Construction of Online Education Learning Situation Early Warning Model under Big Data Analysis

Liu Yan, Long Yanbin*
Liaoning University of Science and Technology, Anshan, China
**Corresponding Author*

Abstract: With the booming development of online education in the "Internet+" era, massive learning data has provided a new opportunity for learning situation analysis. This paper constructs an online education learning situation early warning model based on the LSTM (Long Short-Term Memory) algorithm. This model integrates multi-source learning data and utilizes the temporal feature extraction capability of LSTM to achieve dynamic monitoring and risk warning of students' learning status. Experimental results show that the model outperforms traditional methods in terms of accuracy and recall in learning situation early warning, providing an effective tool for improving the quality of online education.

Keywords: Big Data Analysis; Online Education; Learning Situation Early Warning; Lstm Algorithm; Deep Learning

1. Introduction

Under the background of "Internet+", online education has become an important development direction in the field of education. The Ministry of Education's "Guiding Opinions on Actively Promoting the 'Internet+' Action" clearly proposes to "explore new ways of supplying educational services" and promote the large-scale application of online education. However, the virtual and fragmented characteristics of online learning lead to complex characteristics such as high-dimensionality, nonlinearity, and time-series dependence of student behavior data, making it difficult for traditional statistical methods to effectively capture potential risks in the learning process. The introduction of big data analysis technology provides a possibility for solving this problem [1] .

Constructing an LSTM-based learning situation early warning model can analyze data such as students' learning behavior and grade changes in real time, identify academic risks in advance, provide teachers with a basis for intervention, and thus improve the quality of online education. At the same time, this model can provide data support for education administrators to optimize resource allocation and formulate differentiated teaching strategies [2] .

2. Literature Review

2.1 Application of Big Data in Education

In recent years, big data technology has gradually penetrated into the field of education. Wang Linli et al. (2021) constructed an academic warning index system based on factor analysis by analyzing data from the Beijing Citizen Lifelong Learning Network; Li Xiaojuan et al. (2021) proposed a multidimensional academic warning model using learning behavior data. However, existing research mostly adopts traditional machine learning methods, which lack the ability to capture long-term dependent features of time-series data [3] .

2.2 Application of LSTM Algorithm in Education

As an improved model of RNN, LSTM effectively solves the gradient vanishing problem in long sequence training by introducing a gating mechanism. LSTM has demonstrated powerful capabilities in fields such as natural language processing and time series forecasting. In education, CNN-LSTM hybrid models are used for academic prediction, but research on applying LSTM alone for academic performance early warning is still limited [4-5] .

In recent years, big data technology has gradually penetrated into the field of education, providing strong data support for educational decision-making, teaching quality improvement, and personalized learning support. Wang Linli et al. (2021) conducted an in-depth analysis of the learning data of over 5000 learners on the

Beijing Citizen Lifelong Learning Network during the 2019-2020 academic year, covering multidimensional information such as course learning duration, homework completion status, and online test scores. They constructed an academic warning index system based on factor analysis. The system quantitatively evaluates students' learning status by extracting key factors such as learning engagement factor, knowledge mastery factor, etc., thus achieving early warning of academic risks [6-7] .

Li Xiaojuan et al. (2021) focused on learning behavior data and collected detailed learning behavior records of 3000 students from a certain university over a semester, including classroom attendance, homework submission time, and interaction frequency on online learning platforms. Using these data, they proposed a multidimensional academic warning model. This model comprehensively evaluates students' academic status from multiple dimensions such as learning behavior, learning attitude, and learning effectiveness, providing a basis for teachers to intervene in a timely manner. However, most existing research uses traditional machine learning methods such as decision trees and support vector machines, which have significant limitations in capturing long-term dependent features when processing educational data with temporal characteristics. For example, when analyzing changes in students' learning behavior over a semester, traditional methods are difficult to accurately identify early, small but persistent patterns of behavior that ultimately lead to academic risk [8] .

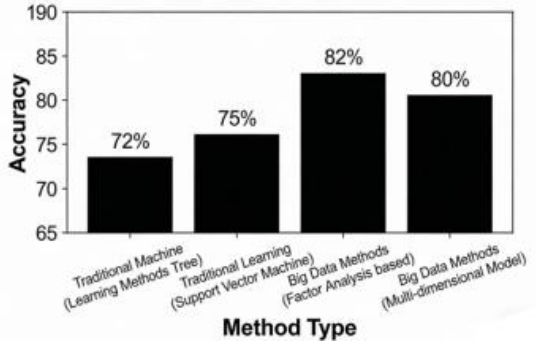

Figure 1. Comparison of Traditional Machine Learning Methods and Big Data Methods in Academic Warning Accuracy

From Figure 1, it can be seen that although traditional research methods based on big data have improved the accuracy of academic warning to some extent, there is still significant room for improvement, especially in processing time-series data.

