Eng Int Syst (2025) 6: 689–699

Engineering
Intelligent Systems

Classroom Teaching Behavior Analysis and Teaching Quality Evaluation Design Based on Deep Learning

Yaqi Lian [1], Feng He [2,*] and Shanshan Wang[3]

[1] *School of Mechanical and Electrical Engineering, Taizhou Vocational & Technical College, Taizhou 318000, China*
[2] *College of Humanities and Law, Zhejiang Gongshang University Hangzhou College of Commerce, Hangzhou 311500, China*
[3] *The Eurasian Institute of Language and Culture, Xi'an FanYi University, Xi'an 710000, China*

With the rapid development of educational technology, the analysis of students' classroom behavior has been an important focus of educational research. Traditional classroom monitoring methods rely on actual observation and cannot provide accurate and comprehensive analysis. Multimodal data analysis combined with deep learning technology provides a new way to solve this problem. This paper proposes a multimodal classroom behavior analysis system based on video, audio and sensor data, which aims to accurately identify and predict the behavior of teachers and students through deep learning models to improve the quality of classroom teaching. The system includes a data acquisition module, data preprocessing module, feature extraction module, behavior analysis and prediction module, and result feedback module. First, classroom behavior data is collected through video, audio and sensors, and the data is preprocessed by denoising, normalization and other operations. Then, a convolutional neural network (CNN) is used to extract image features from video, the MFCC method is used to extract spectral features from audio, and the LSTM model is used to extract time series features from sensor data. Then, the system uses deep neural network (DNN) for behavior classification and LSTM to predict learning status and teaching quality. Finally, the analysis results are fed back to teachers and education managers in the form of reports. Experimental results show that the system performs well in terms of behavior classification and regression tasks, with higher accuracy, precision and F1 score than traditional models, and the system has stable performance in different tasks. This research not only provides new ideas for classroom behavior analysis, but also provides decision support tools for educational administrators.

Keywords: multimodal data, deep learning, behavior analysis, classroom monitoring, educational technology

1. INTRODUCTION

Teaching quality assessment is an indispensable and important part of the modern education system. It aims to comprehensively analyze and evaluate the teaching effectiveness of teachers, students' learning performance and various factors influencing classroom teaching. With the continuous development of information technology, traditional teaching quality assessment methods face many challenges and limitations [1]. Traditional assessment methods rely mainly on actual observation and teacher self-evaluation. These methods are not only time-consuming and labor-intensive, but also are prone to the influence of subjective factors, and cannot objectively and impartially reflect the real situation in the classroom. Therefore, how to evaluate the quality of classroom teaching through more scientific, objective and efficient methods has become an important topic in current educational research [2].

In recent years, deep learning technology has achieved remarkable results in various fields, especially in computer vision, natural language processing and other fields. Deep

*Corresponding author's E-mail: hefeng@zjhzcc.edu.cn

learning can automatically extract features and perform classification and prediction by learning a large amount of data, thereby replacing manual intervention in many tasks and greatly improving processing efficiency and accuracy [3]. In the field of education, the deep learning approach offers new opportunities for the analysis of classroom teaching practices and teaching quality evaluation [4]. By analyzing classroom images, videos and student and teacher interaction data [5, 6], deep learning can efficiently and accurately identify various behaviors in classroom teaching and provide a scientific basis for evaluating teaching quality. Therefore, teaching behavior analysis and quality evaluation technology based on deep learning has gradually become an important tool for improving education quality and management level.

The main reason for analyzing classroom teaching practices is to improve the effectiveness of classroom teaching. Teachers' teaching behaviors, students' learning attitudes, and the quality of the interaction between teachers and students are key factors affecting teaching effectiveness. By applying deep learning technology, various teaching behaviors in the classroom can be analyzed in real time to accurately identify teachers' teaching methods, students' learning status, and classroom interaction patterns. Based on these analysis results, teachers can adjust their teaching strategies in real time [7], optimize classroom teaching, and thus improve teaching quality. At the same time, students can also understand their own learning progress and problems through feedback on learning status data, which helps them improve their learning outcomes. The application of deep learning technology makes classroom teaching no longer a one-way process of imparting knowledge, but a dynamic, interactive, and intelligent feedback process.

In their daily management work, educational administrators usually need to deal with a large amount of teaching data, including classroom videos, student performance, teacher evaluations, etc. The processing and analysis of these data traditionally rely on manual operations, which is not only inefficient, but also makes it difficult to draw accurate and systematic conclusions. However, through deep learning technology, data can be automatically processed and intelligently analyzed. The teaching behavior analysis and evaluation system based on deep learning can automatically identify key behaviors and interaction patterns in the classroom, discover potential teaching problems, and provide specific suggestions for improvement. This intelligent support not only improves the efficiency of education management; it also provides a data-driven scientific basis for education decision-making, and helps to refine EMSs and integrate intelligent technology [8, 9].

This study aims to design a classroom teaching behavior analysis and teaching quality evaluation system based on deep learning. Combining multimodal data sources such as video images and sensor data, the deep learning model automatically analyzes teaching behaviors and student performance in the classroom to provide data support for evaluating teaching quality. The system will be able to monitor classroom behavior patterns, teacher-student interactions, student emotions and other information in real time, and quantitatively evaluate the overall quality of classroom teaching. Through this system, teachers can understand their teaching effectiveness and adjust teaching strategies in a timely manner; educational administrators can fully grasp the classroom teaching scenario, thereby optimizing resource allocation and teaching arrangements. Traditional classroom teaching behavior analysis relies mainly on actual observation, which is both subjective and inefficient. Hence, this study proposes to automate classroom behavior analysis through deep learning technology, use computer vision, speech recognition and other technologies to automatically identify various teaching behaviors of teachers and students, and combine multi-dimensional evaluation indicators to comprehensively evaluate teaching quality. This automated system can not only improve the efficiency of analysis; it also reduces the influence of human factors, provides more objective and accurate evaluation results, and promotes the comprehensive upgrade of teaching quality evaluation [10, 11].

In order to build a complete teaching behavior analysis and evaluation system, this study adopted a variety of deep learning techniques, including convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory networks (LSTM), etc., combined with multimodal information such as classroom videos, audio, sensor data, etc., through data processing, feature extraction, model training and other links.

