Automatic Identification of Collaborative Problem-Solving Phases from Oral Peer Dialogue in Classroom

Wenting Sun[1] and Jiangyue Liu[2]
[1]Humboldt-Universität zu Berlin, Germany
[2]Soochow University, Jiangsu, China

suwentin@hu-berlin.de
ljy@suda.edu.cn

Abstract: Collaborative problem solving (CPS) is a critical competency in the Artificial intelligence (AI) era, requiring the integration of cognitive and social skills through real-time dialogue and coordination. While prior studies have explored CPS behaviours using human-coded text from online platforms, limited research has examined how machine learning (ML) and deep learning (DL) models perform on spoken peer dialogue in face-to-face (F2F) classroom settings. This study investigates the automatic classification of CPS phases using a validated coding framework applied to two classroom tasks—one supported by a GenAI assistant and one not. A total of 7,744 utterances were manually labelled across nine CPS subskills and three broader facets. Six ML and five DL models were evaluated, including lightweight BERT variants combined with various classifiers. Results show that BERT-based models significantly outperform traditional ML approaches. Specifically, BERT+ANN achieved better overall performance in smaller, imbalanced datasets, while BERT+CNN performed better in larger datasets. Reducing label granularity from nine subskills to three facets consistently improved classification accuracy and F1 scores. Both models achieved AUROC scores around 0.90, indicating strong discriminative capability. Several key insights emerged from the findings: Model architecture matters: Simpler classifiers like ANN preserve BERT's semantic representations and offer stable performance, especially in smaller or imbalanced datasets. Task context influences CPS behaviour: Different tasks elicit distinct CPS skill distributions, with task regulation dominating in technical tasks and communicative participation more prevalent in reflective tasks. Label granularity affects performance: Reducing the number of classification labels (e.g., from 9 subskills to 3 facets) significantly improves model accuracy and generalizability. Lightweight models are viable: Even with a reduced-capacity BERT model, competitive performance was achieved, suggesting potential for real-time, resource-efficient deployment in educational settings. This study contributes to educational AI by introducing a novel oral CPS dataset, benchmarking multiple models, and demonstrating the feasibility of lightweight architectures for real-time deployment. Limitations include the small sample size and single-modality input. Future work should explore multimodal features, larger and more diverse classrooms, and teacher-facing dashboards for actionable feedback. The findings support the development of scalable, ethical, and human-centered learning analytics tools that enhance collaborative learning in AI-enhanced education.

Keywords: Spoken dialogue, Face-to-Face classroom, Machine learning (ML), Deep learning (DL), Multi-label classification

1. Introduction

Artificial Intelligence (AI) has transformed many aspects of modern society, from healthcare and finance to education, reshaping how humans learn and collaborate (Markauskaite et al., 2022). However, while AI systems excel in computation and automation, uniquely human abilities—such as collaboration, communication, and joint problem solving—remain essential. Collaborative Problem Solving (CPS) is widely recognized as a core 21st-century competency, integrating cognitive reasoning with social interaction (Andrews-Todd et al., 2023; Flor & Andrews-Todd, 2022). Understanding and supporting CPS has therefore become a central focus of research in educational psychology and AI-based learning analytics (Taylor et al., 2024).

Collaborative problem solving (CPS) is a cognitively and socially demanding process in which individuals coordinate ideas, actions, and dialogue to explore problems and co-construct solutions. As emphasized by Markauskaite et al. (2022), CPS is a critical 21st-century competency, particularly in the era of Artificial intelligence (AI), where human collaboration skills are increasingly relevant for both human-human and human-agent interactions.

While prior research has made significant progress in identifying CPS behaviours through human-coded, text-based data from online learning platforms (e.g., Andrews-Todd et al., 2023; Flor & Andrews-Todd, 2022), much less is known about how Machine learning (ML) and Deep learning (DL) models perform when applied to spoken peer dialogue in face-to-face (F2F) classroom settings. While prior research has focused on text-based interactions from online platforms, spoken peer dialogue in face-to-face classrooms offers richer contextual information and authentic collaborative dynamics. Unlike scripted or simulated environments, real-world classroom dialogue reflects spontaneous coordination, negotiation, and decision-making, which are critical for understanding CPS behaviours.

Moreover, most existing studies rely on manual annotation and focus on online or simulated environments. Few have systematically compared the performance of ML and DL models on authentic, oral CPS data, especially across different task types and AI-supported learning contexts. As CPS behaviours may vary significantly depending on task structure and modality (Andrews-Todd et al., 2023), there is a pressing need to explore how automated models perform in diverse, real-world classroom settings.

To address these gaps, this study investigates the automatic classification of CPS phases from spoken peer dialogue using a range of ML and DL models. Specifically, it:

Applies a validated CPS coding framework to oral data collected from two classroom tasks—one supported by a GenAI assistant and one not.