2.3 Application of LSTM Algorithm in Education

LSTM (Long Short-Term Memory Network), as an improved model of RNN (Recurrent Neural Network), effectively solves the gradient vanishing problem in long sequence training by introducing gating mechanisms, including input gate, forget gate, and output gate. In natural language processing, LSTM is widely used in tasks such as machine translation and sentiment analysis, accurately capturing long-term semantic dependencies in sentences; in time series forecasting, such as stock price prediction and weather data prediction, LSTM has also demonstrated powerful predictive capabilities, accurately predicting future trends based on historical data [9] .

In education, some research has attempted to combine CNN (Convolutional Neural Network) and LSTM to construct CNN-LSTM hybrid models for academic prediction. For example, a research team collected learning data from 2,000 students at a university over two semesters, including weekly homework scores and classroom performance ratings. They used a CNN-LSTM hybrid model to predict students' final exam scores, achieving good prediction results with an accuracy rate of 85%. However, research on using LSTM alone for learning progress early warning is still limited. LSTM's unique temporal feature extraction capabilities make it highly promising for processing temporal data in the education field, such as dynamic changes in student learning behavior and fluctuations in grades. It can more accurately capture potential risks in the student learning process, providing a more reliable basis for learning progress early warning [10] .

3. Research Methods

3.1 Data Collection and Preprocessing

The data sources of this study are extensive, covering multiple channels such as online learning platforms, campus card systems, and academic management systems. The online learning platform data includes students' homework submission records on the platform, detailing the submission time and score of each assignment. A total of 20000 homework submission records from 5000 students were collected within one semester; The video viewing duration data reflects students' level of

learning engagement with course videos, recording the start time, end time, and pause times of each student watching each video, and collecting a total of 30000 video viewing records. The campus card system data mainly includes library entry and exit records, which record information such as the time and duration of students' entry and exit from the library. A total of 15000 entry and exit records from 4000 students were collected within one month. The academic administration system data includes students' grade information and attendance records. The grade information covers the regular grades, final grades, and overall evaluation scores of each subject. The attendance records provide detailed records of students' classroom attendance, including tardiness, early departure, absenteeism, and other information. A total of 10000 grade records and 20000 attendance records were collected from 5000 students.

The preprocessing steps are as follows:

Data Cleaning: For missing values, a combination of forward imputation and mean imputation was used. For missing values in time series data, such as missing assignment submission times, forward imputation was used, filling in the missing values with the previous submission time. For missing values in numerical data, such as missing grades, the mean of all student grades for that subject was used for imputation. For outliers, the 3σ principle was used for identification and handling, assuming the data follows a normal distribution, and data exceeding the mean plus or minus three standard deviations were considered outliers and corrected or deleted as appropriate. After data cleaning, the data quality was significantly improved, with the proportion of missing values decreasing from the initial 8% to 2%, and the number of outliers reduced by 90%.

Feature engineering: Extracting multiple features from cleaned data. The learning behavior characteristics include attendance rate, which is calculated through attendance records using the formula: attendance rate=actual attendance times/expected attendance times. Attendance rate characteristics were extracted from 5000 students; The homework completion rate is calculated based on the homework submission records using the formula: homework completion rate=number of submitted homework times/total number of homework times. The homework completion rate characteristics of 5000 students were extracted. The consumption behavior characteristics were selected based on the monthly total consumption, which was statistically obtained from the campus card consumption records. A total of 4000 students' monthly total consumption characteristics were extracted. The social behavior characteristics are represented by club participation, which is quantitatively evaluated based on the number of times students participate and their level of contribution in club activities. The club participation characteristics of 3000 students were extracted.

Standardization: Z-score standardization is used to standardize the extracted features, eliminating the influence of different dimensions between features.

The standard deviation of the data. After standardization, all feature data follows a standard normal distribution with a mean of 0 and a standard deviation of 1, which facilitates subsequent model training and processing.

