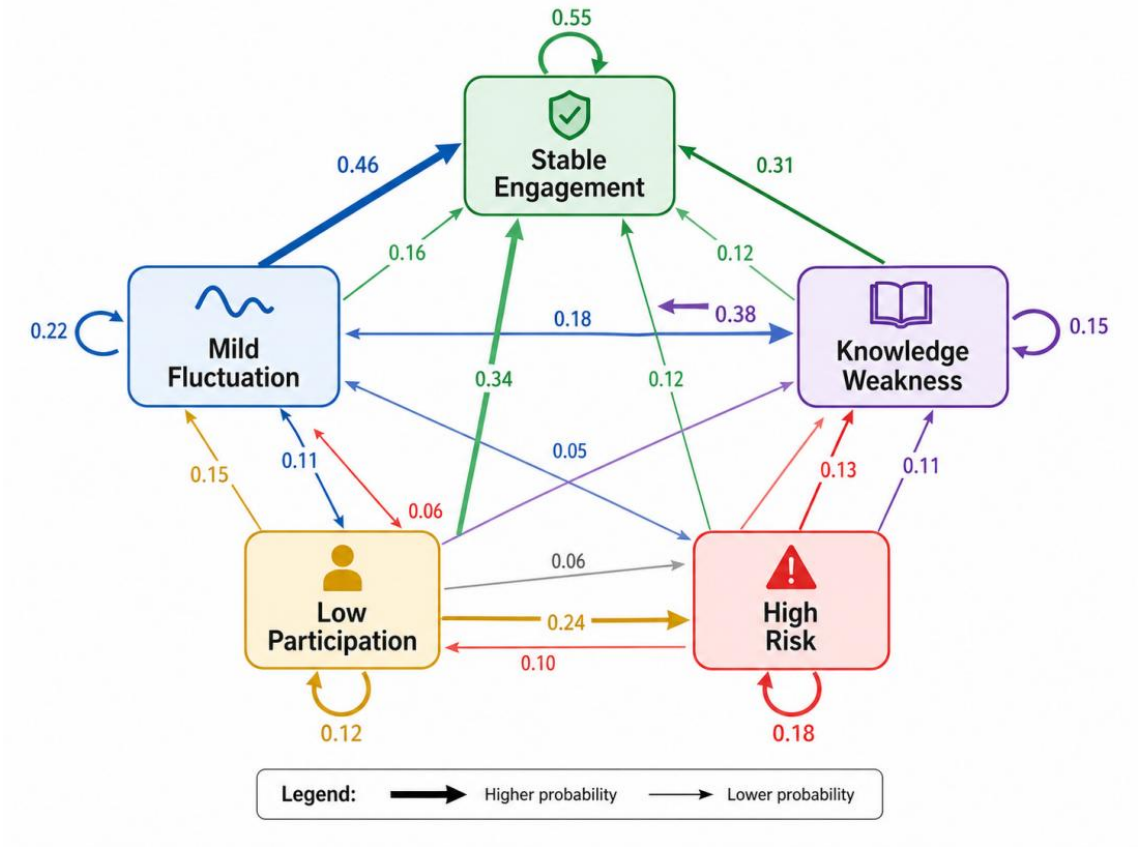

Figure 4: Student learning state transition probability diagram

It can be seen from Figure 4 that after personalized learning support, the overall learning state of students shows a trend of migration from a low-level state to a stable state. The probability of the mild fluctuation state turning to the stable input state reached 0.46, and the probability of the knowledge weak state turning to the mild fluctuation state and the stable input state were 0.38 and 0.31, respectively, indicating that the fault cause analysis, basic exercises and resource supplement could alleviate the knowledge gap. The probability that the low participation state turned to the stable engagement state was 0.34, and the probability that the high risk state continued to remain high risk was reduced to 0.18, indicating that the micro-class resources, interactive tasks and teacher intervention tips recommended by the system could effectively reduce the risk of persistent stragging-behind and enhance the practical effect of personalized learning support. At the same time, the system recommendation accuracy reaches 92.6%, indicating that the resource push mechanism based on learning state prediction and student profile matching can better adapt to the individual differences of students.

