Publication Date: 02 September 2026
Archs Sci. (2026) Volume 75, Issue 6 Pages 76-85, Paper ID 2025609.
https://doi.org/10.68304/as/75609

Research on a Teaching Improvement Model for Higher Vocational Midwifery Education Integrating AI Assessment and Classroom Interaction Analysis

Yan Zhu and Jian Wang
Weinan Vocational and Technical College, Weinan, 714000, China
Corresponding authors: Yan Zhu (e-mail: mujin33765@163.com).

Abstract This study proposes a teaching-improvement model for midwifery education in higher vocational colleges that integrates artificial-intelligence-based assessment with classroom interaction analysis. The model is intended to enhance teaching quality and students' learning experience by using AI techniques to capture dynamic information that is difficult to obtain through conventional end-of-course assessment. A dual-path architecture combining ELMo-BiLSTM-CRF with a Transformer encoder is employed to analyze classroom interaction data across multiple modalities, including speech, text, and behavioral information, thereby supporting emotion recognition, engagement assessment, and feedback on instructional strategies. The study further incorporates a closed-loop mechanism in which AI-generated assessment results are linked to instructional adjustment so that classroom interaction, learner affect, and teaching decisions can be analyzed within a unified framework. The experimental sections report model-performance analyses, group comparisons, and multimodal visualization results. In classroom interaction emotion recognition, the proposed dual-path model achieved an accuracy of 89.7%, a recall of 90.3%, and an F1 score of 90.0%, outperforming the comparative approaches reported in the study. Classroom engagement in the AI-enhanced group reached 85.9%, compared with 68.4% in the control group, while the Emotion Improvement Index increased from 0.21 to 0.43. Additional analyses illustrate how the framework can capture changes in learner emotion and engagement, examine the influence of regulatory mechanisms, and extract information relevant to teaching improvement. Overall, the results support the potential of AI-assisted classroom analytics as a decision-support tool for data-informed improvement in vocational midwifery education.

Index Terms artificial intelligence assessment, classroom interaction analysis, higher vocational midwifery education, ELMo-BiLSTM-CRF, Transformer, emotion recognition, teaching improvement

I. Introduction

In recent years, the continuous development of higher vocational education has placed increasing emphasis on the quality, relevance, and practice-oriented nature of professional training. Midwifery, as an important branch of nursing and maternal-health education, requires students to develop not only disciplinary knowledge but also communication competence, emotional resilience, clinical judgment, and the ability to respond appropriately in complex interpersonal situations. Conventional midwifery teaching in higher vocational colleges generally combines classroom instruction with practical or clinical training; however, teaching evaluation often relies heavily on final examinations, teacher judgment, and limited classroom observation. Such approaches provide only a partial view of the learning process and may fail to capture changes in classroom participation, emotional state, interaction quality, and learning engagement in a timely manner. Consequently, instructors may receive delayed or incomplete feedback about students' learning difficulties, while students' emotional fluctuations and interaction patterns remain insufficiently quantified. These limitations create a need for more objective, data-informed, and responsive approaches to teaching improvement.

The rapid development of artificial intelligence (AI), particularly deep learning, natural language processing (NLP), learning analytics, and multimodal sensing, provides new possibilities for educational assessment. AI-supported systems can process classroom speech, textual interaction, behavioral signals, and other learning traces to identify patterns that are difficult to observe consistently through manual evaluation alone. In healthcare education, virtual simulation, intelligent feedback, adaptive learning, and AI-supported assessment have increasingly been explored as tools for improving learning experiences and professional preparation [1]. Midwifery education is also beginning to consider how AI may support curriculum design, learner feedback, simulation, and professional readiness [2]. Related work in nursing simulation further indicates that AI can be incorporated into simulation-based education to support learning and evaluation [3]. In addition, case-based learning combined with virtual-reality simulation has shown value in midwifery laboratory education, particularly for practical ability, self-directed learning, and comprehensive competence [4]. These developments suggest that AI-assisted classroom analytics may complement, rather than replace, professional teaching judgment.

The educational objective of midwifery training extends beyond the transmission of factual knowledge. Students must develop professional communication skills, psychological stability, collaborative behavior, and the ability to integrate theoretical understanding with clinical reasoning. Classroom interaction therefore plays an important role in knowledge construction, emotional support, and the development of professional confidence. Nevertheless, vocational midwifery classrooms may contain uneven participation, anxiety, hesitation, and differences in learning engagement. When these dynamic states are not captured systematically, instructors may have difficulty determining when and how to adjust teaching strategies. Contemporary discussions of AI in nursing education emphasize both opportunities and challenges, including the potential for individualized learning support, simulation, intelligent feedback, and ethical concerns related to transparency and bias [5]. Empirical and review-based studies also indicate that AI-supported educational interventions can contribute to professional learning when they are appropriately integrated into pedagogical design [6], [7].

The application of AI-based assessment in education has expanded from isolated classification tasks toward broader learning-behavior modeling, simulation, and feedback systems. Systematic evidence from health education shows growing interest among students and academics in AI-supported learning and practice, while also emphasizing the need for appropriate training, governance, and curriculum integration [8]. In obstetric nursing, interactive virtual-reality communication simulation supported by ChatGPT has been investigated as a way to strengthen communication competence and learning experience [9]. AI has also been used for adaptive scenario development in interprofessional education [10] and for scenario-based simulation in medical education [11]. These developments are consistent with broader pedagogical frameworks that combine classroom learning, high-fidelity simulation, and AI-enhanced extended reality [12]. For midwifery specifically, technology-supported childbirth training and gamification have also demonstrated the value of interactive learning environments [13]. However, the direct integration of AI-based classroom interaction analysis with midwifery teaching-improvement decisions remains comparatively underdeveloped.