2. RELATED RESEARCH

2.1 Research on Teaching Behavior Analysis

Teaching behavior analysis refers to the quantitative analysis and quality evaluation of the behavior of teachers and students during the teaching process by systematically observing and recording them. With the continuous development of educational technology, the methods of teaching behavior analysis have also undergone a transformation from traditional actual observation to automated analysis technology. In recent years, an increasing number of studies have attempted to use advanced image processing, sensors, and video surveillance technologies to assist in the analysis and identification of teaching behaviors, thereby improving the efficiency and accuracy of the analysis.

Classroom behavior analysis involves several aspects: teacher behavior, student behavior, and teacher-student interaction. Teacher behavior analysis focuses on the various behaviors displayed by teachers during the teaching process, such as explanation methods, classroom management, body language, etc.; student behavior analysis focuses on observing students' learning attitudes, classroom participation, and concentration; and teacher-student interaction analysis concerns the interaction patterns between teachers and students, including the frequency, form, and effect of communication [12]. With the development of technology, modern teaching behavior analysis increasingly relies on automated tools to improve analysis efficiency and reduce human errors. Traditional teaching behavior analysis methods rely more on manual records by observers, usually including classroom recordings, videos, and self-reports by teachers or students. The most classic teaching behavior analysis tool is the

"classroom observation method", in which researchers record and analyze events that occur in the classroom through a pre-designed behavior classification system [13]. However, this method has its limitations. First, it is extremely time-consuming and labor-intensive. Second, the observer's bias and subjective judgment may affect the results of the analysis. Manual observation may have different interpretations of the same behavior due to different observers' experiences and backgrounds, thus affecting the reliability of the results [14].

With the development of information technology, automated teaching behavior recognition technology based on video and sensors has been widely used. Video surveillance technology can capture classroom teaching activities in real time and use computer vision technology to identify the behaviors of teachers and students. For example, the literature proposes a method to evaluate the teaching effect of teachers by analyzing their expressions and body language captured through video [6]. In addition, sensor-based teaching behavior analysis has gradually become an effective technical means. By using infrared sensors, motion sensors and environmental monitoring sensors, researchers can collect information such as students' seat changes, activity status, and temperature and humidity of the classroom environment in real time, thereby conducting a comprehensive analysis of the teaching process [7]. These automated technologies have greatly improved the efficiency of data collection and the accuracy of analysis.

2.2 Research Progress on Teaching Quality Assessment

Teaching quality assessment is a core factor in education management. Its purpose is to conduct a comprehensive assessment of teachers' teaching effectiveness, students' learning outcomes and various aspects of the teaching process, in order to provide a scientific basis for teaching improvement. Traditional teaching quality assessment methods rely mostly on self-evaluation and questionnaire surveys of teachers and students. In recent years, assessment methods based on big data and artificial intelligence have emerged. Traditional teaching quality assessment methods usually rely on teachers' self-evaluation and students' evaluation. In terms of teacher self-evaluation, teachers will reflect on themselves and put forward suggestions for improvement based on their classroom performance. The students' evaluation of their teachers usually involves questionnaires or scoring systems enabling students to give feedback on their teachers' classroom practices. These methods usually rely on subjective evaluations of teachers and students, and it is difficult to avoid the inconsistencies caused by individual differences [8]. For example, the literature points out that student evaluations are often affected by the course content and the students' personal interests and emotions, which may lead to distorted evaluation results [9].

In recent years, with the rise of big data and artificial intelligence technologies, data-driven teaching quality evaluation methods have received widespread attention. These methods collect multi-dimensional data such as student grades, classroom behavior data, and teacher explanation content, and use statistical analysis or machine learning algorithms to evaluate teaching quality. For example, the literature uses student grades and classroom participation data, combined with machine learning algorithms to predict students' academic performance, thereby evaluating teachers' teaching quality [10]. In addition, some researchers use natural language processing (NLP) technology to perform sentiment analysis on student feedback to further improve the teaching quality evaluation system [11].

2.3 Application of Deep Learning in Education

Deep learning originated from artificial neural networks. It simulates brain processes by building a multi-layer network. It can automatically extract features from massive data and perform pattern recognition. With the improvement of computing power and the increase of data, deep learning has made remarkable achievements in image recognition, speech processing, natural language processing and other fields [15]. In the field of education, deep learning is widely used in teaching behavior analysis, personalized learning recommendation, academic performance prediction and other aspects. In particular, in teaching behavior analysis, deep learning can automatically identify the behavior of teachers and students by learning a large amount of data, thereby providing more accurate teaching evaluation [16]. In recent years, the application cases of deep learning in the field of education have emerged in an endless stream. The literature proposes a classroom behavior analysis method based on deep learning. By analyzing classroom videos, it automatically identifies the teacher's body language, tone and students' reactions, thereby evaluating the effect of teaching interaction [17]. In addition, deep learning is also used for student behavior analysis and academic performance prediction. Researchers use students' online learning data and behavior patterns to predict students' learning outcomes and provide personalized recommendations [18]. For example, the literature uses deep learning algorithms to analyze students' learning logs, predict their academic performance and provide customized learning resources [19].

3. INNOVATIVE TEACHING BEHAVIOR ANALYSIS METHOD BASED ON DEEP LEARNING

With the increasing complexity of the educational environment and the amount of data available, traditional teaching behavior analysis methods can no longer meet the requirements of accuracy and real-time performance in modern education. Innovative teaching behavior analysis methods based on deep learning can effectively meet this challenge. This paper proposes an innovative teaching behavior analysis method based on deep learning, which combines multimodal data fusion, adaptive learning and time series modeling, and improves the accuracy and real-time performance of classroom behavior analysis through feature extraction and multi-level information processing.