5.3 Comparative analysis of multimodal fusion model and baseline model

In order to further verify the comprehensive advantages of the multimodal fusion model, this paper selects LightGBM, CNN, BiLSTM and CNN-bilSTM as the baseline models, and compares them from three indicators: Accuracy, F1 and AUC. LightGBM mainly relies on manually constructed features and can handle structured logs and evaluation data, but its ability to depict visual behavior, text semantics and temporal changes is limited. CNN can extract local behavior combination features, but it is difficult to fully model the continuous evolution of students 'learning states. BiLSTM can capture the temporal dependence, but lacks the fine-grained extraction of local feature patterns. CNN-BiLSTM performs well in

combining local features and temporal features, but the attention to key time slices and key modalities is still not prominent enough. The comparison of the comprehensive performance of different models is shown in Figure 5.

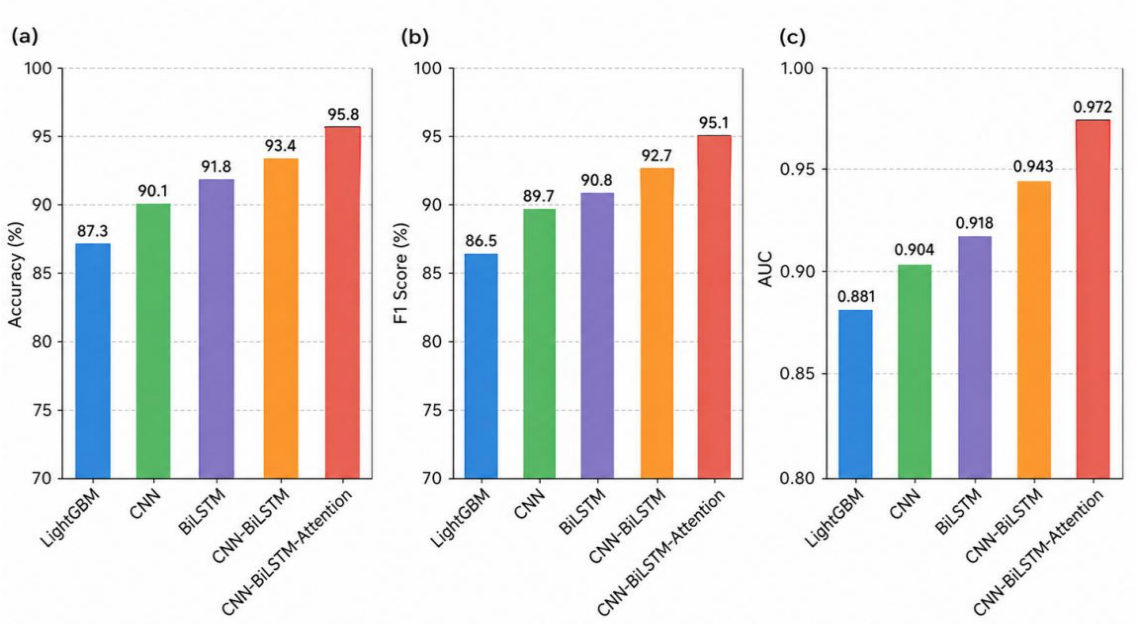

Figure 5: Bar charts comparing the comprehensive performance of different models

It can be seen from Figure 5 that the CNN-BiLSTM-Attention model in this paper achieves the best results on the three indicators, with Accuracy reaching 95.8%, F1 reaching 95.1%, and AUC reaching 0.972. Compared with the LightGBM model, the Accuracy is increased by 8.5 percentage points, F1 is increased by 8.6 percentage points, and AUC is increased by 0.091. Compared with the CNN-BiLSTM model, the Accuracy is increased by 2.4 percentage points, F1 is increased by 2.4 percentage points, and AUC is increased by 0.029. The results show that multi-modal feature fusion can enhance the integrity of student behavior expression, the combination of CNN and BiLSTM improves the ability of local pattern extraction and temporal dependence modeling, and the Attention mechanism further strengthens the contribution of key learning behavior segments, making the model more stable in learning state recognition and risk discrimination.

