Construction of an Intelligent Classroom Teaching Evaluation System Integrating Multimodal Deep Learning

Huang Da[1,2], Hean Liu[1]*

(1.Sehan University, Jeollanam-do, 58447, South Korea;2.Heyuan Polytechnic,Guangdong,517000,China)

Abstract: In the context of rapid development of educational informationization, traditional classroom teaching evaluation methods have become difficult to meet the needs of modern teaching quality improvement due to issues such as strong subjectivity, poor real-time performance, and single data dimensions. This paper constructs an intelligent classroom teaching evaluation system that integrates multimodal deep learning, based on models such as computer vision (ResNet+PoseEstimation), speech processing (CNN+LSTM), and natural language processing (BERT+Transformer). It comprehensively analyzes multimodal data such as students' facial expressions, speech emotions, and classroom speaking content, accurately quantifying student focus, classroom interaction index, and teaching quality. By implementing a hybrid fusion strategy that combines Early Fusion and Late Fusion, effective integration of different modal features is achieved. Experimental results show that this system can objectively, in real-time, and comprehensively feedback the classroom teaching process, providing data-driven personalized improvement suggestions for teachers. The research provides a feasible path for the construction of an intelligent classroom teaching evaluation system, with strong practical application value and promotion significance.

Keywords: Classroom teaching evaluation; Multimodal data; Deep learning; Educational intelligence; Teaching optimization

1.Introduction

In the context of the ongoing promotion of educational informationization, classroom teaching quality has increasingly become the focus of attention in educational research and practice [1]. Teaching evaluation, as an important tool for measuring teaching effectiveness and improving teaching behavior, plays a key role in classroom practice. However, traditional evaluation methods often suffer from strong subjectivity, delayed feedback, and single data dimensions, making it difficult to comprehensively reflect real-time classroom dynamics, thereby affecting the scientificity and targeting of teaching improvements. Methods such as teacher observation, student questionnaires, or post-class grades are primarily relied upon. With the increasing demand for personalized and precise teaching, there is an urgent need for a more objective, real-time, and multidimensional classroom evaluation method [2].

In recent years, the application of artificial intelligence technology, especially deep learning, in educational scenarios has been continuously expanding, bringing new opportunities for classroom teaching evaluation [3]. Computer vision (CV) technology has been widely applied in student behavior recognition, expression analysis, and attention detection [4]; automatic speech recognition (ASR) and speech emotion recognition (SER) have been used to analyze classroom speech content and recognize the tone and emotional state of teachers and students [5]; natural language processing (NLP) has been used to parse classroom Q&A content, assess the quality of questions, and the logic of student responses [6]. The integration of these technologies allows for the capture and modeling of multimodal classroom data, providing strong support for the intelligent and data-driven development of classroom evaluation.

Based on this, this paper aims to achieve multidimensional modeling of classroom teaching activities by integrating three core modes: visual, speech, and text through a hybrid fusion strategy, constructing a classroom teaching evaluation system

* Corresponding author:Hean Liu, Email:33894549@qq.com

that integrates multimodal deep learning technology [7]. This system not only quantifies student focus, participation, and teaching quality but also provides real-time feedback and personalized teaching optimization suggestions, thereby providing practical support for improving classroom teaching efficiency and quality.

2.Research Design and Methods

2.1 The overall architecture of the system

The classroom evaluation system constructed in this study adopts a multimodal deep learning architecture, combining visual, speech, and textual classroom data to achieve precise modeling and assessment of student learning behaviors and teacher teaching behaviors. The overall design of the system is based on a hybrid fusion strategy, integrating the advantages of Early Fusion and Late Fusion to enhance the model's generalization performance and robustness while maintaining the collaborative modeling capability of cross-modal features. [8].

The system's data collection module is deployed in real classroom environments, using high-definition cameras to capture visual data such as students' facial expressions, eye movement trajectories, head orientation, body movements, and teacher behaviors, analyzing students' attention levels, emotions, and interaction situations; high-sensitivity microphones are used to record teachers' spoken language and teacher-student interaction audio, identifying teachers' speech rates, tones, and emotional states, and assessing the classroom atmosphere; at the same time, automatic speech recognition technology is employed to transcribe speech into text in real-time, using natural language processing to analyze the content of teacher questions, student response logic, and keyword coverage, thereby mining the quality of classroom interactions. Data processing is divided into two phases: Early Fusion and Late Fusion. The former completes cross-modal data synchronization preprocessing at the frame, sentence, and temporal levels; the latter extracts visual features using ResNet and OpenPose models, extracts speech emotion and spectral features using CNN and LSTM, and analyzes text semantics using BERT and Transformer. Finally, high-dimensional features are fused through an attention mechanism to achieve classroom behavior classification and teaching evaluation decisions, providing data support for precise teaching optimization.