3.1 Innovative Ideas and Framework

The innovative teaching behavior analysis method proposed in this study utilizes the multimodal learning framework of deep learning. This framework can comprehensively analyze teaching behavior by fusing video, audio, and sensor data. Specifically, the video data is extracted with visual features through convolutional neural network (CNN), the audio data is extracted with speech features through the Mel frequency cepstral coefficient (MFCC), and the sensor data is processed with time series features through the long short-term memory network (LSTM). By means of multimodal feature fusion and an adaptive weighting mechanism, the model can automatically adjust the weights of various modalities in different teaching scenarios, thereby improving the accuracy of the analysis results. In this method, we use convolutional neural network (CNN) to extract features of images in the video layer by layer; at the same time, we extract spectral features from audio signals through the MFCC, and then process the time series information in sensor data through LSTM network. After multimodal feature fusion, behavior classification and prediction are finally performed through deep neural network (DNN). In order to ensure the adaptability and robustness of the model, we introduce the self-attention mechanism to dynamically adjust the weights of each modality during feature fusion [20, 21].

3.2 Data Preprocessing and Feature Extraction

Before multimodal data fusion, each data source needs to be preprocessed and influential features need to be extracted. For video data, the video frames are first sampled, and representative frames are selected for processing. Then, the convolutional neural network (CNN) is used to extract image features. The convolutional layer of CNN can extract information such as the teacher's body language, facial expressions, and the students' visual focus from each frame of the image. The video data processing flow is as follows [22, 23], as shown in Equation (1):

$$F_v = \mathrm{CNN}(I_v) \tag{1}$$

where, F_v is the feature extracted from video data, and I_v is the input video frame.

For audio data, we use MFCC for feature extraction. After the audio signal is preprocessed, the MFCCs of each audio segment are calculated, and further feature learning is performed through a convolutional neural network. The audio data processing flow is [24, 25], as shown in Equation (2).

$$F_a = \mathrm{CNN}(\mathrm{MFCC}(A)) \tag{2}$$

where, F_a is the feature extracted from audio data, and A is the input audio data.

For sensor data, we use the LSTM network to process time series data. LSTM can capture long-term dependencies in data and is particularly suitable for processing time series data. Sensor data is extracted from time series features through the LSTM model, as shown in Equation (3):

$$F_s = \mathrm{LSTM}(S) \tag{3}$$

where, F_s is the feature extracted from sensor data, and S is the input sensor data.

3.3 Multimodal Data Fusion and Adaptive Weighting

After feature extraction, the next step is to fuse multimodal data. Traditional feature fusion methods often use simple concatenation or weighted averaging, but these methods cannot effectively handle the heterogeneity of data in different modalities. Therefore, this paper introduces an adaptive weighting mechanism to automatically adjust the weights of each modality through the self-attention mechanism to optimize the final result of the feature fusion [26].

Assume that the video feature is F_v, the audio feature is F_a, and the sensor feature is F_s, then through the self-attention mechanism, the model dynamically calculates the weight of each modality $\alpha_v, \alpha_a, \alpha_s$. These weights are automatically adjusted based on the contribution of each modality to the behavior analysis. The expression after feature fusion is Equation (4):

$$F_{fused} = \alpha_v F_v + \alpha_a F_a + \alpha_s F_s \tag{4}$$

The calculation of adaptive weights is based on the attention mechanism, in which w_m, the weight representing the modality, can be dynamically calculated with Equation (5):

$$\alpha_m = \frac{\exp(w_m)}{\sum_{i=1}^{3} \exp(w_i)} \tag{5}$$

where w_m is a trainable parameter related to the modality, α_m is the weighting coefficient of the modality, which can be updated and dynamically adjusted during the training process.

3.4 Behavior Recognition and Prediction

After multimodal data fusion, combined with the features of video, audio, and sensor data, we use deep neural networks (DNN) to classify and predict teaching behaviors in the classroom. Assume that the fused features are F_{fused}, By inputting these features into the deep neural network, we can obtain the behavior prediction result y, that is, the identified teaching behavior category.

The calculation process of the deep neural network can be expressed with Equation (6):

$$y = \sigma(W_n \cdot h_n + b_n) \tag{6}$$

where y is the prediction result of the model output, representing different teaching behavior categories; W_n is the weight matrix of the nth layer; h_n is the output of the nth layer b_n is the bias term; σ is the activation function, which is used to introduce nonlinearity. The output of each layer is nonlinearly transformed through the activation function, gradually extracting more feature information, and finally obtaining the prediction of the behavior category.

Table 1 Dataset description.

Dataset	Quantity	Video duration (hours)	Audio duration (hours)	Sensor data points	Category distribution
train	5000	200	200	1M	[0.4, 0.6]
verify	1000	40	40	200K	[0.5, 0.5]
test	1500	60	60	300K	[0.3, 0.7]

During the training process, the network performs gradient descent optimization based on the error between the predicted result and the true label. The loss function usually uses the cross-entropy loss function to measure the performance of the classification task, as shown in Equation (7):

$$L_{class} = -\sum_{i=1}^{C} y_i \log(\widehat{y_i}) \tag{7}$$

where C is the number of behavior categories, y_i is the true label, and $\widehat{y_i}$ is the probability predicted by the model. This loss function optimizes the classification performance of the model by minimizing the difference between the predicted probability and the true label.

In addition, we can use other regression loss functions to predict some continuous behavioral characteristics, such as the student's attention level or the teacher's explanation intensity. The loss function of the regression task is generally the mean square error (MSE), as shown in Equation (8):

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{8}$$

where $\widehat{y_i}$ is the predicted value of the regression task, y_i is the true value, and N is the number of samples. The goal of the regression task is to minimize the difference between the predicted value and the actual value in order to obtain a more accurate estimate of the behavior intensity.