5.4 Model stability generalization ability and feasibility analysis of real-time deployment

In order to verify the robustness and engineering landing ability of the model in complex teaching scenarios, this paper carries out experimental analysis from two perspectives of modal disturbance stability and deployment real-time performance. In the stability test, missing, noise and random disturbance are injected into the log modality, visual modality, text modality and evaluation modality respectively, and the changes in Accuracy, F1 and AUC indicators of the model are calculated to evaluate the adaptability of the multimodal fusion model to incomplete data and abnormal input. The model performance changes under different modal disturbance conditions are shown in Figure 6.

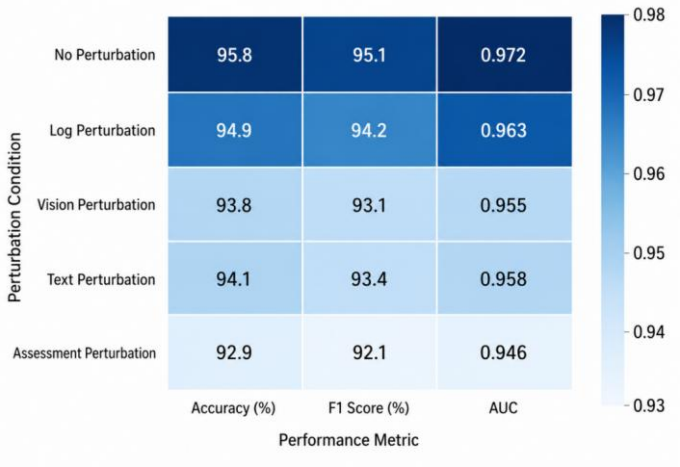

Figure 6: Heat map of model performance variation under different modal disturbance conditions

It can be seen from Figure 6 that under the condition of no disturbance, the model Accuracy, F1 and AUC reach 95.8%, 95.1% and 0.972, respectively. After adding log perturbation, the three indexes decreased to 94.9%, 94.2% and 0.963, respectively. The decrease was more obvious under visual disturbance, which were 93.8%, 93.1% and 0.955, respectively. After text perturbation, it is 94.1%, 93.4% and 0.958; The evaluation modal perturbation had the largest impact, with Accuracy reduced to 92.9%, F1 reduced to 92.1%, and AUC reduced to 0.946, indicating that the evaluation and knowledge mastery information had a stronger supporting effect on the recognition of learning states. In general, the main indicators of the model remain above 92% under various disturbances, indicating that the constructed multimodal fusion mechanism has good stability and cross-scene generalization ability.

In addition to identification stability, whether the system can balance between recommendation effect and response efficiency is also the focus of deployment feasibility analysis. This paper further compares the trade-off relationship between the recommendation accuracy and the average response time of typical models, and plots the Pareto front. Figure 7 shows the relationship between the recommendation accuracy of different models and the system response delay.