Despite these advances, several technical and pedagogical challenges remain. Classroom interaction data are heterogeneous and may include colloquial language, professional terminology, spontaneous speech, non-verbal behavior, and emotionally sensitive expressions. A single data modality may therefore provide an incomplete representation of students' classroom states. Moreover, AI-generated outputs become educationally useful only when they can be translated into interpretable and actionable teaching feedback. Recent reviews of AI-based pedagogies in nursing education emphasize the need to connect technical innovation with sound instructional design and responsible implementation [14]. Evidence from labor-support training further illustrates that interactive and experiential strategies can improve learning performance and core competencies when teaching design is aligned with professional tasks [15]. Personalized adaptive learning has also been examined in nursing and midwifery contexts as a means of tailoring instructional support to learner characteristics and performance [16]. These findings reinforce the importance of combining computational assessment with pedagogical interpretation rather than treating model output as an isolated end point.

To address these issues, this paper develops a teaching-improvement model for higher vocational midwifery education that integrates AI-based assessment with classroom interaction analysis. The proposed framework is designed around three complementary objectives. First, contextual language representations are used to improve the interpretation of classroom interaction text containing ambiguous and domain-specific expressions. Second, multimodal classroom information is incorporated so that textual, speech, video, and behavioral evidence can be considered jointly. Third, assessment results are connected to a feedback loop that supports adjustment of teaching strategies, thereby linking data analysis to classroom decision-making. This orientation is consistent with current efforts to improve AI literacy among healthcare learners and educators [17], with research on the use of AI by midwifery students [18], and with multimodal learning-analytics systems developed for team-based healthcare simulation [19].

At the computational level, the model uses ELMo-based contextual embeddings to reduce ambiguity in classroom interaction text, BiLSTM to capture bidirectional sequential dependencies, and CRF decoding to model dependencies among output labels. These components are intended to support the identification of student emotions, interaction behaviors, and engagement-related states. The technical basis for contextualized word representation and sequence labeling is well established in the NLP literature [23]–[25].

In addition to text-based interaction data, the framework is designed to incorporate classroom audio, video, and facial-expression information. The purpose of this multimodal design is not simply to increase the amount of input data, but to provide complementary evidence about engagement and affect when a single modality is incomplete or ambiguous. Such an approach is aligned with the wider digital transformation of nursing education, in which simulation, multimodal technologies, and digitally mediated learning are increasingly integrated into curriculum development [20].

The proposed framework therefore extends beyond AI-based classification alone. Its intended role is to link assessment outputs with instructional strategies, forming a closed-loop relationship among teachers, students, and the analytical system. Within this loop, model outputs are interpreted as decision-support information rather than as autonomous judgments. This distinction is important in healthcare education, where educational quality, professional accountability, and fairness must remain under human oversight. Broader reviews of technology integration in medical curricula similarly emphasize that successful digital transformation requires not only technological capability but also curriculum redesign, faculty preparation, and responsible implementation [21].

II. Abortion-Related Biomedical Named Entity Recognition

Biomedical named entity recognition (Bio-NER) identifies clinically or biologically relevant entities, such as diseases, drugs, symptoms, procedures, genes, and proteins, from unstructured biomedical text. In the present technical pipeline, abortion-related biomedical text is used as a domain-specific language source for contextual representation learning. The entity-recognition process consists of two fundamental tasks: identifying entity boundaries and assigning an appropriate entity category. Neural sequence-labeling architectures provide a well-established basis for these tasks [24], [25].

A. Overall Model Architecture

Distributed word representations are widely used in natural language processing because they transform discrete lexical items into dense vectors that can be processed by neural models. Conventional static embeddings, however, assign a single vector to a word regardless of context and therefore cannot adequately represent polysemy or context-dependent terminology. This limitation is particularly relevant in biomedical and educational text, where abbreviations, professional terms, and semantically ambiguous expressions occur frequently. Contextualized representations address this problem by generating a word representation that depends on the surrounding sentence. ELMo is one such approach: it derives task-sensitive contextual embeddings from a bidirectional language model and combines information from multiple internal layers [23]. Chinese named-entity recognition also benefits from models that explicitly represent character and lexical structure, as demonstrated by lattice-LSTM approaches [26]. Figure 1 summarizes the ELMo component used in the present architecture.

A corpus of Chinese abortion-related biomedical text was used in the source study to pre-train the Chinese ELMo component. After training, each input token is associated with representations derived from the token layer and the forward and backward language-model layers. These representations are subsequently combined through learned task-specific weights. For a sequence of N words, $(t_1, t_2, \ldots, t_N)$, the forward language model factorizes the sequence probability by predicting the token at position k from the preceding context, as shown in Eq. (1).

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1}). \tag{1}$$

The backward language model is defined analogously, except that each token is predicted from the subsequent context, as shown in Eq. (2).

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N). \tag{2}$$

The bidirectional language model is trained by maximizing the sum of the forward and backward log-likelihoods. The corresponding objective is given in Eq. (3).

$$\sum_{k=1}^{N} \left[\log p\left(t_k \mid t_1, \ldots, t_{k-1}; \theta_x, \vec{\theta}_{LSTM}, \theta_s\right) + \log p\left(t_k \mid t_{k+1}, \ldots, t\right.\right. \tag{3}$$

For a bidirectional language model with L internal layers, ELMo constructs multiple representations for each token by combining the forward and backward hidden states. The representation set at position k is written as

$$\begin{aligned} R_k &= \left\{x_k^{LM}, \left[\vec{h}_{k,j}^{LM}, \overleftarrow{h}_{k,j}^{LM}\right] : j = 1, \ldots, L\right\} \\ &= \left\{h_{k,j}^{LM} : j = 0, \ldots, L\right\}, \end{aligned} \tag{4}$$

where x_k^{LM} denotes the context-independent token representation and $h_{k,j}^{LM} = [\vec{h}_{k,j}^{LM}, \overleftarrow{h}_{k,j}^{LM}]$ denotes the concatenated forward and backward representation at layer j.