This architecture realizes an end-to-end closed-loop design from multi-source perception, feature learning to decision evaluation, with good scalability and practical application prospects, providing efficient and objective support for subsequent teaching optimization.

2.2 Modal feature extraction methods

To achieve precise perception and comprehensive evaluation of classroom teaching activities, the system performs deep feature extraction on three types of modal data: visual, speech, and text, constructing high-dimensional feature vectors that can be used for fusion modeling. Each modal feature extraction method adopts customized deep learning models based on its data characteristics to enhance feature expression capabilities and semantic understanding accuracy.

In the visual modality, the system utilizes the ResNet (Residual Neural Network) architecture to extract key emotional features from students' facial expression images. This network has strong image recognition and fine-grained classification capabilities, suitable for multi-class emotion recognition tasks, such as focus, confusion, fatigue, and joy. At the same time, to identify students' behavioral performance and participation status in the classroom, the system introduces Pose Estimation technology, selecting human keypoint detection models like OpenPose and MediaPipe to dynamically recognize and encode students' body movements (such as raising hands, writing, and bowing) and teachers' teaching methods (such as standing lectures and walking teaching). By combining facial expression and posture information, multidimensional visual behavior features reflecting student engagement and classroom atmosphere can be constructed.

In the speech modality, the system employs Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) to construct a speech feature extraction model. First, spectral features such as Mel-frequency cepstral coefficients (MFCC), Chroma, and Spectral Contrast are extracted from the audio to represent the basic physical properties of speech. Subsequently, the spectral features are input into CNN to extract spatial local features, combined with LSTM for temporal modeling to obtain dynamic information such as teachers' teaching rhythm and students' speech emotional changes. Additionally, the speech recognition front end achieves high-precision automatic speech transcription through Wav2Vec2 or DeepSpeech models, serving as the input basis for the text modality.

In the text modality, based on the classroom dialogue text transcribed from speech recognition results, the system uses BERT (Bidirectional Encoder Representations from Transformers) to perform contextual semantic modeling of the text, identifying semantic structures and keyword distributions. This model, through a bidirectional attention mechanism, can fully understand the meaning of classroom language in context, suitable for tasks such as identifying types of teacher questions and analyzing student response logic. On this basis, the system further utilizes the Transformer structure to analyze dialogue turns, semantic depth, and knowledge point coverage, thereby extracting text features that reflect the quality of classroom interactions and content coverage.

Through the deep feature extraction of the above three types of modalities, the system achieves a multidimensional characterization of classroom behaviors, emotional states, and language content, providing rich raw data support and feature foundation for subsequent multimodal fusion modeling and teaching quality evaluation.