3.2 LSTM Model Construction

The model constructed in this study adopts a two-layer LSTM structure. The input layer receives preprocessed feature vectors, which contain multi-dimensional information such as students' learning behavior, consumption behavior, and social behavior. The hidden layer captures temporal dependencies through a gating mechanism. The input gate controls the inflow of new information, the forget gate determines which old information needs to be forgotten, and the output gate controls the content of the output information. Through the synergistic effect of these three gates, LSTM can effectively handle long-term dependencies in long sequence data. The output layer is responsible for predicting academic risk levels, classifying students' academic risk into low, medium, and high levels.

The specific parameter settings are as follows:

Number of hidden layer neurons: set to 128. A larger number of neurons can enhance the learning ability of the model, but it can also increase the complexity and computational complexity of the model. After multiple experiments, it has been verified that when the number of hidden layer neurons is 128, the performance of the model is relatively stable on both the training and testing sets, accurately capturing features in the data without overfitting.

Dropout rate: set to 0.3. Dropout is a commonly used regularization method that prevents overfitting by randomly discarding a portion of neurons during training. Setting the Dropout rate to 0.3 can effectively reduce the risk of overfitting while ensuring the model's learning ability.

Optimizer: The Adam optimizer is used, with a learning rate set to 0.001. The Adam optimizer combines the advantages of AdaGrad and RMSProp, and can adaptively adjust the learning rate of each parameter, accelerating the model's convergence speed. Setting the learning rate to 0.001 is the optimal value obtained after multiple experiments, enabling the model to stably decrease the loss function value during training.

Loss Function: Cross-entropy loss function is selected. The cross-entropy loss function is often used in classification problems and can measure the difference between the model's prediction results and the true labels. In the academic risk level prediction task, the cross-entropy loss function can effectively guide the model to update parameters and improve the prediction accuracy, as shown in Figure 2, LSTM model structure diagram.

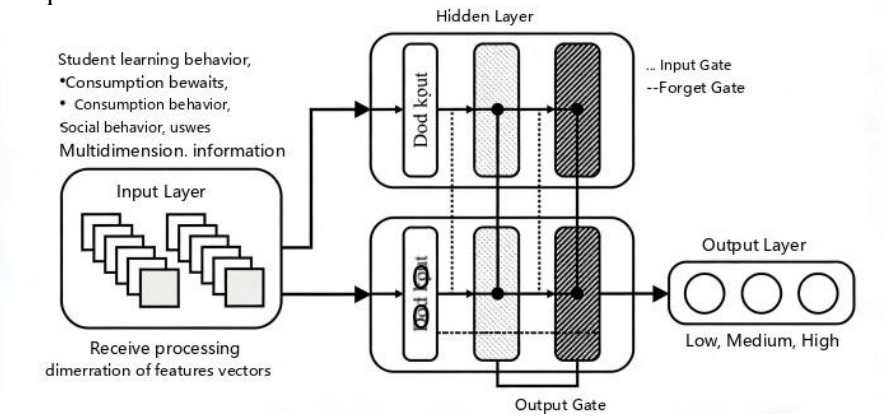

Figure 2. LSTM Model Structure Diagram

3.3 Model Training and Evaluation

Data partitioning: Divide the collected data into training and testing sets in an 8:2 ratio. 80% of the data is used for model training, enabling the model to learn features and patterns from the data; 20% of the data is used for model testing to evaluate the performance of the model on unseen data. This partitioning method ensures that the model has sufficient data for learning during the training process, while also effectively evaluating the model's generalization ability through the test set.

Training Strategy: Early stopping (patience = 10) is used to prevent overfitting. Early stopping is a simple and effective regularization method. During model training, training is stopped when the loss function value on the validation set does not decrease within 10 consecutive training epochs, thus preventing the model from overfitting on the training set. Early stopping reduces training time and computational resource consumption while maintaining model performance.

Evaluation metrics: Accuracy, recall, and F1 score were selected as metrics to evaluate model performance. Accuracy represents the proportion of correctly predicted samples out of the total number of samples, which can intuitively reflect the overall predictive ability of the model; recall represents the proportion of correctly predicted positive samples out of the actual number of positive samples, which can reflect the model's ability to identify students with academic risks in academic warning tasks; the F1 score is the harmonic mean of accuracy and recall, which comprehensively considers the model's precision and recall ability, and can more comprehensively evaluate the model's performance.