For a specific downstream task, the layer representations are combined into a single contextual vector using learned normalized weights, as shown in Eq. (5).

$$ELMo_k^{task} = E\left(R_k; \theta^{task}\right) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} h_{k,j}^{LM}, \tag{5}$$

where s_j^{task} is obtained through a softmax-normalized weighting scheme and γ^{task} is a task-specific scaling parameter.

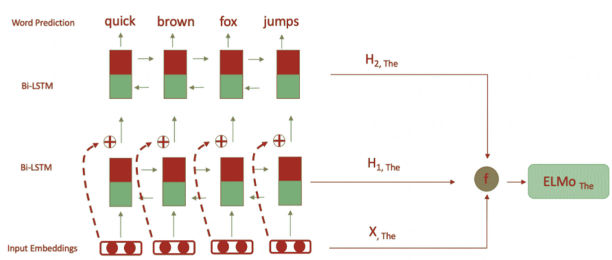

Figure 1: Contextual Representation Process in the ELMo Component

For an input sentence X, the pre-trained ELMo model generates contextual representations for each token. The layer-specific representations are assigned learned weights and combined into a single context-sensitive embedding. This embedding is then supplied to the BiLSTM-CRF sequence-labeling network as an additional feature. The BiLSTM captures bidirectional context, while the CRF layer models dependencies among neighboring labels and returns a globally consistent label sequence. Similar neural sequence-labeling strategies are widely used for named-entity recognition [24], [25], while contextualized biomedical language models such as BioBERT demonstrate the value of domain-specific pre-training for biomedical text mining [28].

B. Classroom Interaction Prediction Mechanism

Figure 2 illustrates the context-prediction mechanism used to obtain an interaction-related target label. In the schematic example, surrounding words provide contextual evidence for the target position. In the classroom application, the target is not the example word itself but an interaction label, such as an affective state or engagement category.

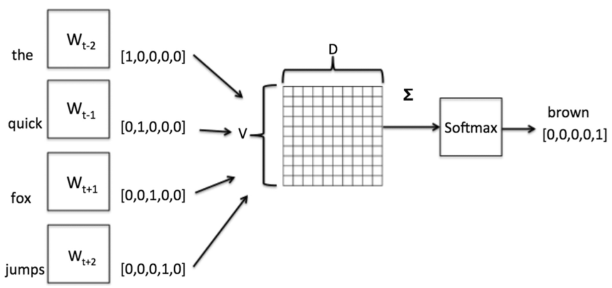

Figure 2: Structure of the Context Prediction Mechanism

The context tokens $W_{t-2}, W_{t-1}, W_{t+1}, W_{t+2}$ are mapped through the embedding matrix V to dense vector representations. The resulting vectors encode local contextual evidence surrounding the target position and provide the input to the attention-weighted context representation.

The contextual vectors are then summed or weighted to form the representation $\boldsymbol{h}_t$:

$$\boldsymbol{h}_t = \sum_{\substack{i=t-k \\ i \neq t}}^{t+k} \alpha_i \boldsymbol{v}_i, \tag{6}$$

where α_i denotes a learned attention weight. This formulation allows the model to assign greater importance to context positions that are more informative for the target interaction label.

Finally, a softmax layer is used to estimate the probability of the target classroom-interaction label:

$$P(y_t \mid context) = \frac{\exp\left(\boldsymbol{u}_{y_t}^{\top} \boldsymbol{h}_t\right)}{\sum_{j=1}^{|Y|} \exp\left(\boldsymbol{u}_j^{\top} \boldsymbol{h}_t\right)}. \tag{7}$$

In the proposed classroom scenario, the model processes interaction data such as transcribed student responses and discussion speech to estimate emotional states and engagement-related behaviors. These predictions are subsequently transferred to the teaching-feedback module. The purpose is threefold: to identify learners who may require additional support, to provide instructors with evidence for adjusting interaction methods and teaching pace, and to establish a closed-loop improvement process in which repeated AI-assisted assessment informs subsequent instructional decisions.

C. Analysis of Transformer Encoder Submodule Interaction in AI Classroom

Figure 3 presents the Transformer encoder submodule incorporated into the proposed teaching-improvement model. The purpose of this branch is to capture long-range dependencies and global contextual relationships within classroom interaction sequences. Its architecture contains an input-embedding layer, Transformer encoder blocks based on multi-head self-attention and feed-forward networks, an output representation layer, and a prediction layer. In combination with the ELMo-BiLSTM-CRF branch, the Transformer provides a complementary global-context pathway. The self-attention architecture follows the general Transformer formulation introduced by Vaswani et al. [29].

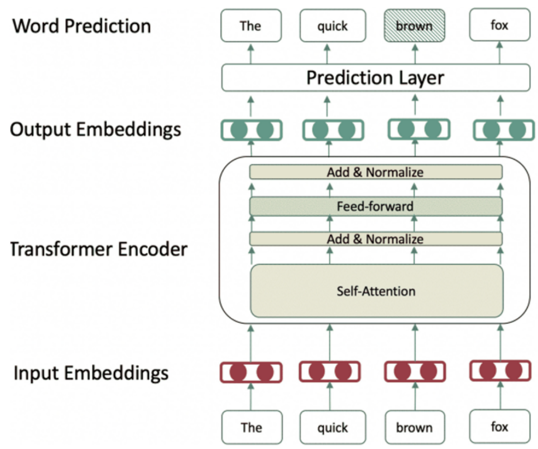

Figure 3: Transformer Encoder Submodule Used for Classroom Interaction Analysis

For a classroom interaction sequence $W = [w_1, w_2, \ldots, w_n]$, each token is first mapped to a dense vector and combined with positional information:

$$\boldsymbol{x}_i = \boldsymbol{E}(w_i) + \boldsymbol{P}_i, \tag{8}$$

where $\boldsymbol{E}$ denotes the word-embedding function and $\boldsymbol{P}_i$ is the positional-encoding vector at position i. This representation allows the encoder to preserve both lexical information and token order.