3.5 Model Training and Optimization

During the training process, we use multi-task learning to optimize, that is, to perform classification and regression tasks in the same model at the same time. Multi-task learning can help the model learn teaching behaviors from multiple perspectives and improve the overall recognition accuracy. To this end, we use different loss functions for each task and balance the training objectives of the classification and regression tasks through weighted coefficients. The final loss function can be expressed as Equation (9):

$$L = \lambda_1 L_{class} + \lambda_2 L_{reg} \tag{9}$$

where λ_1 and λ_2 are hyperparameters that control the weights of the classification task and regression task respectively. These hyperparameters are adjusted through cross-validation during training to find the optimal loss function combination.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Design

The architecture of this system includes five main modules: data acquisition, data preprocessing, feature extraction, behavior analysis and prediction, and result feedback. The data acquisition module collects video and audio data in the classroom through cameras and microphones, and obtains relevant sensor data through environmental monitoring equipment and sensors worn by students. The collected raw data undergoes preprocessing that involves denoising, normalization, and segmentation to ensure data quality and consistency. Video data is segmented into frames, audio data is spectrally converted, and sensor data is organized through time series processing. In the feature extraction module, the system uses deep learning models such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) to extract useful features from video, audio, and sensor data. Video data uses CNN to extract image features, audio data uses MFCC to extract spectral features, and then further extracts features through CNN, and sensor data uses LSTM for time series modeling. Combining these multimodal features, the behavior analysis and prediction module uses deep neural networks (DNNs) to classify behaviors and uses LSTMs for time series prediction, outputting the behavior analysis results of students and teachers, as well as predictions of students' learning status and teaching quality. Finally, the system presents the analysis results in the form of charts and reports to teachers and education administrators, providing them with data support so that they can make more accurate decisions. The entire system is implemented via the Python programming language, using deep learning frameworks such as TensorFlow and PyTorch. At the same time, libraries such as OpenCV, Librosa, Pandas and NumPy are used for data collection and processing, and the result feedback module uses Matplotlib and Seaborn for visualization.

4.2 Experimental Results

Table 1 shows the multimodal datasets used for training, validation, and testing. The dataset includes video, audio, and sensor data, covering information such as the distribution, data volume, resolution, and frame rate of different categories. The training set contains 5,000 samples, with a total video and audio duration of 200 hours, and 1 million sensor data points. The number of samples in the validation set and test set are 1,000 and 1,500 respectively, with video and audio duration of 40 hours and 60 hours. The category distribution in the dataset reflects the diversity of teacher-student interactions, and the

Table 2 Model architecture parameters.

Layer Type	Convolution kernel size	Number of filters	Number of LSTM cells	DNN layer depth	Activation Function	Dropout ratio
CNN-1	3x3	64	-	-	ReLU	0.2
CNN-2	5x5	128	-	-	ReLU	0.3
LSTM	-	-	256	-	tanh	-
DNN-1	-	-	-	2	ReLU	0.4
DNN-2	-	-	-	1	Softmax	-

Table 3 Comparison of feature extraction efficiency.

feature	Extraction time (seconds/frame)	Memory usage (MB)	Feature Dimension	Data preprocessing time (seconds)	Data transfer rate (Mbps)
video	0.05	500	1024	1.5	100
Audio	0.02	100	256	0.5	50
sensor	0.01	20	64	0.2	10

distribution of each category tends to be balanced to ensure that there is no bias in model training. The processing and analysis of audio and sensor data adopts corresponding audio spectrum and time series processing techniques. The diversity and quality of the dataset increase the generalizability of the model in real-world applications, thereby improving the accuracy and effectiveness of the system in classroom behavior analysis.

Table 2 lists the parameter settings for each layer in the model, including convolution kernel size, number of filters, number of LSTM units, DNN layer depth, activation function, and Dropout ratio. The model contains multiple convolutional layers (CNN-1 and CNN-2) to extract features of video data, using ReLU activation function and appropriate Dropout ratio to avoid overfitting. The LSTM layer is used to process time series data, has 256 units, and is modeled by the Tanh activation function. The DNN part contains two layers, one for feature fusion and one for classification, respectively. The first layer uses the ReLU activation function, and the second layer uses the Softmax activation function for multi-classification output. The Dropout ratio setting (0.2 to 0.4) is designed to enhance the robustness of the model and prevent overfitting. With this architecture, the model can effectively process video, audio, and sensor data, capture classroom behavior patterns, and make accurate predictions.

For comparison purposes, Table 3 presents the efficiency of video, audio, and sensor data in feature extraction. The extraction time of video data is 0.05 seconds/frame, the memory consumption is 500MB, and the feature dimension is 1024. The extraction time of audio data is 0.02 seconds/frame, the memory consumption is 100MB, and the feature dimension is 256. The extraction time of sensor data is the shortest, only 0.01 seconds/frame, the memory consumption is 20MB, and the feature dimension is 64. The data preprocessing time also lists the processing time of each data type. The preprocessing time of video data is the longest, and audio and sensor data are more efficient. In terms of data transmission rate, the transmission rate of video is 100Mbps, audio is 50Mbps, and sensor data is 10Mbps. The efficient extraction of these data provides the basis for the system's real-time processing and feedback which can monitor and analyze the learning situation in real time, ensuring the rapid response of data transmission and analysis, especially in the classroom environment.

Figure 1 shows the performance indicators of the model for the classification task, covering several evaluation criteria: accuracy, precision, recall, F1 score and AUC. The performance of each category (such as teacher explanation, student participation, teacher questioning, etc.) shows high accuracy and F1 scores. In particular, teacher explanation (95.2%) and classroom management (96.0%) performed best. The AUC value is also above 0.97, indicating that the model has strong discrimination ability in each category. The values of precision and recall show that the model has a good balance when dealing with different categories, and can effectively distinguish the behavior patterns of teachers and students. The F1 score indicates the robustness of the model when dealing with unbalanced data, and the higher value further confirms the reliability of the model. As shown by these indicators, the model can provide accurate classroom behavior analysis and prediction, supporting teachers and education administrators to better understand classroom dynamics.

Figure 2 shows a regression analysis, where the blue dots represent the attention level data and the green dots represent the lecture intensity data. The figure also shows two regression lines: the red line for the attention level and the orange line for the lecture intensity. As can be seen from the figure, as the X-axis value increases, both the attention level and the lecture intensity increase, although the increasing trend of the lecture intensity is more obvious.

