Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enhancing Cognitive Diagnosis with Group-Aware Graph Augmentation and Self-Supervised Alignment

Xiaomeng Zhang
Bozhou University
Ruxiang Liu
liuruxiang@bzuu.edu.cn
Bozhou University
Piao Shi
Bozhou University
Qun Ren
Bozhou University

Research Article

Keywords: Cognitive Diagnosis, Self-Supervised Learning, Graph Neural Networks, Data Sparsity

Posted Date: June 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10007904/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Enhancing Cognitive Diagnosis with Group-Aware Graph Augmentation and Self-Supervised Alignment

Xiaomeng Zhang[1], Ruxiang Liu[1*], Piao Shi[1], and Qun Ren[1]

[1]Department of Electronics and Information Engineering, Bozhou University, No. 2266, Tangwang Avenue, High-tech Industrial Development Zone, Bozhou, 236800, Anhui, China.

*Corresponding author(s). E-mail(s): liuruxiang@bzuu.edu.cn;
Contributing authors: xiaomeng026@outlook.com;
shipiao@mail.hfut.edu.cn; renqun531@163.com;

Abstract

In the field of intelligent education, cognitive diagnosis aims to assess students' cognitive states by analyzing their historical response data, serving as a foundation for supporting downstream tasks such as personalized learning recommendations. However, in real learning scenarios, the distribution of student response data is often highly imbalanced. Students with fewer responses suffer from data sparsity, leading to inaccurate diagnoses, which poses a major challenge for existing models. To this end, we propose a Group-Aware Cognitive Diagnosis (GACD) framework. The core idea is to group students with similar answering patterns together. Within each group, we artificially creates additional "peer interactions" for students with sparse data, effectively giving the model more information to learn from. Notably, as correct and incorrect responses reflect distinct behavioral patterns, we first decompose the interaction graph into two separate subgraphs based on response types and learn representations from each independently. To ensure accurate modeling of students' cognitive states, we introduce a self-supervised strategy to align node representations between the original and augmented graphs. Finally, the learned representations are integrated into existing cognitive diagnosis models. Extensive experiments on three real-world datasets demonstrate that the GACD framework significantly improves prediction performance.

Keywords: Cognitive Diagnosis, Self-Supervised Learning, Graph Neural Networks, Data Sparsity

1 Introduction

The rapid integration of artificial intelligence into educational contexts is facilitating the emergence of intelligent education [1–3]. As a fundamental component of this field, cognitive diagnosis [4] aims to assess students' cognitive states based on their historical response data, thereby supporting downstream tasks such as exercise recommendation [5] and course recommendation [6].

Early cognitive diagnosis models were developed based on educational measurement theories, including Item Response Theory (IRT) [7] and DINA [8]. These approaches rely on manually designed interaction functions to model the relationship between students' cognitive states and item responses, yet they often fall short in capturing complex, non-linear interactions. With advances in deep learning, neural networks have increasingly replaced handcrafted interaction functions, leading to significant improvements in performance [9]. Subsequent research has focused on leveraging deep learning techniques to extract latent information from students' daily learning behaviors, enriching the representation of learner characteristics [10–13]. These efforts have contributed to the formation of versatile and scalable cognitive diagnostic frameworks [12].

However, in real learning scenarios, students' exercise selections are often subjective, leading to a biased distribution in response data [14]. Specifically, students tend to choose simpler items or exercise types with which they are already familiar and proficient [15]. Furthermore, online education platforms typically design their item pools to emphasize key and fundamental concepts, resulting in most students engaging with mainstream problems while only a minority attempt more obscure or challenging exercises. This subjectivity in exercise choice creates significant disparities in both the quantity and types of exercises and knowledge concepts that different students engage with. As shown in Table 1 ,the number of exercises completed by the students varies greatly across the three regular datasets. A detailed description of these datasets can be found in 4.1.1. In deep neural network-based cognitive diagnosis models, students with limited interaction history pose a significant challenge due to the data sparsity problem. Without sufficient response data [16], the model lacks the necessary evidence to reliably infer the student's knowledge state.

Table 1 Statistical summary of student exercise counts across the three datasets. The mean is rounded.