The core computational components are multi-head self-attention and a position-wise feed-forward neural network. Together, these components allow the model to integrate evidence from multiple positions while learning nonlinear transformations of the contextual representation.

The self-attention mechanism estimates the relevance of each sequence position to the others and forms a context-dependent representation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V, \tag{9}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the key dimension used for scaling.

The multi-head formulation performs several attention operations in parallel so that different relational patterns can be represented simultaneously:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V), \tag{10}$$

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^O. \tag{11}$$

The attention output is combined with a residual connection and normalized:

$$\boldsymbol{H}' = LayerNorm\left(X + MultiHead(Q,K,V)\right). \tag{12}$$

A feed-forward network then applies an additional nonlinear transformation:

$$\boldsymbol{H}'' = FFN(\boldsymbol{H}') = \max\left(0, \boldsymbol{H}'W_1 + b_1\right)W_2 + b_2. \tag{13}$$

This stage increases the representational capacity of the encoder after contextual information has been aggregated through attention.

1) Output representation

The contextual representation is transferred to the prediction stage after the feed-forward output is combined with the residual pathway:

$$\boldsymbol{H} = LayerNorm\left(\boldsymbol{H}' + \boldsymbol{H}''\right). \tag{14}$$

The resulting matrix retains sequence-level contextual information and provides the input features required for classroom interaction classification.

2) Prediction layer

For token- or interaction-level classification, a softmax layer can be used to estimate the probability of each classroom interaction label:

$$P(y_i \mid W) = \frac{\exp\left(\boldsymbol{u}_{y_i}^\top \boldsymbol{h}_i\right)}{\sum_{j=1}^{|Y|} \exp\left(\boldsymbol{u}_j^\top \boldsymbol{h}_i\right)}. \tag{15}$$

For sequence-labeling tasks in which neighboring output labels are not independent, CRF decoding can replace independent softmax decoding so that the final prediction accounts for transitions among labels and returns a globally optimized label sequence.

Within the overall framework, the input layer represents transcribed classroom interaction text, the Transformer encoder captures global contextual dependencies, and the output representation is passed to the AI-assessment module to estimate student emotion and interaction-related labels. The ELMo-BiLSTM-CRF branch provides complementary contextual and sequential modeling, while the Transformer branch strengthens long-range feature interaction. Pre-trained Transformer models such as BERT and biomedical variants such as BioBERT further demonstrate the effectiveness of contextual pre-training for language understanding and domain-specific text analysis [27], [28]. The dual-path design is therefore intended to combine local sequence sensitivity with global contextual modeling.

III. Data and Methods

A. General data

The source manuscript reports that 500 abortion patients admitted to the participating hospital were included as study subjects. Participants were allocated to a control group and an experimental group according to odd and even consultation serial numbers, with 250 participants in each group [21]. The reported age range in the control group was 23–34 years, with a mean age of 28.7 ± 4.6 years, whereas participants in the experimental group were 22–37 years old, with a reported mean age of 26.4 ± 10.1 years. The source further states that participants provided informed consent after admission, that the study was approved by the hospital medical ethics committee, and that complete medical records were available. Because these participant details are clinically oriented and differ from the educational sample described elsewhere in the manuscript, the original records should be checked before publication.

B. Evaluation Indicators

The source manuscript reports the use of the Self-Rating Anxiety Scale (SAS) and Self-Rating Depression Scale (SDS) to assess psychological status. For the SDS, scores of ≤ 53 were interpreted as indicating no depression, scores of 53–62 as mild depression, scores of 63–72 as moderate depression, and scores of > 72 as severe depression. Higher scores therefore indicate more severe depressive symptoms. The corresponding SAS score was interpreted in the same direction, with higher values indicating greater anxiety. The repeat-abortion or repeat-miscarriage rate was also recorded and compared between the two groups. These clinical outcomes are reproduced from the supplied manuscript and should be reconciled with the classroom-education outcomes described in the abstract and later experimental sections.

IV. Results

A. Comparison of the Repeat Miscarriage Rate Between the Two Groups

In the control group, 21 participants (8.40%) experienced repeat miscarriage, whereas 2 participants (0.80%) in the experimental group experienced repeat miscarriage. The reported rate was therefore lower in the experimental group. Because the source manuscript does not provide the inferential test for this comparison in the accompanying text, the result is presented descriptively rather than as an independently verified claim of statistical significance.

B. Comparison of SDS and SAS Scores of Patients in the Two Groups

Before treatment, the differences in SDS and SAS scores between the experimental and control groups were not statistically significant ($p > 0.05$), as reported in Table 1. After treatment, both measures were lower in the experimental group than in the control group, and the table reports $p < 0.001$ for the between-group comparisons after treatment.

Table 1 therefore provides the numerical basis for the reported changes in anxiety- and depression-related scores.

Table 1: Comparison of SDS and SAS Scores Between Two Groups of Patients (Score, $\overline{x} \pm s$)

Group	SDS score		SAS score	
	Before treatment	After treatment	Before treatment	After treatment
Control group	58.75±2.88	55.12±5.98*	61.57±4.01	59.78±6.01*
Experimental group	58.23±3.45	32.01±5.31*	62.12±5.11	44.23±5.13*
t value	1.5101	49.2115	1.5897	32.9912
p value	0.1423	0.0000	0.1128	0.0000

Note: An asterisk indicates a within-group difference relative to the corresponding pre-treatment value at $p < 0.05$, as reported in the source manuscript.

C. Performance Analysis

A Chinese biomedical corpus was constructed by combining translated English biomedical material with manually annotated Chinese text. The purpose of this auxiliary experiment was to evaluate the domain-specific entity-recognition and relation-extraction components used in the language-processing pipeline. The named-entity recognition model employs contextualized representation together with BiLSTM-CRF sequence labeling, while comparison models include BiLSTM+CRF, Lattice LSTM, and BERT+BiLSTM+CRF. The principal hyperparameters are summarized in Table 2, and the corresponding entity-recognition and relation-extraction results are reported in Table 3.