Compares the performance of multiple ML and DL models, including lightweight BERT variants, across different tasks and label granularities.

Evaluates the feasibility of using automated models for CPS detection in real-world educational settings.

By bridging the gap between human-centered pedagogy and automated learning analytics, this study contributes to the development of scalable and context-aware AI tools for classroom use.

For this purpose, two research questions (RQ) were proposed: RQ1: How effectively can traditional Machine Learning (ML) and Deep Learning (DL) models classify CPS phases from spoken peer dialogue in classroom settings? RQ2: How does model performance vary across different task contexts (AI-supported vs. non-AI-supported) and label granularities? In the following, section 2 reviews related work on CPS frameworks and AI-based behaviour detection. Section 3 details the dataset, annotation process, and model configurations. Section 4 presents experimental results. Section 5 discusses implications for CPS analytics in both AI and real-world educational settings, and Section 6 concludes with limitations and future directions.

2. Related Work and Research Questions

The literature reviewed in this section was selected based on relevance to three key dimensions: (1) empirical or computational studies on CPS; (2) application of ML/DL models to dialogue or behaviour classification; and (3) publication recency (2018–2025) to ensure inclusion of recent advances. Both supporting and contrasting perspectives were considered to provide a balanced understanding.

2.1 Collaborative Problem Solving (CPS) Skills and Frameworks

Collaborative problem solving (CPS) involves the coordination of ideas, actions, and dialogue among group members in real time. According to Andrews-Todd et al. (2023), CPS can be categorized into three core facets: Communicative participation (e.g., maintaining communication, sharing information), Social regulation (e.g., monitoring, negotiating), and Task regulation (e.g., planning, executing). This framework together with other CPS analysis frameworks like indicator-based mapping by Sun et al. (2022) have been validated in digital and online collaborative environments, but their application to oral, face-to-face classroom settings remains underexplored.

2.2 AI-Based CPS Behaviour Detection

Recent advances in AI have enabled the automated detection of CPS behaviours using both traditional ML and DL models. Traditional ML models such as logistic regression (LR), random forest (RF), and support vector machines (SVM) have been widely used for text classification tasks. However, DL models—particularly those based on pretrained language models like BERT—have shown superior performance in capturing contextual semantics (Stewart et al., 2023; Wong et al., 2025).

While most existing studies focus on text-based interactions in online platforms, emerging research emphasizes the importance of analysing spoken dialogue transcripts in classroom settings. Unlike scripted or simulated environments, real-world classroom dialogue reflects spontaneous coordination, negotiation, and decision-making, which are critical for identifying CPS behaviours (Taylor et al., 2024; Wong et al., 2025). Moreover, oral peer dialogue in face-to-face classrooms offers authentic insights into collaborative dynamics that are often absent in digital platforms (Flor & Andrews-Todd, 2022).

2.3 Research Gap and Contribution

Despite growing interest in CPS analytics, few studies have applied fine-grained CPS coding schemes to spoken peer dialogue in real-world classrooms. Furthermore, comparative evaluations of multiple ML and DL models on the same dataset remain limited. This study addresses these gaps by:

- Applying a validated CPS framework to human-labelled oral transcripts from two distinct classroom tasks.
- Comparing the performance of six ML and five DL models, including lightweight BERT variants.
- Investigating how model performance varies across tasks and label granularities.

3. Methodology

3.1 Data Collection and Context

Data were collected from 38 student groups across two undergraduate courses in China:

- Lesson Plan Assessment Task (24 groups, 49 participants): Conducted in an Educational Technology course, where students revised lesson plans using a GenAI assistant (Wenxinyiyan, based on ERNIE).
- Internet Control Message Protocol (ICMP) Networking Task (14 groups, 28 participants): Conducted in an Engineering course, where students collaborated on IP packet transmission using ICMP (a network layer protocol used for error reporting and diagnostics in networking).

Both tasks were conducted in face-to-face classroom settings, with audio recordings capturing peer dialogue. The lesson plan task involved human-AI collaboration, while the ICMP task did not involve AI assistance. These two datasets were also combined to evaluate model robustness across contexts. To explore how model performance across tasks, three datasets were organized: one from the Lesson plan assessment task, one from the ICMP task, one combining the prior two datasets.