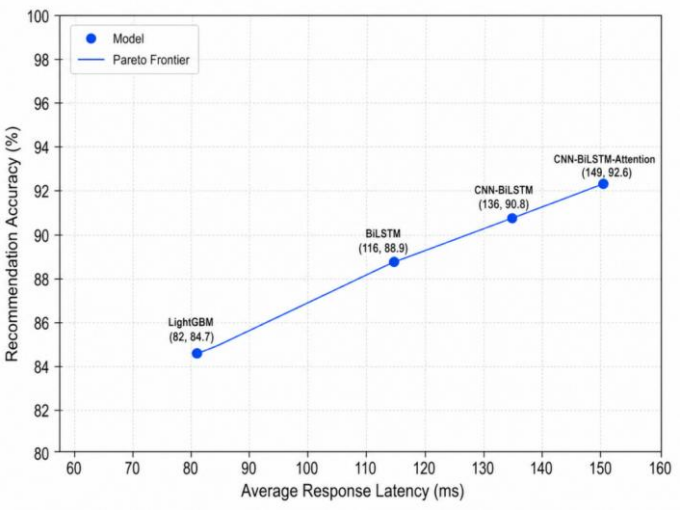

Figure 7: Pareto front plot of recommendation accuracy and system response delay

It can be seen from Figure 7 that the average response time of LightGBM model is the lowest, only 82 ms, but the recommendation accuracy is only 84.7%. The recommendation accuracy of BiLSTM model was improved to 88.9%, and the response delay was increased to 116 ms. The CNN-BiLSTM model can achieve 90.8% recommendation accuracy at 136 ms. Although the response delay of CNN-BiLSTM-Attention model is 149 ms, which is slightly higher than that of some baseline models, the recommendation accuracy of CNN-bilSTM-attention model reaches 92.6%, which is located in the Pareto front optimal region, indicating that it can still meet the real-time response requirements of online learning support system while maintaining high recommendation quality. Based on Figure 6 and Figure 7, it can be seen that the proposed model has strong robustness in the modal disturbance environment, and it also takes into account the recognition accuracy, recommendation effect and operation efficiency in the actual deployment, which has good engineering application value.

6 Discussion

Multimodal data fusion provides a more complete computing foundation for the analysis of college students 'behavior than single log data. The log of learning management platform can reflect students 'login frequency, resource access path, page stay time and assignment submission rhythm. Classroom visual data can supplement students 'concentration state, posture changes and participation performance. Different modalities form students' learning portraits from the behavioral, cognitive and emotional levels, which makes the system have stronger discrimination ability in identifying low participation, high risk and weak knowledge states. Experimental results show that the model combined with CNN, BiLSTM and Attention mechanism is better than the baseline model in terms of Accuracy, F1 and AUC, indicating that local behavior feature extraction, bidirectional temporal dependence modeling and key segment weighting have a good complementary effect. CNN can identify the behavior combination pattern in a short window, BiLSTM can track the continuous changes of learning states, and Attention mechanism can highlight key modes and key time slices, making student state recognition closer to the real learning process.

From the perspective of personalized learning support, the value of the system is not only to judge the current state of students, but also to transform the recognition results into executable learning support strategies. The experimental results showed that after personalized feedback, students with mild fluctuations, weak knowledge and low participation tended to transfer to a stable state of engagement, indicating that resource recommendation, error cause analysis, interactive tasks and teacher intervention tips could have a positive effect on the improvement of learning status. Compared with the traditional way of pushing learning resources, the support mechanism based on student portrait and state prediction can more accurately match the current needs of students. For students with weak knowledge, the system preferentially pushed basic exercises, error cause analysis and similar problem groups. For low participation students, the system focused on interactive tasks, discussion topics and collaborative learning resources. For high risk students, a remedy path and a teacher-side warning are generated. This state-driven feedback logic can reduce invalid recommendations and improve the pertinence of the supported content on the student side.

The model stability experiment further shows that multimodal fusion does not simply depend on a certain kind of data. When the log, visual, text and evaluation modalities are disturbed, the main indicators of the model remain at a high level, indicating that there is a certain information compensation ability between different modalities. In contrast, the performance degradation caused by the evaluation mode perturbation is more obvious,

indicating that knowledge mastery, error type and stage evaluation results are still important bases for judging learning risk. In the subsequent system design, the association modeling between knowledge graph, error network and course objectives can be further strengthened, so that the learning state recognition not only stays at the behavior surface level, but also goes deep into the knowledge structure level, so as to improve the explanation depth of personalized learning support.

From the perspective of deployment feasibility, although the computational complexity of CNN-BiLSTM-Attention model is higher than that of light-weight models such as LightGBM, its response time is still within the acceptable range of online learning support systems, which can meet the real-time requirements of teacher-end monitoring, student-end recommendation and learning risk warning. Considering the characteristics of concurrent access, long course cycle and continuous data update in university teaching scenarios, the system operation overhead can be reduced by model pruning, feature caching, edge reasoning and incremental update in the future. At the same time, learning behavior data involves privacy protection and algorithm fairness, and the system needs to maintain constraints on data desensitization, permission control, interpretable feedback and manual review mechanisms to avoid model judgment being directly equivalent to student ability evaluation. In general, the method in this paper forms a relatively stable technical link between recognition accuracy, recommendation support and system deployment, but it still needs to be improved in cross-school data verification, long-term tracking experiment and ethical governance mechanism.