Table 2: Hyperparameter Settings for Chinese Named Entity Recognition

Parameter	Set value	Parameter	Set value
Character embedding size	100	LSTM hidden	100
Character dropout	0.6	Sentence length	400
LSTM layer	1	Regularization	1e-8
Learning rate	0.016	Learning-rate decay	0.06

As shown in Table 3, ELMo+BiLSTM+CRF achieved an F1 score of 84.81% on the reported Chinese entity-recognition task. The value is slightly lower than the BERT+BiLSTM+CRF F1 score of 84.89%, although ELMo+BiLSTM+CRF achieved the highest recall among the listed entity-recognition models. Accordingly, the revised interpretation avoids claiming that ELMo+BiLSTM+CRF is uniformly superior across all metrics. Lattice-LSTM provides an appropriate Chinese NER comparison because it explicitly incorporates lexical information into recurrent sequence modeling [26], while BERT provides a contextual Transformer baseline [27].

Table 3: Chinese Entity Recognition and Relation Extraction Results

Task Type	Model	Precision (%)	Recall %	F_1 score (%)
Entity Identification	BiLSTM+CRF	82.31	82.10	81.99
	Lattice LSTM	84.56	83.45	83.97
	BERT+BiLSTM+CRF	84.95	84.77	84.89
	ELMo+BiLSTM+CRF	84.32	85.76	84.81
Relationship Extraction	BiLSTM+Attention	79.29	78.62	79.01
	BiLSTM+Attention+Feature	81.35	80.46	80.98

For biomedical relation extraction, the study uses a bidirectional long short-term memory network with an attention mechanism and introduces a stroke-based feature at the input layer. The corresponding hyperparameters are listed in Table 4. According to Table 3, the BiLSTM+Attention+Feature configuration achieved an F1 score of 80.98%, compared with 79.01% for BiLSTM+Attention without the additional feature. The reported improvement is therefore 1.97 percentage points. This result supports the interpretation that the added feature contributed useful information for the reported Chinese relation-extraction task, while avoiding the stronger claim that the feature would necessarily generalize to other corpora or domains.

Table 4: Hyperparameter Settings for Chinese Relation Extraction

Parameter	Set value	Parameter	Set value
Batch size	100	LSTM hidden	200
Embedding size	100	Sentence length	100
LSTM layer	1	Regularization	1e-8
Dropout	0.016	Learning-rate decay	0.016

D. Comparison of Model Prediction Performance

Table 5 compares the performance of the evaluated methods for classroom interaction emotion recognition in terms of accuracy, recall, and F1 score.

Table 5: Performance Comparison of Different Methods in Classroom Interaction Emotion Recognition

Model	Accuracy (%)	Recall (%)	F1 score (%)
Manual evaluation	72.1	70.4	71.2
BiLSTM+CRF	83.4	82.9	83.1
Transformer Encoder	85.6	84.7	85.1
ELMo-BiLSTM-CRF + Transformer	89.7	90.3	90.0

The results in Table 5 show a clear performance progression across the four evaluated approaches. Manual evaluation produced an accuracy of 72.1%, a recall of 70.4%, and an F1 score of 71.2%. The BiLSTM+CRF model increased the F1 score to 83.1%, indicating that sequence modeling and structured decoding capture information that is not consistently represented by manual assessment alone. The Transformer encoder achieved an F1 score of 85.1%, consistent with the ability of self-attention to model longer-range contextual relations. The proposed dual-path ELMo-BiLSTM-CRF+Transformer model obtained the highest reported values: 89.7% accuracy, 90.3% recall, and a 90.0% F1 score. Relative to the Transformer encoder, the absolute F1 improvement is 4.9 percentage points. The complementary architecture provides a plausible explanation for this increase because contextualized ELMo embeddings reduce lexical ambiguity, BiLSTM-CRF models local sequential dependencies and label transitions, and the Transformer strengthens global contextual interaction. These results indicate that the combined model is promising for classroom interaction emotion recognition, although external validation on additional cohorts would be required before broader generalization. The importance of rigorous educational evaluation is consistent with recent mixed-methods work on nursing-program quality and student competence [22].

Table 6 compares classroom engagement and the Emotion Improvement Index (EII) between the control group and the AI-enhanced experimental group after course completion. Classroom engagement was 68.4% in the control group and 85.9% in the experimental group, corresponding to an absolute difference of 17.5 percentage points. The EII increased from 0.21 to 0.43, which is approximately twice the control-group value. These descriptive differences are consistent with the intended function of the framework: identifying emotion and interaction patterns and returning interpretable feedback that can support adjustment of classroom strategies. However, Table 6 does not report dispersion estimates, confidence intervals, or inferential test results. For that reason, the revised text describes the observed differences without labeling them statistically significant. The results should be interpreted as evidence from the reported sample rather than as proof of a universal effect.

Table 6: Comparison of Engagement and Emotion Improvement Index Between the Experimental and Control Groups After Course Completion

Group	Classroom Engagement (%)	Emotion Improvement Index (EII)
Control Group	68.4	0.21
Experimental Group (AI-enhanced Model)	85.9	0.43

Figure 4 illustrates a published example of racial bias in contextual clinical language modeling. For the prompt "[RACE] pt became belligerent and violent. sent to [TOKEN] [TOKEN]," the displayed SciBERT outputs differ across racial descriptors: the model associates White/Caucasian descriptors with "hospital" more often in the illustrated example and Black/African American descriptors with "prison." This example originates from prior work on bias in clinical contextual word embeddings [30]; it is not an independently generated result of the present classroom experiment. Its relevance here is methodological: AI-based educational assessment may reproduce undesirable associations if training data, representation learning, or evaluation procedures contain systematic bias. Consequently, classroom analytics should incorporate dataset review, subgroup performance checks, human oversight, and explicit fairness auditing before model outputs are used to support teaching decisions.