3.2 Data Annotation

All audio recordings were manually transcribed and segmented into meaning units, where each segment represents a coherent CPS behaviour. Transcripts were labelled using the CPS coding scheme proposed by Andrews-Todd et al. (2023), which includes 3 CPS facets and 9 CPS subskills: Communicative Participation (Maintaining communication, Sharing information, Establishing shared understanding), Social Regulation (Negotiating, Monitoring), Task Regulation and Activity (Exploring and understanding, Representing and formulating, Planning, Executing). This coding approach captures both cognitive and regulatory dimensions of CPS. Example annotation:

- *Student A*: “In the summary, emphasize the key challenges…” → Label: Planning
- *Student B*: “The AI revised it for me.” → Label: Sharing Information

3.3 Model Selection and Configuration

We evaluated six Machine Learning (ML) models—Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Random Forest (RF)—and five Deep Learning (DL) models—Bidirectional Encoder Representations from Transformers (BERT), BERT combined with Artificial Neural Network (BERT+ANN), BERT with Convolutional Neural Network (BERT+CNN), BERT with Long Short-Term Memory (BERT+LSTM), and BERT with Bidirectional LSTM (BERT+BiLSTM).

ML refers to algorithms that learn patterns from data through statistical inference and manual feature engineering—that is, selecting and designing input variables that help the model make accurate predictions. DL, by contrast, is a subset of ML that relies on multi-layer neural networks to automatically learn hierarchical representations from raw data, often requiring less manual preprocessing.

Among the DL models, BERT (Bidirectional Encoder Representations from Transformers) uses a transformer architecture to capture contextual meaning from text by considering both left and right word sequences. For this study, we adopted a lightweight BERT variant—`uer/chinese_roberta_L-2_H-128`—which includes only two transformer layers and 128 hidden units. This configuration significantly reduces computational cost while maintaining sufficient representational power for our classification tasks.

Considering that a detailed review of algorithmic mechanisms is beyond the scope of this paper, readers may refer to two comprehensive reviews on ML and DL in educational contexts: Deep Learning Techniques in

Educational Data Mining (Lin et al., 2025) and Unlocking the Power of Machine Learning in E-Learning (Salem & Shaalan, 2025).

All models were trained using stratified 10-fold cross-validation. For BERT-based models: Tokenization: Max sequence length = 300, Hyperparameters: Learning rate = 5e-5, batch size = 16, epochs = 10. These were selected based on prior best practices and confirmed through grid search.

3.4 Evaluation Metrics

To assess model performance, as mentioned in the review by Naidu et al. (2023), we adopted multiple evaluation metrics, including: Macro-F1 Score, Accuracy, Precision, Recall, Area under the receiver operating characteristic curve scores (AUROC).

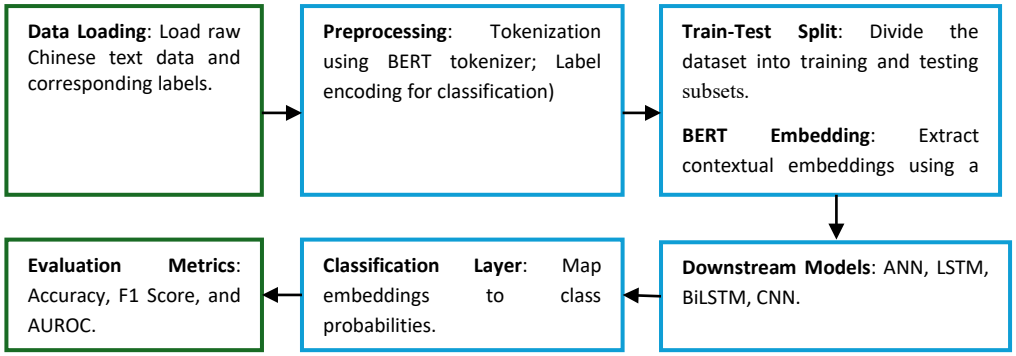

Figure 1: Data analysis workflow (Note: ML model approaches did not have BERT embedding step)

All code solutions and one examples of automatic classification results can be found on the OSF link:

https://osf.io/xqe4m/?view_only=5862e658a2804e4e900041b356647aba

4. Results

4.1 Descriptive Analysis of CPS Skill Distribution

Table 1 presents the distribution of CPS skills across the two classroom tasks: the lesson plan assessment (with GenAI assistance) and the ICMP packet operation (without GenAI). Students demonstrated a wide range of CPS skills in both tasks, covering all nine subskills defined in the coding framework by Andrews-Todd et al. (2023).

Notably, the most frequently observed CPS skills differed between the two tasks, diverging from patterns reported in prior studies conducted in online environments (Andrews-Todd et al., 2023; Flor & Andrews-Todd, 2022). In the ICMP task, which involved hands-on technical collaboration, the most frequent skills were Executing (30.46%) and Planning (16.65%), reflecting a strong emphasis on task regulation. In contrast, the lesson plan assessment task showed higher frequencies of Sharing Information (19.33%) and Maintaining Communication (17.87%), aligning more closely with communicative participation.