Degree	Assist2017	Neurips2020	Junyi
Maximum	805	150	513
Minimum	2	15	13
Mean	228	76	41
Median	204	75	28

Research has found that students present diverse group characteristics in the cognitive dimension of exercise measurement [17]. According to the cognitive styles, they

can be divided into field-independent learners and field-dependent learners [18]. Field-independent students are good at analytical exercises (such as mathematical proof and graphical reasoning), while field-dependent students are better at situational exercises (such as case analysis and open discussion). From the perspective of memory capacity difference, students with high memory capacity have significant advantages in multistep reasoning exercises (mathematical comprehensive exercises, logic chain exercises) [19]. From the perspective of knowledge representation, language learners get high scores in text description exercises, and visual learners will improve their accuracy in solving chart and schematic exercises [20]. In terms of metacognition strategy differences, students with high metacognition ability are better at open exercises (such as designing experiments and demonstrating ideas) because they can dynamically adjust problem solving strategies [21]. Students' performance may lead them to prefer exercises they excel at, which could bias response outcomes. The Latent Class Analysis (LCA) of students can help to further explore learners' heterogeneity, reveal hidden cognitive structure patterns, and provide effective evidence for low-frequency response data enhancement.

To address these issues, we propose a **G**roup-**A**ware **C**ognitive **D**iagnosis framework, named **GACD** to mitigate the impact of imbalanced response data distribution by enhancing the homogeneity of student latent groups. Specifically, we begin by clustering students based on their node representations to partition them into distinct subgroups. Within each student subgroup, we augment interactions between low-degree student nodes and other students, thereby enhancing both the intra-group connectivity and the emphasis on low-degree students. Notably, since correct and incorrect responses reflect different behavioral patterns, we first partition the graph into two separate subgraphs based on response types and perform representation learning on each independently. To ensure accurate modeling of students' cognitive states, we propose a self-supervised strategy to align node representations between the original and augmented graphs. Finally, we incorporated the learned representations into existing cognitive diagnostic models and effectively improved their performance. Extensive experiments on real datasets have demonstrated our proposed GACD framework achieves state-of-the-art performance on three real-world public datasets. Our main contributions are as follows:

- We focus on the impact of student latent class and proposed a graph augmentation method based on student groups. By enhancing the connections of low-degree student nodes within subgroups, we improved the topological structure and focused on disadvantaged students.
- We propose a cognitive diagnostic framework named GACD, which ensures that node representations on the augmented graph remain consistent with the original graph through a self-supervised alignment strategy, while separately capturing different student response behavior patterns.
- Extensive experiments on three public datasets demonstrate that our framework achieves optimal performance, validating the importance of student latent class analysis for cognitive diagnostic models.

2 Related Work

2.1 Cognitive diagnosis

Cognitive diagnosis aims to infer students' cognitive states based on their response data. IRT [7] and DINA [8] were early mainstream cognitive diagnostic models.IRT conceptualizes student ability as a unidimensional latent trait θ and employs logistic functions to model the probabilistic relationship between this latent trait and item responses. Furthermore, IRT models dynamically incorporate item parameters—including difficulty, discrimination, and guessing parameters —to optimize model fit across diverse educational measurement contexts [4]. Unlike the IRT, DINA [8] models student proficiency as binary mastery vectors via the Q-matrix (exercise-knowledge correlation matrix), where 1 indicates full command of a knowledge concept and 0 denotes non-mastery. This approach explicitly quantifies noise factors (i.e., slip and guessing) within its response function. With the advancement of deep learning [22, 23], research primarily manifests in the following two paradigms. One is to utilizing neural networks as diagnostic tools, where comprehensive input variables (e.g., students' cognitive states, item parameters difficulty/discrimination, and contextual factors) are nonlinearly mapped to response patterns; second is to leveraging deep learning to mine latent feature vectors from response logs and the Q-matrix, thereby enriching the embedding representations of students and items. Recent research has primarily focused on the latter aspect. For instance, Ma et al. [10] integrate the intrinsic correlations between students' cognitive states and knowledge concepts. Gao et al. [11] unify interaction and structural relationships through a multi-layer graph of student-exercise-knowledge concept relations. Qian et al. [12] refine cognitive representations by decomposing interaction graphs into response-signal-based subgraphs. Liu et al. [13] streamlines node neighborhood aggregation on interaction graphs to enable real-time diagnosis. However, most existing cognitive diagnosis models (CDMs) overlook the critical issue of imbalanced response data distribution.