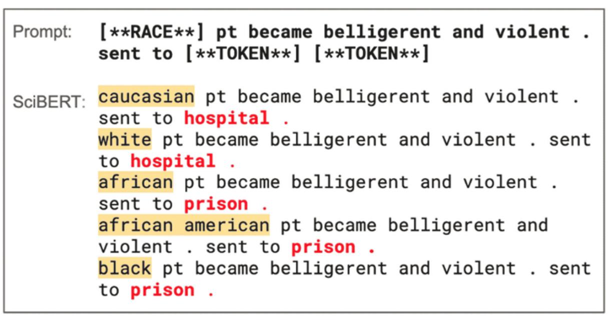

Figure 4: Illustration of Racial Bias in SciBERT Fill-in-the-Blank Predictions

Figure 5 presents a supplementary analysis of teaching experience and educational level across six classroom dimensions: learning atmosphere, classroom management, instructional clarity, activating instruction, differentiated instruction, and teaching–learning strategies. The plotted estimates indicate that the direction and magnitude of associations vary across teacher subgroups and classroom dimensions. In particular, the displayed coefficients suggest stronger positive associations between teaching experience and several instructional dimensions for male teachers, whereas the education-level coefficients are more heterogeneous. These patterns should be interpreted cautiously because the source manuscript does not provide the full sampling, regression, or model-specification details required for independent verification of the plotted coefficients. Within the proposed AI-assisted teaching framework, the figure is therefore used as an example of how instructor-level variables could be incorporated into diagnostic analysis rather than as evidence that demographic characteristics deterministically produce particular teaching outcomes.

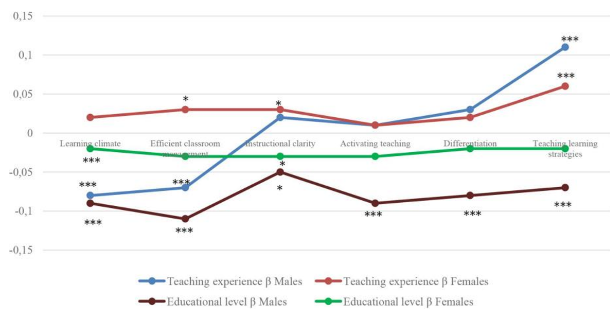

Figure 5: Differential Associations of Teaching Experience and Educational Level with Classroom Dimensions

Figure 6 shows a random-slope representation of the relationship between perceived classroom engagement and eight emotional states: joy, hope, pride, anger, anxiety, shame, helplessness, and burnout. The displayed trends associate greater engagement with higher positive-emotion scores and lower scores for several negative emotions, while the strength of the relationship varies across individuals. Anxiety and helplessness show comparatively pronounced downward trends in the plotted analysis, whereas burnout changes less strongly. The relevance of this analysis to the proposed framework is that classroom emotion should be modeled dynamically rather than treated as a fixed learner characteristic. An AI-assisted system can use repeated interaction evidence to identify changing affective patterns, but such outputs should be interpreted together with pedagogical context and, where appropriate, direct communication with students.

Figure 7 examines the relationship between person-mean-centered classroom participation and several emotional states while considering internal and external regulation as moderators. The displayed models indicate that higher participation is generally associated with more positive affect and lower negative affect, but that the slopes vary across levels of regulation. This pattern is important for teaching improvement because students with similar participation levels may nevertheless differ in emotional response and self-regulatory capacity. Within the proposed framework, regulatory information can therefore be treated as contextual evidence when interpreting emotion-recognition outputs. Rather than generating identical interventions for all learners, the system can support differentiated recommendations while leaving final pedagogical decisions to the instructor. This interpretation is also consistent with the broader emphasis in adaptive learning on tailoring support to learner characteristics and performance [16].

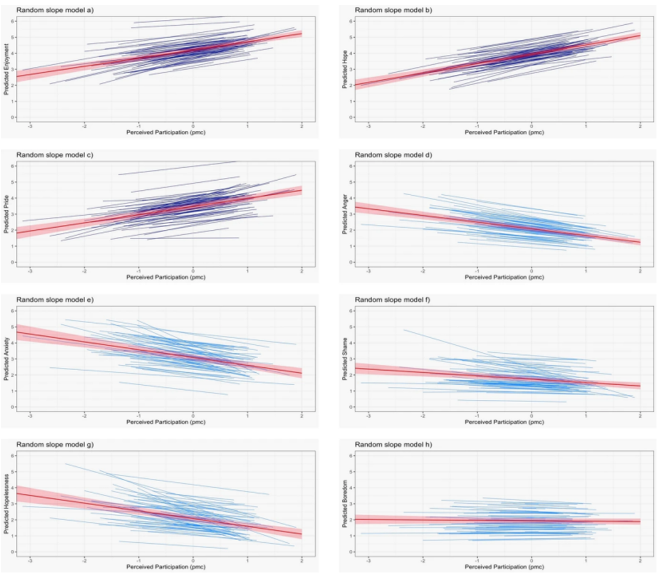

Figure 6: Random-Slope Analysis of the Relationship Between Classroom Participation and Student Emotions