This variation suggests that task type and modality (face-to-face vs. online) significantly influence the manifestation of CPS behaviours. The combined dataset showed a relatively balanced distribution across the three CPS facets, with Task Regulation accounting for the largest proportion (53.38%).

Table 1: Summary of reported aggregate CPS skills by facet and task

CPS skills	Lesson plan task	ICMP task	Combined
Communicative participation	**1831(52.52%)**	**921(21.63%)**	**2752(35.54%)**
Maintaining communication	623(17.87%)	192(4.51%)	815(10.52%)
Sharing information	647(19.33%)	547(12.85%)	1194(15.42%)
Establishing shared understanding	561(16.09%)	182(4.27%)	743(9.59%)
Social regulation	**563(16.15%)**	**295(6.93%)**	**858(11.08%)**
Negotiating	255(7.31%)	48(1.13%)	303(3.91%)
Monitoring	308(8.83%)	247(5.8%)	555(7.17%)
Task regulation and activity	**1092(31.33%)**	**3042(71.44%)**	**4134(53.38%)**

CPS skills	Lesson plan task	ICMP task	Combined
Exploring and understanding	189(5.42%)	374(8.78%)	563(7.27%)
Representing and formulating	213(6.11%)	662(15.55%)	875(11.30)
Planning	198(5.68%)	709(16.65%)	907(11.7%)
Executing	492(14.11%)	1297(30.46%)	1789(23.1%)
Sum	3486	4258	7744

4.2 Model Performance Analysis

4.2.1 Performance on 9-Category CPS classification

Tables 2 and 3 summarize the performance of all models on the lesson plan assessment task and ICMP packet operation task datasets, respectively. Across both tasks, BERT+ANN slightly outperformed other models in terms of Accuracy and AUROC, indicating better overall performance and discriminative ability. BERT+CNN also performed competitively, particularly in the ICMP dataset, where it achieved higher recall and precision.

As illustrated in Figure 2 and Figure 3, the confusion matrices visualize the classification patterns of the best-performing models for each dataset. Darker cells indicate higher prediction accuracy. The numbers 0–8 in both figures correspond to the CPS subskill labels: Maintaining Communication (0), Sharing Information (1), Establishing Shared Understanding (2), Negotiating (3), Exploring and Understanding (4), Representing and Formulating (5), Planning (6), Executing (7), and Monitoring (8).

Figure 2 shows the confusion matrix for the BERT+ANN model on the lesson plan assessment dataset. The strong diagonal dominance indicates generally accurate classification across CPS subskills, with the highest accuracy for Maintaining Communication (0), followed by Sharing Information (1) and Executing (7).

Figure 3 presents the confusion matrix for the BERT+CNN model on the ICMP packet operation dataset. Here, the Executing (7) label is predicted with particularly high accuracy, reflecting the task's procedural and technical nature.

Together, these visualizations demonstrate that the models capture distinct task-specific CPS behaviour patterns, with clearer identification of task-regulation phases in the ICMP task and communicative facets in the lesson plan task.

Among traditional ML models, logistic regression with TF-IDF achieved the best performance, outperforming more complex models like random forest and SVM. This suggests that while DL models are superior overall, well-tuned ML baselines remain viable for certain tasks.

Table 2: Model Performance Comparison by Evaluation Metric on lesson plan assessment dataset

Models	Accuracy	Precision	Recall	F1	AUROC
BERT	0.4957	0.3351	0.36	0.3212	0.8544
BERT+LSTM	0.5244	**0.5815**	0.4226	0.41	0.8714
BERT+ANN	**0.6074**	0.5742	**0.5492**	0.5497	**0.9039**
BERT+BiLSTM	0.5057	0.5656	0.3809	0.365	0.8549
BERT+CNN	0.5845	0.5779	0.5442	**0.5511**	0.9027
Logistic Regression +IF-IDF	0.5431	0.5217	0.4915	0.502	0.8601
Naïve Bayes+IF-IDF	0.4384	0.3001	0.2869	0.2344	0.8111
Random Forest+IF-IDF	0.5043	0.4971	0.3833	0.3838	0.8321
K-Nearest Neighbors+IF-IDF	0.4527	0.3894	0.3663	0.3575	0.7344
SVM+IF-IDF	0.5345	0.5241	0.4861	0.4992	0.8506
Decision Tree+IF-IDF	0.342	0.2764	0.279	0.2766	0.5979

Table 3: Model Performance Comparison by Evaluation Metric on ICMP packet operation dataset