Table 4 lists the performance indicators for the regression tasks in terms of MAE, RMSE, R^2, MAPE, and Spearman correlation coefficient. The regression tasks focus on predicting the attention level and teaching intensity in the classroom. The MAE of attention level is 0.05, RMSE is 0.1, R^2 is 0.9, and MAPE is 0.08, indicating that the model performs well in predicting students' attention level. The performance of teaching intensity is also excellent, with MAE of 0.04, RMSE of 0.08, R^2 of 0.95, and MAPE of 0.06, showing that the model can accurately predict the intensity of

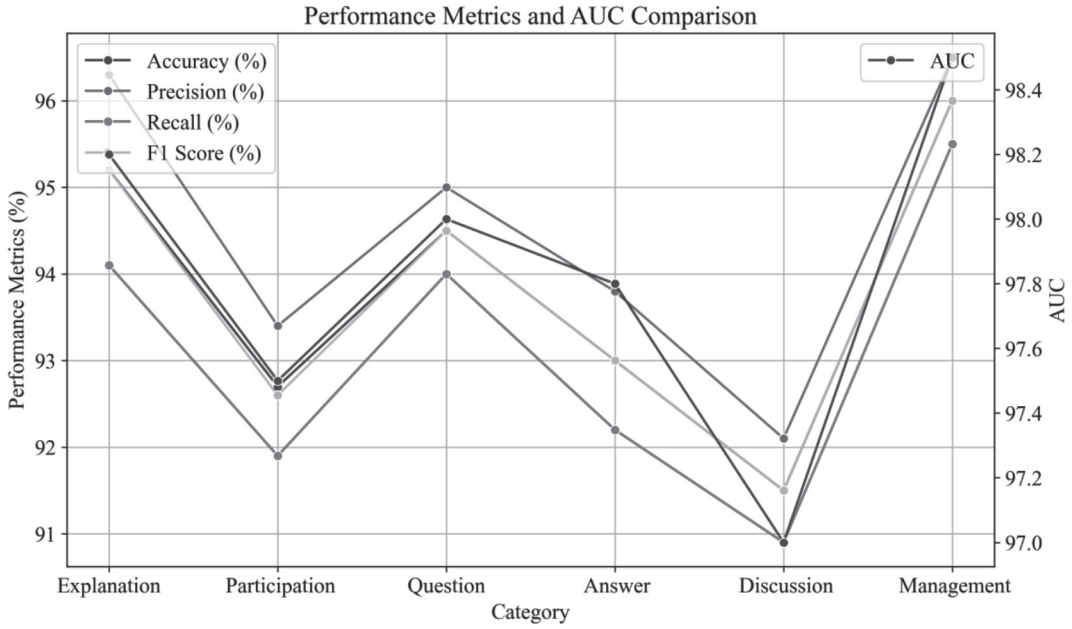

Figure 1 Classification task performance indicators.

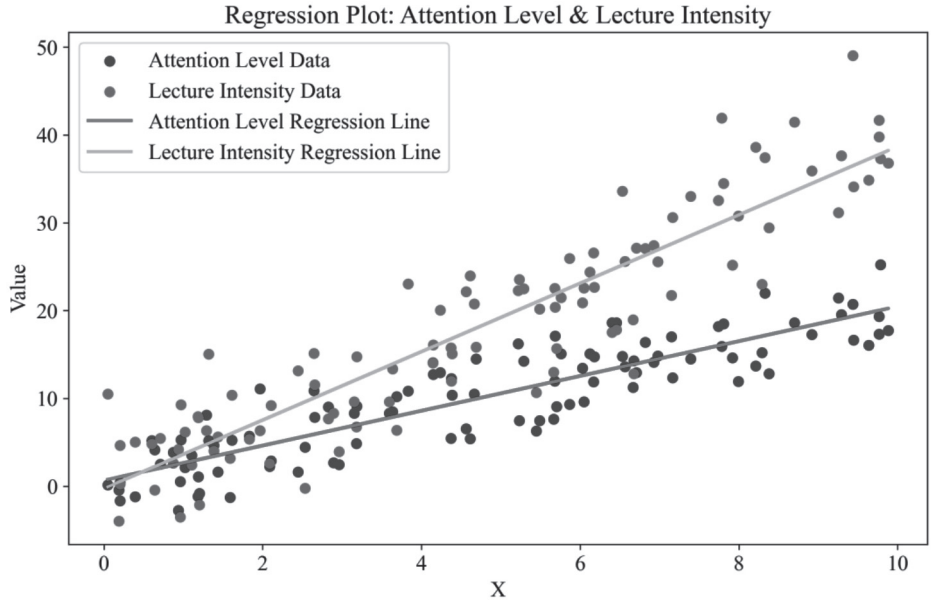

Figure 2 Regression analysis.

Table 4 Regression task performance indicators.

Target	MAE	RMSE	R^2	Mean Absolute Percentage Error (MAPE)	Spearman correlation coefficient
Attention Level	0.05	0.1	0.9	0.08	0.92
Explain the intensity	0.04	0.08	0.95	0.06	0.94

teacher teaching. The Spearman correlation coefficients (0.92 and 0.94) further confirm the high correlation between the predicted results and the actual values. The accuracy of these regression tasks is of great significance for the evaluation and improvement of classroom teaching quality and can provide effective feedback to teachers.

Figure 3 shows the performance comparison between the Baseline model and the proposed model. The proposed model outperforms the Baseline model in terms of accuracy (95% vs. 88%) and F1 score (94% vs. 87%), and also performs better in terms of MAE and RMSE indicators. The AUC value increased from 0.9 to 0.98, further proving the superiority of the proposed model in multi-category classification tasks. In terms of inference time, the proposed model has an inference time of 80ms, slightly lower than the 100ms of the Baseline. Although the number of parameters has increased to 1 million,