7 Conclusion

Focusing on the needs of college students 'behavior analysis and personalized learning support, this paper proposes an intelligent system framework based on multimodal data fusion. The system converts heterogeneous data such as learning log, classroom vision, discussion text, evaluation performance and resource access into a unified behavior representation, and uses the CNN-BiLSTM-Attention model to complete learning state recognition and risk judgment. Experimental results show that the proposed model outperforms baseline methods such as LightGBM, CNN, BiLSTM and CNN-bilSTM in action recognition tasks, with Accuracy, F1 and AUC reaching 95.8%, 95.1%and 0.972, respectively. Under the condition of modal perturbation, the main index still remains above 92%, which indicates that the model has good stability and generalization ability. The personalized support experiment showed that the probability of students with mild fluctuations turning to a stable investment state reached 0.46, and the probability of maintaining a high-risk state decreased to 0.18, indicating that resource recommendation and teacher-side intervention tips could improve the learning state. In the future, knowledge graph, edge reasoning and cross-school data verification can be further combined to improve the system interpretation ability, deployment efficiency and adaptability of educational scenarios.

Author's Profile

Linglan Gao was born in Fuzhou, Fujian, P.R. China, in 1998. She received a master's degree from The Education University of Hong Kong. She is currently working at the Institute of Education, Fuzhou University of International Studies and Trade. Her primary research focus is Higher Education.

References

[1] Tahiru F. AI in Education: A Systematic Literature Review[J]. Journal of Cases on Information Technology, 2021, 23(1): 1-20. DOI: 10.4018/JCIT.2021010101.

[2] Crompton H, Burke D. Artificial intelligence in higher education: the state of the field[J]. International Journal of Educational Technology in Higher Education, 2023, 20: 22. DOI: 10.1186/s41239-023-00392-8.

[3] Hardaker G, Glenn L E. Artificial intelligence for personalized learning: a systematic literature review[J]. International Journal of Information and Learning Technology, 2025, 42(1): 1-14. DOI: 10.1108/IJILT-07-2024-0160.

[4] Giannakos M, Cukurova M. The role of learning theory in multimodal learning analytics[J]. British Journal of Educational Technology, 2023, 54(5): 1246-1267. DOI: 10.1111/bjet.13320.

[5] Worsley M, Martinez-Maldonado R, D'Angelo C. A New Era in Multimodal Learning Analytics: Twelve Core Commitments to Ground and Grow MMLA[J]. Journal of Learning Analytics, 2021, 8(3): 10-27. DOI: 10.18608/jla.2021.7361.

[6] Dubovi I. Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology[J]. Computers & Education, 2022, 183: 104495. DOI: 10.1016/j.compedu.2022.104495.

[7] Sghir N, Adadi A, Lahmer M. Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022)[J]. Education and Information Technologies, 2023, 28: 8299-8333. DOI: 10.1007/s10639-022-11536-0.

[8] Cerezo R, Lara J A, Azevedo R, Romero C. Reviewing the differences between learning analytics and educational data mining: Towards educational data science[J]. Computers in Human Behavior, 2024, 154: 108155. DOI: 10.1016/j.chb.2024.108155.

[9] Khosravi H, Shabaninejad S, Bakharia A, Sadiq S, Indulska M, Gašević D. Intelligent Learning Analytics Dashboards: Automated Drill-Down Recommendations to Support Teacher Data Exploration[J]. Journal of Learning Analytics, 2021, 8(3): 133-154. DOI: 10.18608/jla.2021.7279.

[10] Rets I, Herodotou C, Gillespie A. Six Practical Recommendations Enabling Ethical Use of Predictive Learning Analytics in Distance Education[J]. Journal of Learning Analytics, 2023, 10(1): 149-167. DOI: 10.18608/jla.2023.7743.

[11] Yusuf A, Noor N M, Bello S. Using multimodal learning analytics to model students' learning behavior in animated programming classroom[J]. Education and Information Technologies, 2024, 29(6): 6947-6990. DOI: 10.1007/s10639-023-12079-8.