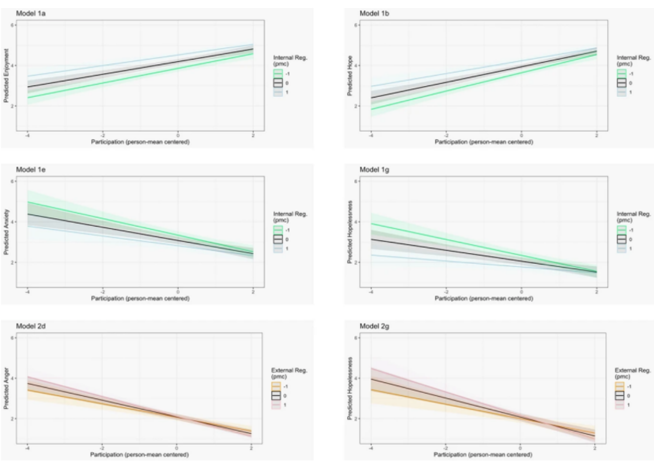

Figure 7: Moderating Effects of Internal and External Regulation on Classroom Participation and Student Emotions

V. Conclusion

This study presents an AI-assisted teaching-improvement framework for higher vocational midwifery education that combines contextual language modeling, sequence labeling, Transformer-based global context modeling, and classroom interaction analysis. The proposed architecture is designed to address a practical limitation of conventional educational evaluation: many important classroom states, including engagement, emotional fluctuation, and interaction quality, evolve continuously and may not be adequately represented by end-of-course examinations or occasional manual observation. By integrating ELMo-based contextual representation with BiLSTM-CRF and a Transformer encoder, the framework combines local sequential information, structured label dependencies, and long-range contextual relationships. The reported classroom emotion-recognition experiment yielded an F1 score of 90.0% for the combined model, compared with 85.1% for the Transformer encoder and 83.1% for BiLSTM+CRF. The reported group comparison also showed higher classroom engagement and a larger Emotion Improvement Index in the AI-enhanced group. Supplementary visual analyses further illustrate the importance of fairness auditing, teacher characteristics, learner engagement, emotion, and regulatory mechanisms when interpreting AI-supported educational data.

The results should nevertheless be interpreted within the limitations of the supplied study record. In particular, the manuscript contains two distinct empirical strands: classroom-oriented educational experiments and a clinically oriented dataset involving abortion patients. Because the relationship between these strands is not fully documented in the source manuscript, the underlying study records should be checked before final publication. In addition, some supplementary figures require fuller methodological documentation, including sample definitions, statistical-model specifications, and data provenance. These limitations do not diminish the usefulness of the computational framework, but they are important for reproducibility and for preventing overinterpretation of descriptive results. Future work should therefore validate the framework on clearly documented midwifery-student cohorts, report subgroup and fairness metrics, define the multimodal data-processing pipeline in greater detail, and evaluate whether AI-generated feedback leads to sustained improvements in learning and professional competence. Under these conditions, AI-assisted classroom analytics may provide a valuable decision-support layer for evidence-informed, personalized, and ethically responsible vocational midwifery education.

Author Contributions

Both authors contributed to the conception and design of the study, methodological development, data analysis, interpretation of the results, and preparation of the manuscript. Both authors reviewed and approved the final version of the manuscript for publication.

Conflicts of Interest

The authors declare that there are no conflicts of interest related to this work.

Funding

This research was supported by the Vocational Education Teaching Reform Project in Shaanxi Province (Project No. 2025SZX769) and the Higher Vocational Education Teaching Reform Project in Shaanxi Province (Project No. 23GZ026).

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to applicable ethical and institutional requirements.

Ethics Approval

The manuscript reports that the study protocol was reviewed and approved by the relevant hospital medical ethics committee.

Informed Consent

The manuscript reports that informed consent was obtained from all participants prior to their participation in the study.

Declaration of Generative AI and AI-Assisted Technologies in the Preparation of the Manuscript

Generative artificial intelligence tools were used solely to assist with language editing, revision, and improvement of the presentation of the manuscript. The authors reviewed and verified all AI-assisted content and take full responsibility for the accuracy, integrity, and final content of the manuscript.

References

[1] G. Georgieva-Tsaneva, I. Serbezova, and S. Beloeva, “Application of virtual reality, artificial intelligence, and other innovative technologies in healthcare education (nursing and midwifery specialties): Challenges and strategies,” *Education Sciences*, vol. 15, no. 1, Art. no. 11, 2025, doi: 10.3390/educsci15010011.

[2] A. Kranz and H. Abele, “The impact of artificial intelligence (AI) on midwifery education: A scoping review,” *Healthcare*, vol. 12, no. 11, Art. no. 1082, 2024, doi: 10.3390/healthcare12111082.

[3] M. M. K. Chan, A. W. H. Wan, D. S. K. Cheung, E. P. H. Choi, E. A. Chan, J. Yorke, and L. Wang, “Integration of artificial intelligence in nursing simulation education: A scoping review,” *Nurse Educator*, vol. 50, no. 4, pp. 195–200, 2025, doi: 10.1097/NNE.0000000000001851.

[4] L. Zhao, X. Dai, and S. Chen, “Effect of the case-based learning method combined with virtual reality simulation technology on midwifery laboratory courses: A quasi-experimental study,” *International Journal of Nursing Sciences*, vol. 11, no. 1, pp. 76–82, 2024, doi: 10.1016/j.ijnss.2023.12.009.

[5] G. Glauberman, A. Ito-Fujita, S. Katz, and J. Callahan, “Artificial intelligence in nursing education: Opportunities and challenges,” *Hawai’i Journal of Health & Social Welfare*, vol. 82, no. 12, pp. 302–305, 2023.

[6] O. O. Akinyemi, G. B. Ogbeye, O. E. Adegbilero-Iwari, and A. S. Afolalu, “AI-driven assessment of educational interventions: Enhancing midwives’ competence in using condom-catheter intrauterine balloon tamponade—Review,” *ABUAD Journal of Engineering Research and Development*, vol. 8, no. 1, pp. 151–160, 2025.

[7] M. K. K. Rony et al., “Nursing educators’ perspectives on the integration of artificial intelligence into academic settings,” *SAGE Open Nursing*, vol. 11, Art. no. 23779608251342931, 2025, doi: 10.1177/23779608251342931.