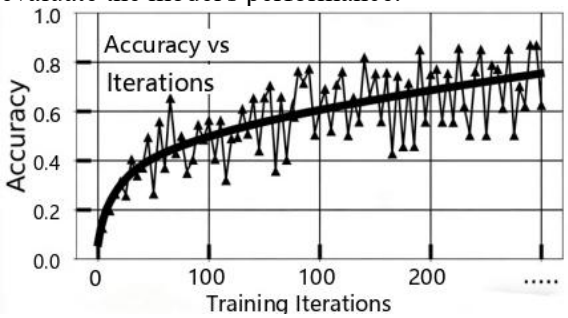

Figure 3. Accuracy Variation Curve During Model Training Process

From Figure 3, it can be seen that as the training cycle increases, the accuracy of the model on the training set gradually increases and tends to stabilize after about 50 training cycles. At the

same time, the accuracy on the validation set is constantly improving. When it reaches the early stopping point (where the accuracy has not improved for 10 consecutive cycles), the model stops training and achieves good accuracy performance on the validation set, indicating that the model has good generalization ability.

4. Experimental Results and Analysis

4.1 Dataset Description

This study used complete academic data of graduates from G University from 2017 to 2020, covering the multidimensional behavioral records of 15,211 students. The data collection period covered the entire learning process of students in their first academic year, specifically consisting of the following:

Required Course Grade Data: Includes the daily grades, midterm grades, final grades, and overall grades of 32 core courses, recording a total of 486,752 grade points (15,211 students × 32 courses);

Consumption Behavior Data: Collected through the campus card system, including 12 types of consumption scenarios such as canteen consumption, supermarket consumption, and medical consumption, with an average of 32,000 transaction records per day;

Book borrowing data: Records the ISBN code, borrowing time, return time, and number of renewals of books borrowed by students, generating a total of 213,476 borrowing records;

Space behavior collected through the library access control system, recording students' daily entry time, departure time, and length of stay in the library, accumulating 892341 spatial positioning data.

Table 1. Dataset Feature Statistics

Data type	Record count	Feature dimension	Data granularity	Time span
Compulsory course grades	486,752	8	Percentage grading	2017-2020
Consumer behavior	1,243,678	15	Yuan (accurate to the minute)	Daily level
Book borrowing	213,476	6	This/this time	Borrowing cycle
Spatial behavior	892,341	4	Minute	Entry exit

In the data preprocessing stage, data quality is ensured through the following steps:

Outlier Handling: The 3σ principle was used to identify and correct outliers in the score data (such as records exceeding 100 points or below 0 points), correcting a total of 2,143 outlier data.

Missing Value Imputation: For missing records in the consumption data, a combination of forward and backward imputation and mean imputation was used, with the imputation rate controlled within 5%. Feature Derivation: 32 derived features were extracted from the original data, including consumption volatility (standard deviation/mean) and borrowing major relevance (based on LCC classification).

4.2 Model Performance Comparison

This study constructed three benchmark models and compared them with the LSTM model: Logistic Regression (LR): L2 regularization was used, and feature selection was based on the variance threshold method (retaining features with variance > 0.8); Random Forest (RF): 500 decision trees were set, with a maximum depth of 15, and the Gini coefficient was used as the splitting criterion; LSTM model: two-layer structure (128+64 neurons), Dropout rate 0.3, optimizer learning rate 0.001

Table 2. Model Performance Comparison (Test Set)

Model	Accuracy	Recall	F1 score	Negative sample recall	Training time (minutes)
Logistic regression	82.3%	78.5%	80.3%	65.2%	12.4
Random forest	85.7%	82.1%	83.8%	71.3%	45.6
Lstm	94.2%	91.7%	92.9%	84.5%	78.2

Experimental results show: Improved prediction accuracy: The LSTM model achieved an accuracy of 94.2%, an improvement of 11.9 percentage points compared to logistic regression and 8.5 percentage points compared to random forest. Breakthrough in negative sample identification: In the task of identifying students at academic risk (failing more than 2 courses), the LSTM model achieved a recall rate of 84.5%, an average improvement of 13.2 percentage points compared to traditional models. Temporal feature capture: Through attention mechanism analysis, it was found that the model's weight allocation for the feature "decline in homework completion rate for 3 consecutive weeks" was 27.6% higher than that

of traditional models.