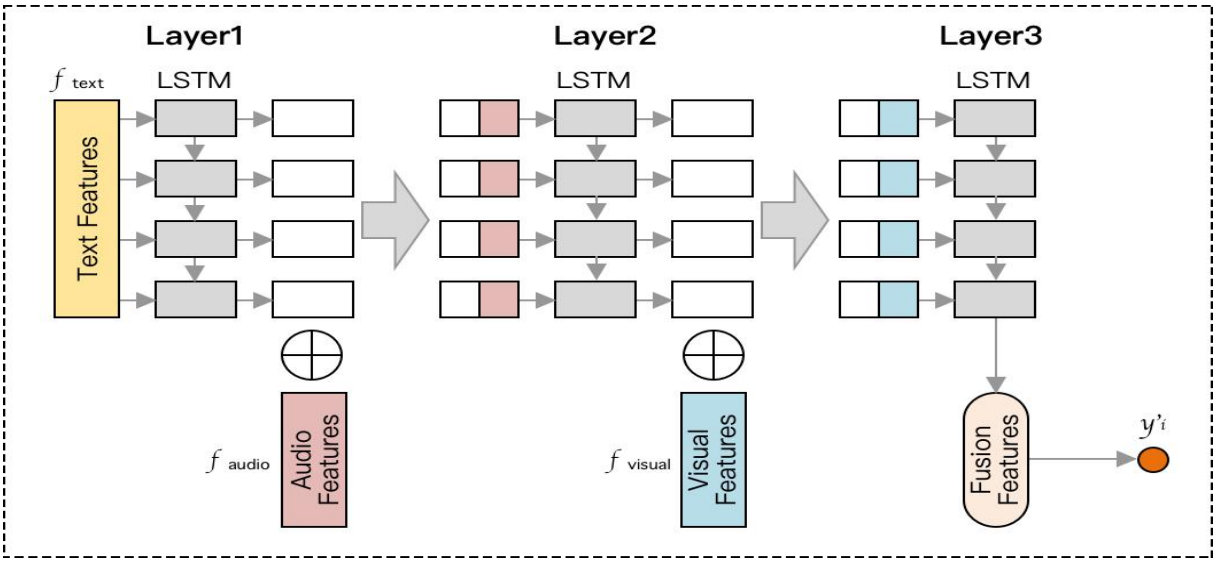
2.3 Multimodal Fusion Strategy

To achieve efficient perception and comprehensive analysis of classroom teaching behaviors, this system is designed and implemented based on the extraction of visual, speech, and text modal features, integrating deep learning and attention mechanisms through a Hybrid Fusion strategy. Multi-level information integration and modeling are performed during the input phase (Early Fusion) and decision phase (Late Fusion) to maximize the synergistic advantages of each modal feature, enhancing the model's generalization ability and evaluation accuracy.

In the Early Fusion phase, the system first completes the time synchronization and frame-level alignment of each modal data. Visual frames, speech streams, and text segments are mapped according to a unified timestamp, ensuring semantic consistency of multimodal features within the same time period. On this basis, feature concatenation is employed to initially fuse low-level feature vectors from ResNet, CNN+LSTM, and BERT models, constructing a fused representation vector as the joint input for the deep neural network. Early Fusion enhances the model's perception of cross-modal relationships, particularly suitable for assessing the comprehensive state of students at a given moment (e.g., visual focus + speech silence + short responses), aiding in fine-grained classroom behavior recognition.

In the Late Fusion phase, each modal feature is first modeled for high-level semantics through independent networks. The visual modality extracts feature maps via CNN/ResNet, the speech modality outputs temporal encodings through LSTM, and the text modality outputs contextual representations via BERT/Transformer. Subsequently, an attention mechanism or gated fusion network is used to jointly model these high-dimensional semantic representation vectors, achieving collaborative integration of multimodal semantics at the decision layer. Compared to Early Fusion, Late Fusion offers stronger interpretability and robustness, effectively avoiding modal noise interference while retaining important structural information from each modality.

Figure 1. Hybrid Fusion Model Diagram



Hybrid Fusion combines the cross-modal interaction advantages of Early Fusion with the model stability of Late Fusion, enhancing the system's ability to perceive complex classroom behaviors and improving fault tolerance for heterogeneous data, as shown in Figure 1. Experimental results indicate that the hybrid fusion structure outperforms single fusion strategies in classification tasks for key indicators such as student focus and classroom interaction quality, validating its applicability and superiority in multimodal classroom teaching evaluation.

3.Empirical Application and Result Analysis

3.1 Data Experiments and Evaluation Metrics

To verify the effectiveness of the constructed multimodal deep learning classroom teaching evaluation system, this study selected certain specialized classes from a university as empirical subjects, constructing a multimodal dataset in real teaching scenarios. Through comparative experiments and metric analysis, the system's performance and evaluation effects were comprehensively tested.

Regarding data sources, data collection was conducted in information technology classrooms and pedagogy classes, covering a total of approximately 60 teachers and students, with 12 classes collected, each lasting about 45 minutes. High-definition cameras, surround microphones, and voice collection devices were deployed in fixed teaching areas, synchronously recording the entire classroom process through legally authorized means. The collected data mainly includes:

Visual modality data: including video clips of student expressions, sequences of facial images, full-body posture behaviors (such as raising hands, writing, and bowing heads), and teacher movement trajectories;

Speech modality data: Includes recordings of teachers' lectures, teacher-student dialogues, and separated individual student speech segments;

Text modality data: Classroom speech transcribed into text using a speech recognition system (Wav2Vec2), and manually proofread for text analysis tasks.

In terms of experimental design, the study sets up three fusion strategy comparison experiments: Early Fusion, Late Fusion, and Hybrid Fusion. Each model uses uniform training parameters, batch size, and learning rate, and is trained and tested under the same dataset partition. The experimental tasks are divided into three sub-modules: student focus recognition, classroom interaction classification, and teacher lecture style assessment. Each task has manually annotated "gold standard" labels for supervised model learning and performance evaluation.