Models	Accuracy	Precision	Recall	F1	AUROC
BERT	0.5012	0.4159	0.3427	0.3558	0.7926
BERT+LSTM	0.5365	0.4197	0.4075	0.408	0.8178
BERT+ANN	**0.5751**	0.4865	0.4374	**0.4496**	0.853
BERT+BiLSTM	0.4965	0.4305	0.3522	0.3599	0.8082
BERT+CNN	0.5435	**0.5082**	**0.4387**	0.4421	**0.8593**
Logistic Regression +IF-IDF	0.4396	0.4162	0.2726	0.2881	0.7902
Naïve Bayes+IF-IDF	0.395	0.4237	0.1909	0.1759	0.7364
Random Forest+IF-IDF	0.4387	0.3738	0.2985	0.311	0.7636
K-Nearest Neighbors+IF-IDF	0.2485	0.3017	0.199	0.194	0.5844
SVM+IF-IDF	0.4401	0.3985	0.2896	0.3088	0.7823
Decision Tree+IF-IDF	0.3624	0.293	0.2664	0.2719	0.6003

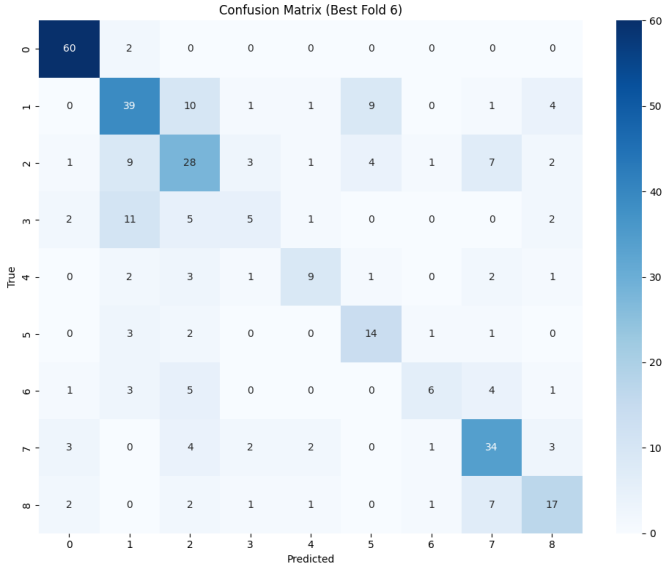

Figure 2: Confusion matrix of the BERT+ANN in on lesson plan assessment dataset

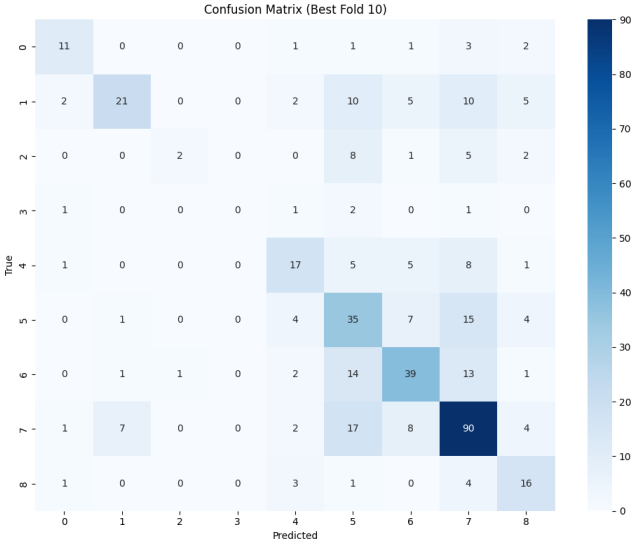

Figure 3: Confusion matrix of the BERT+CNN in on ICMP packet operation dataset

4.2.2 Performance on combined dataset

Table 4 shows model performance on the combined dataset. While performance slightly declined compared to task-specific datasets, BERT+CNN and BERT+ANN remained the top-performing models. This supports the finding that task-specific training yields better results, likely due to reduced contextual variability.

Table 4: Model Performance Comparison by Evaluation Metric Combined Dataset (9 CPS Skills)

Models	Accuracy	Precision	Recall	F1	AUROC
BERT	0.4806	0.4929	0.4109	0.4125	0.8265
BERT+LSTM	0.5123	0.5104	0.4595	0.4631	0.8563
BERT+ANN	0.5277	0.504	0.4958	0.4925	0.8678
BERT+BiLSTM	0.5078	**0.5415**	0.4322	0.4351	0.8446
BERT+CNN	**0.5368**	0.504	**0.507**	**0.5016**	**0.8743**
Logistic Regression +IF-IDF	0.4464	0.4483	0.3768	0.3737	0.8017
Naïve Bayes+IF-IDF	0.3611	0.4673	0.2393	0.2338	0.7771
Random Forest+IF-IDF	0.4038	0.3678	0.3424	0.3347	0.7771
K-Nearest Neighbours+IF-IDF	0.2484	0.3043	0.2358	0.2092	0.6183
SVM+IF-IDF	0.4327	0.4244	0.3723	0.3659	0.802
Decision Tree+IF-IDF	0.3226	0.281	0.2916	0.278	0.5901

4.2.3 Impact of label granularity: 3 vs. 9 categories

Table 5 compares the performance of BERT+ANN and BERT+CNN across both tasks using 3-category (facet-level) and 9-category (skill-level) classification. In all cases, reducing the number of labels led to substantial improvements in accuracy, precision and F1 score. For example, BERT+CNN achieved an F1 score of 0.6879 on the lesson plan dataset with 3 labels, compared to 0.5511 with 9 labels.