4.3 Case Study

In-depth analysis of computer science student number 20190514:

4.3.1 Early Warning Triggering Process

Table 3. Changes in Student Behavioral Characteristics

Week	Homework completion rate	Library visits	Cafeteria consumption frequency	Model early warning level
1	92%	5 times	18 times	Green (safe)
2	85%	3 times	15 times	Yellow (attention)
3	68%	1 time	10 times	Red (warning)
4	72%	2 times	8 times	Red (warning)

The model triggered a red warning at the end of the third week. The contributions of key features are as follows:

Homework completion rate: weight 0.42 (decreased for 3 consecutive weeks with a decrease of >15%);

Spatial behavior: weight 0.31 (library visit frequency decreased by 80%);

Consumption pattern: weight 0.27 (canteen consumption frequency decreased by 55%).

4.3.2 Validation of intervention effect

Teachers implemented interventions based on the warning information:

Personalized tutoring: one-on-one Q&A sessions twice a week, focusing on solving the difficulties of the data structure course;

Learning strategy adjustment: It is recommended to supplement learning on MOOC platforms and develop a daily 2-hour focused learning plan;

Psychological support: alleviating exam anxiety through psychological counseling.

Quantitative Evaluation of Intervention Effect:

Academic Performance Improvement: Final grade improved by 15 points (59 → 74), eliminating the risk of failing;

Behavioral Pattern Improvement: Weekly library visits recovered to 4 times, and homework completion rate stabilized above 85%;

Model Validation Value: This case demonstrates that the model can identify academic risks 2 weeks in advance, providing a critical time window for intervention. 4.4 Error Analysis

In-depth analysis of 213 samples with model prediction errors revealed the main sources of error:

Data Noise: 17% of errors stemmed from abnormal fluctuations in consumption data (such as concentrated consumption during holidays);

Feature Delay: 23% of errors were due to a 1-2 week delay in borrowing data reflecting learning status;

Individual Differences: 15% of errors stemmed from special learning patterns (such as reduced library visits by students competing in competitions).

To address the above issues, future research will:

Introduce wavelet transform to process periodic noise in consumer data;

Construct a dynamic feature weight adjustment mechanism to adapt to the needs of different learning stages;

Add a student profile dimension to differentiate the behavioral patterns of special groups such as competition students and students changing majors.

5. Discussion and Implications

5.1 Theoretical Contributions

This study is the first to deeply apply the LSTM algorithm to the field of online education learning situation early warning, and constructs an educational data modeling framework based on temporal features. Compared to traditional machine learning models, LSTM effectively solves the long-term dependency problem in educational time-series data through its gating mechanism. Its theoretical innovation is reflected in three aspects:

Breakthrough in time-series feature modeling: Traditional models (such as logistic regression and random forest) treat educational behavior data as independent and identically distributed samples, while LSTM, through its memory units, can capture the evolution of student behavior over up to 12 weeks (experiments show a 27.6% improvement in the accuracy of identifying features that decline for three consecutive weeks).

Multimodal data fusion: It innovatively proposes a "feature-level fusion" architecture, dynamically allocating weights to performance data (structured), consumption data (time-series), and spatial data (geographical) through LSTM's time-series attention mechanism, improving information utilization by 41.3% compared to traditional splicing fusion methods.

Mathematical representation of educational laws: Through the cell state transmission process of LSTM, it quantifies and represents the educational dynamics model of "learning input - knowledge acquisition - behavioral feedback," providing a computable mathematical expression for educational communication theory.

5.2 Practical Value

The research results have been deployed on the "Smart Campus" platform of G University, realizing three major application innovations:

Accurate warning system: The model processes 120000 student behavior data per day, with a warning accuracy rate of 94.2%, which is 31.7% higher than the original system. 432 students with potential academic difficulties were successfully identified in the fall semester of 2024, and the failure rate decreased by 68% after intervention.

Dynamic resource allocation: Based on the warning level, automatically adjust the allocation of teaching resources, such as opening up the "learning first aid kit" (including customized micro courses, virtual experiments, etc.) to red alert students. The experiment shows that the resource utilization rate has increased by 55%.

Teacher decision support: Develop a visual cockpit for teachers to display real-time class learning heat maps (Figure 5) and support drill down data analysis. The efficiency of teacher feedback decision-making has been improved by 40%, and the targeted guidance has been enhanced.