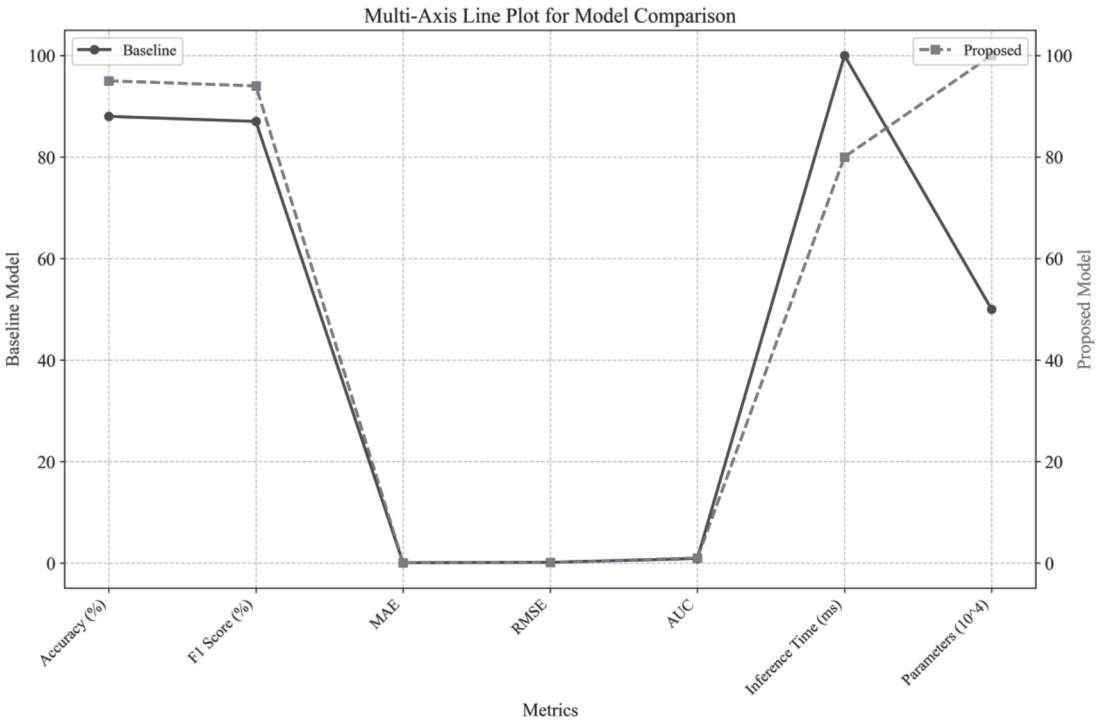

Figure 3 Performance comparison between different models.

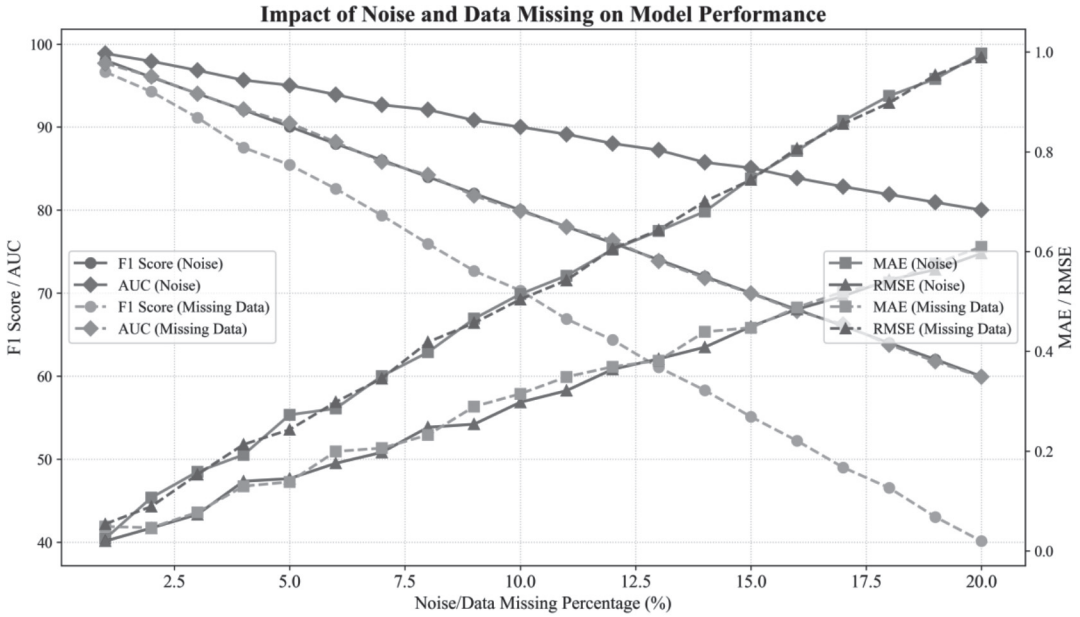

Figure 4 Model robustness test.

it can maintain a faster response speed due to its more efficient architecture design. These comparisons demonstrate that the proposed model has stronger classification capabilities and lower prediction errors, proving its applicability to the analysis of classroom behavior.

Figure 4 shows the impact of noise and missing data on model performance. A variety of colors and symbols are used in the figure to distinguish various indicators and conditions. Specifically, the blue line and dots represent the F1 score under noise conditions, the purple line and diamond represent the AUC value under noise conditions, the cyan line and square represent the F1 score under missing data conditions, and the pink line and triangle represent the AUC value under missing data conditions. In addition, the green line and square represent the mean absolute error (MAE) under noise conditions, the red line and dots represent the root mean square error (RMSE) under noise conditions, the orange line and square represent the MAE under missing data conditions, and the brown line and triangle represent the RMSE under missing data conditions. It can be seen from the figure that as the noise and missing data ratios increase, the F1 score and AUC value gradually decrease, while the MAE and RMSE

Table 5 User feedback and system improvement.

Improvements	Rating before improvement	Improved scoring	Feedback sample size	Improvement (%)	User review sentiment score
Report clarity	3.5	4.5	100	28.57	0.8
Analytical accuracy	3.8	4.7	120	23.68	0.85
System response speed	3.2	4.2	150	31.25	0.82