Table 5: Comparison of BERT+ANN and BERT+CNN on 3 vs. 9 CPS Categories

Models	dataset	CPS	Accuracy	Precision	Recall	F1	AUROC
BERT+ANN	Lesson plan task	9 CPS skills	0.6074	0.5742	0.5492	0.5497	**0.9039**
		3 CPS facets	0.7135	0.6807	**0.6849**	0.68	0.8567
	ICMP task	9 CPS skills	0.5751	0.4865	0.4374	0.4496	0.853
		3 CPS facets	0.7606	0.633	0.584	0.603	0.8008
	Combined	9 CPS skills	0.5277	0.504	0.4958	0.4925	0.8678
		3 CPS facets	0.7226	0.6524	0.6398	0.6442	0.8206
BERT+CNN	Lesson plan task	9 CPS skills	0.5845	0.5779	0.5442	0.5511	0.9027
		3 CPS facets	0.7307	0.7085	0.6742	**0.6879**	0.873
	ICMP task	9 CPS skills	0.5435	0.5082	0.4387	0.4421	0.8593
		3 CPS facets	**0.8028**	**0.7539**	0.5739	0.6257	0.8049
	Combined	9 CPS skills	0.5368	0.504	0.507	0.5016	0.8743
		3 CPS facets	0.729	0.6975	0.6219	0.6432	0.8184

5. Discussion and Implication

5.1 Summary of Findings

This study investigated the effectiveness of ML and DL models in classifying CPS phases from spoken peer dialogue in face-to-face classroom settings. It was found that DL models, particularly BERT-based architectures, outperformed traditional ML models across classification tasks. Among them, BERT+ANN achieved the most stable and accurate performance, especially in smaller and imbalanced datasets, while BERT+CNN showed advantages in larger datasets with more diverse samples. Models trained on task-specific datasets performed

better than those trained on combined datasets. Additionally, reducing the number of classification labels (e.g., from 9 subskills to 3 facets) improved model performance, suggesting that label granularity and data homogeneity are critical factors. These findings are consistent with prior research (Andrews-Todd et al., 2023; Wong et al., 2025; Xu et al., 2025) and extend them by applying CPS frameworks to spoken classroom dialogue, a modality that has received limited attention in automated CPS analytics.

5.2 Interpretation of Model Behaviour

The superior performance of BERT+ANN can be attributed to its architectural simplicity and semantic preservation. As noted in prior studies (Dablain et al., 2024), adding a simple fully connected layer to BERT outputs tends to yield more stable results than complex structures like CNN or LSTM, particularly when data is limited or noisy. ANN layers effectively retain BERT's contextual embeddings without introducing structural redundancy. In contrast, CNNs excel at extracting local linguistic patterns, which may enhance recall in specific CPS categories. However, this comes at the cost of disrupting BERT's global semantic structure, leading to slightly lower overall accuracy. This trade-off was evident in our results and aligns with findings from sentiment classification research (Chen et al., 2022), which highlight CNN's strengths in local feature extraction but limitations in modelling long-range dependencies.

Moreover, both BERT+ANN and BERT+CNN achieved AUROC scores around 0.90, indicating strong discriminative capabilities in multi-class classification. This reinforces the robustness of BERT-based models in capturing the nuanced cognitive and regulatory elements of CPS.

5.3 Model Efficiency and Practical Deployment

A notable aspect of this study is the use of a lightweight BERT variant (uer/chinese_roberta_L-2_H-128), which contains only 2 transformer layers and 128 hidden units. Despite its reduced capacity, the BERT+ANN model achieved performance comparable to larger models used in previous studies (Wong et al., 2025). This suggests that lightweight BERT models, when paired with appropriate classifiers, can offer a cost-effective and deployable solution for real-time CPS analytics in classroom settings. This finding is particularly relevant for educational environments with limited computational resources, where deploying full-scale language models may not be feasible.