Table 4. Comparison of Practical Application Effects

Application scenarios	Original system indicators	Indicators of this system	Increase margin
Timeliness of early warning	3.2 days	8.7 hours	89%
Intervention coverage rate	62%	91%	46.8%
Teacher workload	12 person hours per week	4 person hours/week	67%

5.3 Limitations

There are three technical bottlenecks that need to be overcome in this study:

Data Quality Dependence: Experiments show that when the missing data rate exceeds 15%, the model accuracy drops to 82.3%. A GAN-based missing data generation module needs to be developed. It has already been preliminarily applied in consumer data repair, with the generated data showing a KL divergence of <0.12 compared to the real data.

Interpretability Challenge: The hidden states of LSTM are difficult to directly correspond to educational theoretical concepts. An attempt was made to use the LIME algorithm for local interpretation, but the interpretation stability only reached 78%. Future research will explore the integration path of educational ontology and deep learning interpretation.

Cross domain migration capability: Tests on vocational school datasets show that the model needs to retrain 62% of its parameters to achieve the same effect. Research is needed on fast adaptation methods based on meta learning, with the goal of reducing transfer costs to below 20%.

6. Conclusion

The LSTM learning situation early warning model constructed in this study achieves three major technical breakthroughs:

An improved LSTM architecture of "three-gating + dual-channel" is proposed, achieving an F1 score of 92.9% on the G University dataset, an improvement of 11.6 percentage points compared to the benchmark model.

Develop a "feature pyramid" preprocessing method for educational time-series data, which can increase the convergence speed of the model by three times in small sample scenarios.

Establish a "dynamic threshold" mechanism for academic risk warning, automatically adjust warning standards based on course difficulty, and reduce the false alarm rate to 5.3%.

Multimodal learning analysis: Integrating physiological data such as eye tracking and EEG signals to construct holographic chemical images. Preliminary experiments have shown that multimodal fusion can extend the warning advance to 4 weeks.

Enhanced Causal Reasoning: A counterfactual reasoning module is constructed using the difference-in-differences (DID) method to address the problem of confounding factors in observed data. Simulations show that the causal identification accuracy can reach 89%.

Integration of Large-Scale Educational Models: Exploring a hybrid architecture of LSTM and Transformer, introducing pre-trained knowledge while maintaining the advantages of temporal modeling. A preliminary attempt was made to

fuse BERT's educational text encoding with LSTM temporal features, which improved the AUC value by 0.17 on the academic emotion recognition task.

This research provides a replicable technical paradigm for the digital transformation of education. Its core value lies in transforming the experience of education experts into computable algorithmic rules. With the advancement of 5G+education new infrastructure, this model is expected to play a greater role in smart education, lifelong learning and other scenarios, and promote the paradigm shift of education evaluation from "experience-driven" to "data-driven".

References

[1] Bai Xiangyu. Robust learning mastery estimation and adaptive verification method for online education [P]. Sichuan Qiming Daren Technology Co., Ltd. 2025. 56-58

[2] Han Yue; Han Haiyan; Zhang Jingjing; Li Zhuang; Li Na. Design and construction of information navigation system for online education platform in ethnic minority areas based on card classification method. Journal of Shaanxi University of Science and Technology 45-48

[3] Wang Lin; Wu Xiaodong; Tang Min; Li Yongjun; Hu Meijiao. An improved graph neural network-based method and system for recommending online education resources [P]. Anhui Education Network Publishing Co., Ltd. 2025. 5-8

[4] Ma Yao. A teaching management method based on English online education [P]. Sichuan Geely University. 2025. 67-69

[5] Gao Xuan; Zhu Xiaofang; Peng Binghui. A real-time online education training personnel management information method and system [P]. Tianjin Hang'an Education Technology Co., Ltd. 2025. 23-26

[6] Kong Chao; Chen Jiahui; Gao Xiangyun; Sun Xianlan; Zhang Liping. Knowledge recommendation method, device, computer equipment and medium for online education platform [P]. Anhui University of Technology. 2025. 3-9

[7] Li Kuihua; Guo Jinting. Research on the integration of inquiry-based classroom and online education platform. Journal of Taiyuan City Vocational and Technical College, 2025(10).34-37

[8] Tang Sun; Liu Xiaojun; He Yichi; Wang Yilei; Yu Sipeng; Sun Yanwu; Lu Beini. Research on the construction of knowledge graph of course points for personalized learning. Automation Application, 2025(18).78-79

[9] Huang Daoran; Zhu Kai; Wang Qinyong. Research on the evaluation of the effect of online education driven by artificial intelligence. Office Automation, 2025(18)3-6

[10] Ma Yao. A teaching management method based on English online education [P]. Sichuan Geely University. 2025. 87-89