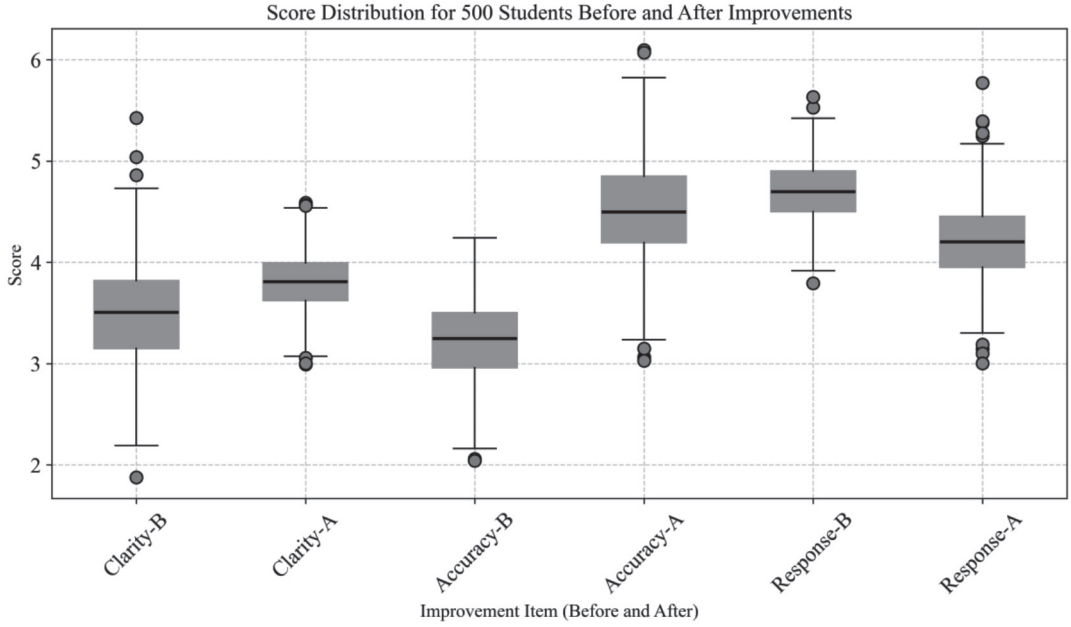

Figure 5 User feedback and system improvement.

gradually increase. This shows that the performance of the model will be affected by noise and missing data. Specifically, the impact of noise on model performance is more significant than is the influence of missing data, especially under high noise ratios, the decrease in F1 score and AUC value is more obvious. In addition, the impact of missing data on model performance also increases with the increase of missing ratios, but its impact is slightly lower than that of noise.

Table 5 shows the effects of user feedback and system improvements. User ratings have significantly improved as a result of better report clarity, analysis accuracy, and system response speed. For example, the report clarity rating increased from 3.5 to 4.5, an improvement of 28.57%. Analysis accuracy increased by 23.68%, and system response speed increased by 31.25%. The sentiment scores of user comments were high, indicating that the improvement measures received positive feedback from users. These improvements further optimized the user experience of the system and enhanced the system's practical application value.

As can be seen from Figure 5, for clarity and accuracy, the median and interquartile range of the improved score distribution (A) are higher than the score distribution (B) before the improvement, which shows that the improvement measures have increased students' scores. In particular, the improvement effect of accuracy is the most significant. The improved score distribution not only has a higher median, but also has fewer outliers, indicating that the improvement measures have a significant effect on the improvement of accuracy. For responsiveness, the median of the improved score distribution (A) is slightly higher than that of the score distribution (B) before the improvement, but the interquartile range and outliers have not changed much, indicating that the improvement measures have a certain effect on improving responsiveness, but not as significant as clarity and accuracy.

4.3 Discussion

The advantage of this study is that the proposed multimodal classroom behavior analysis system is innovative and practical. The system combines video, audio and sensor data, uses deep learning technology for behavior analysis and prediction, and provides teachers and education administrators with a more accurate tool for classroom monitoring. First, the article introduces the system architecture and technical details in detail, involving multiple modules such as data acquisition, preprocessing, feature extraction, behavior analysis and prediction, and makes full use of deep learning models such as CNN, LSTM and DNN, reflecting the current application prospects of artificial intelligence technology in the field of education. Secondly, the article verifies the superiority of the model in different tasks (such as classification tasks and regression tasks) through experimental results, provides rich quantitative data support, and proves the actual effect of the system in improving the accuracy of classroom behavior analysis and teaching quality evaluation. In comparison with traditional models, the proposed model shows stronger classification ability and produces fewer

errors. However, this study has several some shortcomings. First, although the experimental results show that the system performs well in multi-task learning, its model architecture is relatively complex, and the increase in inference time and parameter quantity may impose certain computational burdens, especially on resource-constrained devices, which may face performance bottlenecks. Secondly, although the model can still maintain good robustness under noise interference and data loss conditions, the performance of the system in extreme environments has not been fully verified. In some complex classroom environments, such as a large number of simultaneous interactions or a noisy background, the performance of the system may be affected. Finally, although the article provides positive improvement data in terms of user feedback and system improvements, the long-term stability, scalability and adaptability of the system to different educational scenarios have not been deeply explored.

5. CONCLUSION

The analysis of traditional classroom behavior relies on personal observation, which has problems such as low accuracy and poor efficiency, and cannot provide real-time feedback on the specific behaviors of teachers and students. In order to overcome these challenges, this paper proposes a classroom behavior analysis system based on multimodal data (video, audio and sensor data) and deep learning, aiming to improve the accuracy and real-time performance of classroom behavior recognition and prediction, and thus optimize classroom teaching management. The system comprises five modules: data acquisition, data preprocessing, feature extraction, behavior analysis and prediction, and result feedback. For data acquisition, the system collects various data in the classroom through video, audio and sensor devices. Video data is obtained through cameras, audio data is collected by microphones, and sensor data comes from environmental monitoring equipment and sensors worn by students. The data preprocessing module performs denoising, normalization, segmentation and other operations on the raw data to ensure data quality and consistency. The feature extraction module uses deep learning methods to extract significant features from different types of data: video data uses convolutional neural networks (CNNs) to extract image features, audio data uses MFCC methods to extract spectral features, and sensor data uses LSTM models to extract time series features. The behavior analysis and prediction module is a combination of deep neural networks (DNN) for behavior classification, and uses the LSTM model for time series prediction, ultimately predicting students' learning status and the instructor's teaching quality. The result feedback module presents the analysis results in the form of a report for reference by teachers and education administrators. The experimental results show that the proposed system performs well in multiple tasks. Specifically, in the classification task, the accuracy, precision, and F1 scores of behaviors such as teacher explanation, student participation, and teacher questioning all reached more than 95%, and the prediction results in the regression task also had a high R^2 value, indicating that the system can accurately predict classroom teaching quality and the students' learning status. In addition, compared with the traditional model, the model proposed in this paper is a significant improvement in terms of accuracy, F1 score, and inference time, showing stronger robustness and efficiency.