[12] Corza-Vargas V M, Martinez-Maldonado R, Escalante-Ramírez B, Olveres J. Students' Ethical, Privacy, Design, and Cultural Perspectives on Visualizing Cognitive-Affective States in Online Learning[J]. Journal of Learning Analytics, 2024, 11(3): 24-40. DOI: 10.18608/jla.2024.8483.

[13] Fazil M, Rísquez A, Halpin C. A Novel Deep Learning Model for Student Performance Prediction Using Engagement Data[J]. Journal of Learning Analytics, 2024, 11(2): 23-41. DOI: 10.18608/jla.2024.7985.

[14] Khenkar S G, Jarraya S K, Allinjawi A, Alkhuraiji S, Abuzinadah N, Kateb F A. Deep Analysis of Student Body Activities to Detect Engagement State in E-Learning Sessions[J]. Applied Sciences, 2023, 13(4): 2591. DOI: 10.3390/app13042591.

[15] Trakunphutthirak R, Lee V C S. Application of Educational Data Mining Approach for Student Academic Performance Prediction Using Progressive Temporal Data[J]. Journal of Educational Computing Research, 2022, 60(3): 742-776. DOI: 10.1177/07356331211048777.

[16] Alamri R, Alharbi B. Explainable Student Performance Prediction Models: A Systematic Review[J]. IEEE Access, 2021, 9: 33132-33143. DOI: 10.1109/ACCESS.2021.3061368.

[17] Albreiki B, Zaki N, Alashwal H. A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques[J]. Education Sciences, 2021, 11(9): 552. DOI: 10.3390/educsci11090552.

[18] Rodríguez-Hernández C F, Musso M, Kyndt E, Cascallar E. Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation[J]. Computers and Education: Artificial Intelligence, 2021, 2: 100018. DOI: 10.1016/j.caeai.2021.100018.

[19] Bhardwaj P, Gupta P K, Panwar H, Siddiqui M K, Morales-Menendez R, Bhaik A. Application of Deep Learning on Student Engagement in E-learning Environments[J]. Computers & Electrical Engineering, 2021, 93: 107277. DOI: 10.1016/j.compeleceng.2021.107277.

[20] Gupta S, Kumar P, Tekchandani R K. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models[J]. Multimedia Tools and Applications, 2023, 82(8): 11365-11394. DOI: 10.1007/s11042-022-13558-9.

[21] Alnasyan B, Basheri M, Alassafi M. The power of Deep Learning techniques for predicting student performance in Virtual Learning Environments: A systematic literature review[J]. Computers and Education: Artificial Intelligence, 2024, 6: 100231. DOI: 10.1016/j.caeai.2024.100231.

[22] Kaouni M, Lakrami F, Labouidya O. The Design of An Adaptive E-learning Model Based on Artificial Intelligence for Enhancing Online Teaching[J]. International Journal of Emerging Technologies in Learning, 2023, 18(06): 202-219. DOI: 10.3991/ijet.v18i06.35839.

[23] Al-Zahrani A M, Alasmari T. Learning Analytics for Data-Driven Decision Making: Enhancing Instructional Personalization and Student Engagement in Online Higher Education[J]. International Journal of Online Pedagogy and Course Design, 2023, 13(1): 1-18. DOI: 10.4018/IJOPCD.331751.

[24] Saleem R, Aslam M. A Multi-Faceted Deep Learning Approach for Student Engagement Insights and Adaptive Content Recommendations[J]. IEEE Access, 2025, 13: 69236-69256. DOI: 10.1109/ACCESS.2025.3561459.

[25] Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning[J]. Electronic Markets, 2021, 31(3): 685-695. DOI: 10.1007/s12525-021-00475-2.

[26] Sarker I H. Machine learning: algorithms, real-world applications and research directions[J]. SN Computer Science, 2021, 2(3): 160. DOI: 10.1007/s42979-021-00592-x.

[27] Prabowo H, Hidayat A A, Cenggoro T W, Rahutomo R, Purwandari K, Pardamean B. Aggregating Time Series and Tabular Data in Deep Learning Model for University Students' GPA Prediction[J]. IEEE Access, 2021, 9: 87370-87377. DOI: 10.1109/ACCESS.2021.3088152.