Additionally, to improve the reliability of the experimental results, this study employs 5-fold cross-validation in multiple rounds of experiments and sets up a noise scenario interference test group to simulate a non-ideal data environment in real teaching, in order to test the model's robustness and adaptability.

3.2 Experimental results

To comprehensively evaluate the performance of the constructed multimodal deep learning classroom evaluation system, this study analyzes from two dimensions: model classification accuracy and classroom teaching evaluation effectiveness. The experiments focus on student focus recognition, classroom interaction level classification, and teacher lecture style assessment as core tasks, comparing the performance of models under different fusion strategies, and demonstrating the system's interpretability and application potential with teaching examples.

In terms of model performance evaluation, the main metrics used are Accuracy, F1-score, and AUC (Area Under Curve). The performance of each fusion model on the test set is shown in Table 1:

Table 1. Performance comparison of different fusion models in key evaluation tasks

Model type	Accuracy	F1-score	AUC
Early Fusion	82.3%	0.781	0.841
Late Fusion	85.7%	0.816	0.873
Hybrid Fusion	89.1%	0.842	0.901

Note: The experiment is based on 5-fold cross-validation, covering three sub-tasks: student focus, classroom interaction levels, and teacher lecturing style.

The results show that the Hybrid Fusion model outperforms both Early Fusion and Late Fusion models

on all three core metrics, indicating its strong semantic expression ability and resistance to interference while maintaining information interaction between modalities, making it suitable for complex and dynamic classroom environments.

Table 2. Accuracy of Teaching Task Classification (using Hybrid Fusion as an example)

Sub-task Type	Recognition Accuracy
Student Focus Classification	87.60%
Classroom Interaction Level Classification	91.40%
Teacher Teaching Style Score	Correlation with Manual Evaluation: 0.71

Note: The teacher teaching style score is a composite index, with a Pearson correlation coefficient of 0.71 between it and student questionnaire feedback scores.

In the student attention recognition task, the system can accurately determine students' attention states based on visual behaviors (such as eye movement trajectories and expression changes), vocal participation (such as active speaking), and classroom dialogue text features. Test results show that the Hybrid Fusion model achieves an accuracy rate of over 85% in recognizing states such as "focused," "distracted," and "confused," demonstrating strong real-time feedback capabilities.

In the classroom interaction index classification task, the model scores and classifies interaction levels by statistically analyzing teacher questioning frequency, teacher-student dialogue turns, and student raising hands/responding behaviors. The results indicate that the recognition accuracy for high-interaction classrooms reaches 91.4%, effectively distinguishing between "one-way lecturing," "low interaction participation," and "high-frequency interactive" classroom modes, providing an objective basis for optimizing teaching behaviors.

In the teacher teaching style assessment task, the system conducts comprehensive scoring modeling by analyzing indicators such as speech rate (WPM), intonation fluctuations, keyword coverage, and question openness. The teaching quality scores of teacher samples show a positive correlation with student post-class questionnaire feedback (Pearson coefficient of 0.71), further validating the rationality and credibility of the model evaluation.

Additionally, combined with the system's visualization module, personalized classroom feedback reports can be generated for teachers, including "Student Attention Heatmap," "Classroom Interaction Structure Diagram," and "Teaching Language Logical Structure Diagram," providing data support for subsequent teaching reflection and improvement. This system has demonstrated stable and accurate performance across multiple teaching evaluation tasks, validating the practicality and scalability of integrating multimodal deep learning models in intelligent classroom analysis.

3.3 Application Feedback and Teaching Optimization Suggestions

Based on the aforementioned model analysis results and classroom data mining, the multimodal classroom evaluation system constructed in this study not only possesses high evaluation accuracy and stability but also provides targeted teaching optimization suggestions for frontline teachers and educational managers, thereby promoting classroom teaching towards a more efficient, interactive, and intelligent direction. The specific optimization suggestions and system application value are mainly reflected in the following three aspects:

First, improve the quality of teachers' instruction by optimizing teaching language expression and classroom rhythm control. The system can comprehensively generate a "teaching style profile" based on teachers' speech data (such as speed, tone, and pause frequency) and textual data (like keyword coverage and questioning techniques) for the comprehensive design of teaching language expression. If teachers exhibit phenomena such as fast speech, monotonous tone, or insufficient coverage of content, the system will automatically suggest appropriate adjustments to the language rhythm of key knowledge points, enrich emotional expression, or provide supplements. Additionally, for teaching segments with low questioning frequency and a high proportion of closed questions, the system will prompt the addition of open-ended and heuristic questions to strengthen cognitive challenges in the classroom, enhancing students' deep thinking and expression abilities.