The significance of this system is that it provides an automated and precise tool for classroom behavior analysis, which can help teachers and education administrators better understand classroom dynamics and optimize teaching strategies and management decisions. At the same time, the system provides new ideas for the application of multimodal data in the field of education and encourages the integration of artificial intelligence and educational technology.

FUNDING

This study was supported by Research on the Evaluation Model of Digital Teaching Ability of Vocational College Teachers under the Background of Industry-Education Integration, Chinese Vocational Education Research Project of Zhejiang Province (2025), Project Number: ZJCV2025B05.

REFERENCES

1. Nusseck M, Richter B, Spahn C, Echternach M. Analyzing the vocal behavior of teachers during classroom teaching using a portable voice accumulator. Logopedics Phoniatrics Vocology. 2018; 43(1):1–10.
2. Jiang JX, Zeng L, Kue J, Li H, Shi Y, Chen CP. Effective teaching behaviors in the emergency department: A qualitative study with Millennial nursing students in Shanghai. Nurse Education Today. 2018; 61:220–224.
3. Chen G, Zhang Q, Han Y. Innovation of college English teaching methods based on deep learning. Engineering Intelligent Systems. 2025; 33:355–364.
4. Zang S. Optimization strategy for teacher performance management in higher vocational colleges based on deep learning. Engineering Intelligent Systems. 2025; 33(3):311–320.
5. Wang SY. Online learning behavior analysis based on image emotion recognition. Traitement Du Signal. 2021; 38(3): 865–873.
6. Ren TY, Dong YF, Wu D, Chen K. Design of direct teaching behavior of collaborative robot based on force interaction. Journal of Intelligent & Robotic Systems. 2019; 96(1):83–93.
7. Wang HJ, He MZ, Zeng CL, Qian L, Wang J, Pan W. Analysis of learning behavior in immersive virtual reality. Journal of Intelligent & Fuzzy Systems. 2023; 45(4):5927–5938.
8. Trevisan DF, Becerra L, Benitez P, Higbee TS, Gois JP. A review of the use of computational technology in applied behavior analysis. Adaptive Behavior. 2019; 27(3):183–196.
9. Gong TY, Wang JM. A data-driven smart evaluation framework for teaching effect based on fuzzy comprehensive analysis. IEEE Access. 2023; 11:23355–23365.
10. Yang BH, Yao ZP, Lu H, Zhou YQ, Xu JK. In-classroom learning analytics based on student behavior, topic and teaching characteristic mining. Pattern Recognition Letters. 2020; 129:224–231.
11. Epifânio JC, da Silva LF. Embracing applied behavior analysis on a serious game design document model. IEEE Access. 2023; 11:72070–72087.

12. Ye JJ, Zhao YK, Teng ZS, Ye HW, Yuan Q, Nie X. Behavior analysis and formative assessments in online oral medicine education during the COVID-19 pandemic. World Journal of Clinical Cases. 2023; 11(21):5063–5072.
13. Li SL, Chai HQ. Recognition of teaching features and behaviors in online open courses based on image processing. Traitement Du Signal. 2021; 38(1):155–64.
14. Chen QG. User-based fraud behavior analysis and research on e-learning platforms: An empirical study. Journal of Nonlinear and Convex Analysis. 2019; 20(6):1155–1171.
15. Fu LP. A flexible group decision-making method for informationization teaching ability evaluation of university teachers integrating EXPTODIM and GRA under interval neutrosophic environment. Journal of Intelligent & Fuzzy Systems. 2024; 46(1):1245–1258.
16. Choshi M, Eaton MK, McDaniel VP. Rethinking student evaluation of teaching: A policy analysis. Journal of Professional Nursing. 2023; 49:165–170.
17. Leijten P, Gardner F, Melendez-Torres GJ, van Aar J, Hutchings J, Schulz S, et al. Meta-analyses: key parenting program components for disruptive child behavior. Journal of the American Academy of Child and Adolescent Psychiatry. 2019; 58(2):180–190.
18. Green CA, Chu SN, Huang E, Chern H, O'Sullivan P. Teaching in the robotic environment: Use of alternative approaches to guide operative instruction. American Journal of Surgery. 2020; 219(1):191–196.
19. Wang L, Zhen ZQ, Wo TY, Jiang B, Sun HL, Long X. A scalable operating system experiment platform supporting learning behavior analysis. IEEE Transactions on Education. 2020; 63(3):232–239.
20. Zheng QY, Chen ZZ, Wang MK, Shi YW, Chen SH, Liu Z. Automated multimode teaching behavior analysis: a pipeline-based event segmentation and description. IEEE Transactions on Learning Technologies. 2024; 17:1717–1733.
21. Elcoro M. Teaching machines, the history of personalized learning. Journal of the Experimental Analysis of Behavior. 2024; 122(1):87–97.
22. Gang Z, Zhu WJ, Hu BL, Jie C, Hui H, Qing X. A simple teacher behavior recognition method for massive teaching videos based on teacher set. Applied Intelligence. 2021; 51(12):8828–8849.
23. Cheng JQ, Wang XM. Artificial intelligence based on effectiveness of inverted classroom teaching of college sports. Journal of Intelligent & Fuzzy Systems. 2021; 40(2):3755–3765.
24. Howarth EJ, O'Connor DB, Panagioti M, Hodkinson A, Wilding S, Johnson J. Are stressful life events prospectively associated with increased suicidal ideation and behavior? A systematic review and meta-analysis. Journal of Affective Disorders. 2020; 266: 731–742.
25. Zhou HZ, Fan LJ, Wu CA, Luo AF, Mo CQ, He GH, et al. Understanding the associations among parents teaching safety rules to children, safety behaviors and unintentional injuries in Chinese preschool children. Preventive Medicine. 2019; 118: 98–103.
26. Xue Y, Qin JF, Su SB, Slowik A. Brain storm optimization based clustering for learning behavior analysis. Computer Systems Science and Engineering. 2021; 39(2):211–219.