[8] S. Shishehgar, P. Murray-Parahi, E. Alsharaydeh, S. Mills, and X. Liu, “Artificial intelligence in health education and practice: A systematic review of health students’ and academics’ knowledge, perceptions and experiences,” *International Nursing Review*, vol. 72, no. 2, Art. no. e70045, 2025, doi: 10.1111/inr.70045.

[9] P.-J. Chen and W.-K. Liou, “ChatGPT-driven interactive virtual reality communication simulation in obstetric nursing: A mixed-methods study,” *Nurse Education in Practice*, vol. 85, Art. no. 104383, 2025, doi: 10.1016/j.nepr.2025.104383.

[10] A. Baser and H. Şahin, “Utilizing artificial intelligence for adaptive scenario development in interprofessional education,” *Journal of Medical Education and Family Medicine*, vol. 2, no. 1, pp. 37–49, 2025, doi: 10.62425/jmefm.1663445.

[11] K. Zheng, Z. Shen, Z. Chen, C. Che, and H. Zhu, “Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education,” *BMC Medical Education*, vol. 24, Art. no. 1003, 2024, doi: 10.1186/s12909-024-05977-z.

[12] K. Mikkonen, S. Y. Liaw, L. Spirgienė, A. Subočius, P. Ignatavičius, T. Blažauskas, and O. Riklikienė, “Multidimensional pedagogical framework for interprofessional education: Blending classroom, high fidelity and extended reality simulation,” *Nurse Education Today*, vol. 154, Art. no. 106838, 2025, doi: 10.1016/j.nedt.2025.106838.

[13] F. Behmanesh, Z. Gholamnia-Shirvani, S. Ghaffarifar, H.-A. Nikbakht, and S. Nazmi, “Designing, implementing, and evaluating a gamification approach in childbirth training for midwifery students using computer and mobile application,” *BMC Medical Education*, vol. 25, Art. no. 1050, 2025, doi: 10.1186/s12909-025-07370-w.

[14] M. Gerdes, A. Bayne, K. Henry, B. Ludwig, L. Stephenson, A. Vance, J. Wessol, and S. Winston, “Emerging artificial intelligence-based pedagogies in didactic nursing education: A scoping review,” *Nurse Educator*, vol. 50, no. 1, pp. E7–E12, 2025, doi: 10.1097/NNE.0000000000001746.

[15] C.-Y. Chang, H.-J. Jen, and J. C. Yang, “Integrating scenario game-based learning with the experiential learning strategy to facilitate nursing students’ learning performance and core competencies in labor support training,” *Interactive Learning Environments*, vol. 32, no. 10, pp. 7170–7185, 2024, doi: 10.1080/10494820.2024.2308092.

[16] A. Naciri, L. Benyahya, M. Radid, and G. Chemsi, “Personalized adaptive learning among midwifery and nursing students: A scoping review,” in *Personalized Learning Through Adaptive Systems and Intelligent Tutoring*, A. Benabid, I. El Imadi, and G. Chemsi, Eds. Hershey, PA, USA: IGI Global Scientific Publishing, 2025, pp. 205–218, doi: 10.4018/979-8-3373-0740-4.ch008.

[17] S. Jones and R. C. Pratchett, “A higher education intervention to promote artificial intelligence literacy in the healthcare workforce,” in *Artificial Intelligence Applications in Higher Education: Theories, Ethics, and Case Studies for Universities*, H. Crompton and D. Burke, Eds. New York, NY, USA: Routledge, 2025, pp. 179–194.

[18] S. Ningsih, Y. Syufi, and A. Lakehu, “Midwifery students’ perceptions of utilizing artificial intelligence (AI) for English language learning (ELL),” *Elsya: Journal of English Language Studies*, vol. 7, no. 1, pp. 90–100, 2025, doi: 10.31849/elsya.v7i1.22565.

[19] V. Echeverria et al., “TeamVision: An AI-powered learning analytics system for supporting reflection in team-based healthcare simulation,” in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2025, Art. no. 309, pp. 1–22, doi: 10.1145/3706598.3713395.

[20] T. Horta Reis Da Silva and A. Mitchell, “Integrating digital transformation in nursing education: Best practices and challenges in curriculum development,” in *Digital Transformation in Higher Education, Part B: Cases, Examples and Good Practices*, M. D. Lytras, A. C. Serban, A. Alkhaldi, S. Malik, and T. Aldosemani, Eds. Bingley, U.K.: Emerald Publishing Limited, 2024, pp. 57–101, doi: 10.1108/978-1-83608-424-220241004.

[21] S. T. Masoumian Hosseini et al., “Are we ready to integrate modern technologies into the medical curriculum for students? A systematic review,” *Discover Education*, vol. 4, Art. no. 114, 2025, doi: 10.1007/s44217-025-00521-7.

[22] M. Yildirim and M. Sahin, “Evaluating undergraduate nursing education and student competencies: A mixed-methods study using the input-process-output framework,” *BMC Nursing*, vol. 24, Art. no. 760, 2025, doi: 10.1186/s12912-025-03358-5.

[23] M. E. Peters et al., “Deep contextualized word representations,” in *Proceedings of NAACL-HLT*, 2018, pp. 2227–2237, doi: 10.18653/v1/N18-1202.

[24] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” in *Proceedings of NAACL-HLT*, 2016, pp. 260–270, doi: 10.18653/v1/N16-1030.

[25] X. Ma and E. Hovy, “End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1064–1074, doi: 10.18653/v1/P16-1101.

[26] Y. Zhang and J. Yang, "Chinese NER using lattice LSTM," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 1554–1564, doi: 10.18653/v1/P18-1144.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[28] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[29] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[30] H. Zhang, A. X. Lu, M. Abdalla, M. B. A. McDermott, and M. Ghassemi, "Hurtful words: Quantifying biases in clinical contextual word embeddings," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 110–120, doi: 10.1145/3368555.3384448.