Second, enhance students' classroom focus and support personalized classroom guidance. The system identifies students' attention levels in real-time and automatically generates a "student focus report" by integrating data from visual expression features, head orientation, and vocal responses, thereby improving students' classroom focus. For students who have been less participative or have scattered attention for a long time, the system can prompt teachers to actively ask questions, design interactive tasks, and enhance participation motivation during the teaching process. Furthermore, teachers can utilize post-class focus trend graphs to analyze the correlation between teaching content and changes in student attention, allowing for reasonable design of teaching rhythm and content structure.

Third, promote the data-driven transformation of teaching behavior reflection and research decision-making. The system supports the aggregation and analysis of historical evaluation data from multiple classroom samples, which can depict the evolution trends of teachers' teaching styles, the changing paths of classroom interaction levels, and the long-term trajectories of student participation, providing objective evidence for teaching research groups to conduct collective lesson preparation and heterogeneous analysis. At the same time, educational managers can identify high-frequency points of teaching issues based on classroom evaluation data from classes and grades, and develop targeted plans for teaching research, training, and quality improvement, thereby promoting the overall level of classroom teaching quality management in schools.

This research not only achieves precise quantification and multidimensional analysis of classroom teaching activities but also empowers teaching improvement practices through data-driven insights, showcasing significant potential for broader application. It holds important value in constructing an intelligent teaching evaluation ecosystem and promoting the digital transformation of education.

4.Conclusion and Outlook

This research addresses the issues of subjectivity, poor real-time performance, and single-dimensional information in traditional classroom teaching evaluations by constructing a multimodal deep learning-based classroom teaching evaluation system. It integrates three types of technologies: computer vision, speech processing, and natural language processing, and achieves comprehensive assessments of student focus, classroom interaction levels, and teacher instruction quality through a multi-layer Hybrid Fusion model. Experimental results indicate that the model proposed in this study demonstrates good accuracy and robustness in key evaluation tasks related to student focus, classroom interaction levels, and teacher instruction quality, possessing practical value and promising prospects for promotion. Its main contributions include: first, the introduction of multimodal deep learning fusion strategies into teaching evaluation for more objective and intelligent assessments; second, the formation of a technical path from data collection, feature extraction to fusion modeling, resulting in good generalization capabilities for the system; third, the ability to automatically generate real-time feedback and teaching optimization suggestions for classrooms, providing decision-making data for precise teaching and teaching management.

Although this research has achieved some phased results, certain limitations still exist. First, the number of samples selected for the study is relatively small, and the distribution of subjects and educational stages is quite singular; second, in some complex environments, the model's response speed and interactive feedback still have room for improvement.

Future research will expand in the following areas: first, studying multimodal small sample learning and transfer learning technologies to enhance the model's generalization capabilities; second, introducing causal inference and explainable AI mechanisms to make the logic of teaching evaluation results clearer and more accurate; third, integrating and deploying the teaching platform with this system to improve applications in intelligent classrooms and educational big data analysis. This research provides theoretical support and technical pathways for constructing a new intelligent classroom teaching evaluation system, offering beneficial explorations for future educational quality enhancement and educational evaluation reform.

References

[1] Haleem A. Understanding the role of digital technologies in education: A review[J]. 2022. ,2022.

[2] Jiao T, Guo C, Feng X, Chen Y, Song J. A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications[J]. Computers, Materials & Continua, 2024, 80(1): 1–35.

[3] Gao Y. Deep learning-based strategies for evaluating and enhancing university teaching quality[J]. Computers and Education: Artificial Intelligence, 2025, 8: 100362.

[4] Q L, X J, R J. Classroom Behavior Recognition

Using Computer Vision: A Systematic Review[J]. PubMed, 2025[2025-04-09].

[5] Southwell R, Pugh S, E. Margaret Perkoff, Clevenger C, Bush J, Lieber R, Ward W, Foltz P, D'Mello S. Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms[J]. 2022. Zenodo, 2022[2025-04-09].

[6] Yaneva V, Von Davier M. Advancing Natural Language Processing in Educational Assessment[M]. 1. New York: Routledge, 2023[2025-04-09].

[7] Li S, Tang H. Multimodal Alignment and Fusion: A Survey[A]. arXiv, 2024[2025-04-09].

[8] Pawłowski M, Wróblewska A, Sysko-Romańczuk S. Effective Techniques for Multimodal Data Fusion: A Comparative Analysis[J]. Sensors, 2023, 23(5): 2381.