5.4 Implications for Model Selection and Evaluation

The results offer several insights for researchers and practitioners designing automated CPS detection systems:

- **Model Architecture**: Simple classifiers (e.g., ANN) are preferable when stability and semantic fidelity are prioritized. CNNs may be used to boost recall in specific categories but should be applied cautiously to avoid semantic distortion.
- **Label Design**: Reducing the number of classification labels improves model performance by increasing sample size per class and reducing ambiguity. This highlights a trade-off between analytical granularity and classification accuracy.
- **Dataset Composition**: Task-specific datasets yield better performance than aggregated datasets, emphasizing the importance of contextual consistency in CPS modelling.
- **Evaluation Strategy**: Macro-F1 can be prioritized in imbalanced multi-label settings, alongside AUROC to assess ranking capabilities.

5.5 Theoretical and Practical Implications

From a theoretical perspective, this study reinforces the context-dependent nature of CPS behaviours, demonstrating that different tasks elicit distinct patterns of communication, regulation, and execution. The use of spoken dialogue as a data source provides a more ecologically valid basis for understanding CPS, capturing authentic scaffolding cycles and decision-making processes.

Practically, the findings have implications for teacher-facing analytics and intelligent tutoring systems (ITS). By identifying optimal moments for intervention—such as when students encounter conceptual thresholds or exhibit regulatory breakdowns—teachers can better support group learning dynamics. These insights can also inform the design of ITS that incorporate teacher-informed scaffolding strategies, enhancing their responsiveness and pedagogical alignment.

Comparative Perspective: CPS in AI-Mediated and Analogue Contexts. While the present study focuses on the automatic detection of CPS phases in face-to-face classroom dialogue, it is worth briefly reflecting on how collaborative problem solving manifests differently in AI-mediated and analogue (human-only) settings.

In AI-mediated contexts, collaboration is often enhanced by digital tools or intelligent agents that can scaffold communication, monitor progress, or provide adaptive feedback. Such environments offer several advantages, including scalability, traceability of learner interactions, and the potential for real-time analytics to support both learners and teachers (Stewart et al., 2023; Taylor et al., 2024). However, these benefits often come at the cost of reduced spontaneity and limited access to non-verbal cues—such as gesture, tone, or facial expression—which are crucial for authentic social coordination (Wong et al., 2025). This tension partly motivated our exploration of face-to-face group discussions in which students collaborated to provide feedback on AI-generated lesson plans (the lesson plan assessment task). Such a setting allows the integration of AI's advantages in providing rapid, high-quality suggestions while simultaneously fostering opportunities for students to negotiate meaning, share perspectives, and resolve conflicts through direct interpersonal interaction.

In contrast, analogue or face-to-face CPS provides richer social and emotional context, allowing for immediate negotiation of meaning and interpersonal regulation. These interactions tend to be more fluid and context-sensitive but are also more difficult to capture and analyse automatically (Flor & Andrews-Todd, 2022).

Understanding the complementary strengths of these two worlds can help researchers and educators design AI tools that augment—rather than replace—the natural dynamics of human collaboration. This comparative view also offers an accessible entry point for AI researchers who may be less familiar with the nuances of classroom-based CPS.

6. Conclusion and Further studies

This study demonstrates the feasibility and effectiveness of using machine learning (ML) and deep learning (DL) models to automatically classify collaborative problem solving (CPS) phases from spoken peer dialogue in face-to-face classroom settings. By applying a validated CPS framework to two distinct classroom tasks—one supported by GenAI and one not—this research provides empirical evidence that BERT-based models, particularly BERT+ANN and BERT+CNN, can achieve strong performance in multi-label CPS classification.

Contributions. This study makes several contributions to the field of educational AI and learning analytics: It pioneers the use of automated CPS classification in oral, face-to-face classroom dialogue, addressing a gap in current research. It evaluates a wide range of ML and DL models on the same annotated dataset, offering a comprehensive performance benchmark. It introduces a novel dataset collected from both GenAI-supported and traditional classroom tasks, enabling future comparative studies. It provides open-source code and model configurations to support replication and further development by the research community.

Limitations. Despite its contributions, the study has several limitations: the dataset size is relatively small, which may limit the generalizability of the findings; only textual transcripts were used; a lightweight BERT model was used, which may lack the representational power of full-scale models.

Future research could address these limitations by: Expanding to larger and more diverse classroom contexts, including different age groups, subjects, and cultural settings. Incorporating multimodal features (e.g., pitch, pauses, gestures) to enhance the detection of social and emotional cues in CPS. Exploring real-time deployment of CPS detection models in classrooms, including teacher-facing dashboards for actionable feedback. Comparing performance with full-capacity BERT models and other large language models (LLMs) to assess trade-offs between accuracy and efficiency.

Ethics Declaration: Ethical clearance was not required for this study.

AI Declaration: AI tools (specifically ChatGPT by OpenAI) were used during the writing process to assist with grammar refinement and language polishing. All research design, data analysis, and interpretation were conducted solely by the authors.

References

Andrews-Todd, J., Jiang, Y., Steinberg, J., Pugh, S. L. & D'Mello, S. K. (2023) 'Investigating collaborative problem solving skills and outcomes across computer-based tasks', Computers & Education, 207, 104928.

Chen, Z., Wang, Y., Zhu, Y. & Chen, S. (2022, January) 'Attention-based CNN and BiLSTM hybrid model for aspect-level sentiment classification', ICETIS 2022; 7th International Conference on Electronic Technology and Information Science, pp. 1–6. VDE.

Dablain, D., Jacobson, K. N., Bellinger, C., Roberts, M. & Chawla, N. V. (2024) 'Understanding CNN fragility when learning with imbalanced data', Machine Learning, 113(7), pp. 4785–4810.

Flor, M. & Andrews-Todd, J. (2022) 'Towards automatic annotation of collaborative problem-solving skills in technology-enhanced environments', Journal of Computer Assisted Learning, 38(5), pp. 1434–1447.

Lin, Y., Chen, H., Xia, W., Lin, F., Wang, Z. and Liu, Y., (2025). 'A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining'. Data Science and Engineering, pp.1-27. https://doi.org/10.1007/s41019-025-00303-z

Markauskaite, L., Marrone, R., Poquet, O., Knight, S., Martinez-Maldonado, R., Howard, S. et al. (2022) 'Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI?', Computers and Education: Artificial Intelligence, 3, 100056.

Naidu, G., Zuva, T. & Sibanda, E. M. (2023, April) 'A review of evaluation metrics in machine learning algorithms', Computer Science On-line Conference, pp. 15–25. Cham: Springer International Publishing.

Salem, M. and Shaalan, K., (2025). 'Unlocking the power of machine learning in E-learning: A comprehensive review of predictive models for student performance and engagement'. Education and Information Technologies, pp.1-24. https://doi.org/10.1007/s10639-025-13526-4

Stewart, A. E. B. et al. (2023) 'CPSCoach: The Design and Implementation of Intelligent Collaborative Problem Solving Feedback', in Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O. C. & Dimitrova, V. (eds.) Artificial Intelligence in Education. AIED 2023, Lecture Notes in Computer Science, vol. 13916. Cham: Springer. https://doi.org/10.1007/978-3-031-36272-9_58

Sun, C., Shute, V.J., Stewart, A.E., Beck-White, Q., Reinhardt, C.R., Zhou, G., Duran, N. and D'Mello, S.K., (2022) 'The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment'. Computers in Human Behavior, 128, p.107120. https://doi.org/10.1016/j.chb.2021.107120

Taylor, M., Barthakur, A., Azad, A., Joksimovic, S., Zhang, X. & Siemens, G. (2024) 'Quantifying collaborative complex problem solving in classrooms using learning analytics', ACM Learning Analytics & Knowledge Conference (LAK'24), pp. 551–562. New York, NY: ACM. https://doi.org/10.1145/3636555.3636913

Wong, K., Wu, B., Bulathwela, S. & Cukurova, M. (2025) 'Rethinking the Potential of Multimodality in Collaborative Problem Solving Diagnosis with Large Language Models', arXiv preprint, arXiv:2504.15093.

Xu, J., Liu, C., Tan, X. et al. (2025) 'General information metrics for improving AI model training efficiency', Artificial Intelligence Review, 58, p. 289. https://doi.org/10.1007/s10462-025-11281-z

Appendix 1

Table: List of Acronyms and Definitions

Acronym	Full Form	Definition
AI	Artificial Intelligence	Computational systems capable of performing tasks that typically require human intelligence, such as perception, reasoning, and learning
ML	Machine Learning	A subset of AI focusing on algorithms that learn patterns and relationships from data through training and evaluation
DL	Deep Learning	A subfield of ML that employs multi-layer artificial neural networks to automatically extract hierarchical features from large datasets
BERT	Bidirectional Encoder Representations from Transformers	A transformer-based deep learning model that captures bidirectional contextual relationships between words in a sentence
CNN	Convolutional Neural Network	A deep learning architecture designed to capture local spatial or sequential patterns through convolution operations, widely used for text and image classification
LSTM	Long Short-Term Memory	A type of recurrent neural network capable of learning long-range dependencies in sequential data through gated cell mechanisms
SVM	Support Vector Machine	A supervised ML algorithm that constructs optimal hyperplanes to separate data into classes with maximum margin
RF	Random Forest	An ensemble ML method that aggregates multiple decision trees to improve classification and reduce overfitting
ICMP	Internet Control Message Protocol	A network-layer protocol used for error reporting and diagnostics in computer networks, commonly utilized in network engineering tasks