2.2 Latent class analysis

LCA is a statistical method for identifying latent subgroups within a population [24]. It operates on the premise that associations among observed variables (manifest variables) can be explained by underlying categorical variables. The earliest application of LCA in cognitive diagnosis is the Mixed Rasch Model (MRM) [25], which integrates the Rasch model with latent class structures. Subsequent research has developed multiple variants of mixture IRT models based on diverse problem contexts and theoretical assumptions. For instance, Smit et al. [26] demonstrated that incorporating covariates related to latent classifications enhances class separation and reduces standard errors of parameter estimation. Vermunt et al. [27] applied multilevel mixture IRT models to mathematical assessments. De Jong et al. [28] implemented multilevel multidimensional mixture IRT models in cross-cultural studies to analyze latent variable structures. Crucially, mixture models bridge a methodological gap: traditional IRT handles continuous latent variables but not categorical ones, while LCA captures

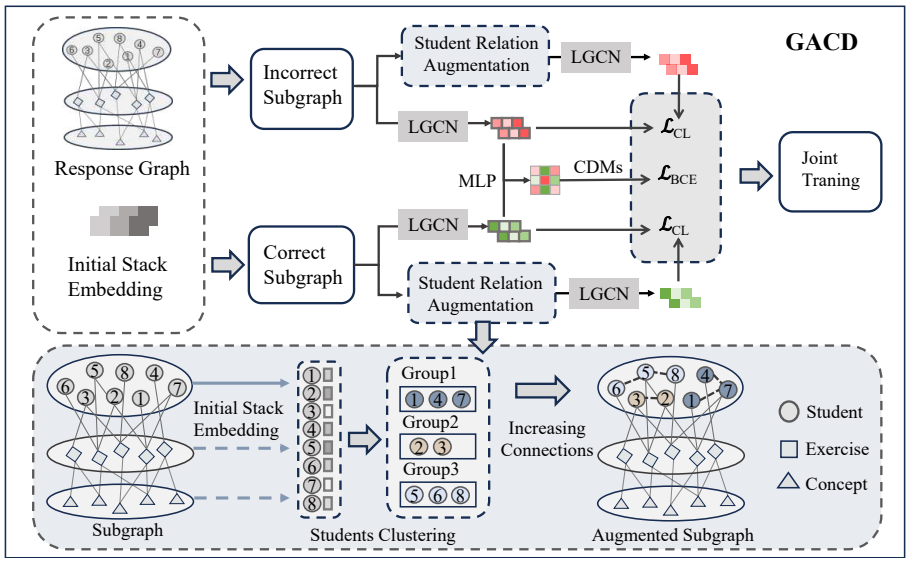


Fig. 1 An overview of GACD framework. The lower part of the diagram details the group-based enhancement process. Notably, in the correct and incorrect subfigures, the same student may be assigned to different groups.

categorical latent variables but ignores continuous traits [29]. This integration significantly improves model-data fit. However, in deep cognitive diagnosis frameworks, explicit utilization of LCA remains scarce. Although some models implicitly learn student group characteristics through parameter iteration, they generally lack systematic analysis and exploitation of latent class structures.

3 Methodology

In this section, we present the proposed GACD framework shown in Fig 1. The framework is designed to alleviate the data sparsity issue and the oversmoothing problem by exploiting the homophily among students. The overall architecture consists of four components: (1) Response Graph Decomposition, (2) Clustering-aware Structural Augmentation, (3) Dual-stream LightGCN Encoding with Contrastive Alignment, and (4) Multi-view Fusion and Diagnosis.

3.1 Response Graph Decomposition

Following the established graph-based cognitive diagnosis paradigm, we first construct a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ based on the response logs $\mathcal{T}$ and the Q-matrix Q. To explicitly capture the distinct semantic implications of correct and incorrect responses, we decompose the original interaction matrix I into two disjoint matrices: I_R (containing only correct responses) and I_W (containing only incorrect responses).

Consequently, we construct two specific subgraphs: the **Right Subgraph** ($\mathcal{G}_R$) and the **Wrong Subgraph** ($\mathcal{G}_W$). Their adjacency matrices, A_R and A_W, are formulated

as:

$$A_R = \begin{pmatrix} 0 & I_R & 0 \\ I_R^\top & 0 & Q \\ 0 & Q^\top & 0 \end{pmatrix}, \quad A_W = \begin{pmatrix} 0 & I_W & 0 \\ I_W^\top & 0 & Q \\ 0 & Q^\top & 0 \end{pmatrix} \tag{1}$$

where Q denotes the association between exercises and concepts. This decomposition allows the model to learn student mastery patterns from success and failure contexts independently.

3.2 Clustering-aware Structural Augmentation

To mitigate the sparsity of student-exercise interactions and enhance the representation of tail users, we propose a **Conditional Edge Generation** mechanism. This module operates identically on both latent student spaces of $\mathcal{G}_R$ and $\mathcal{G}_W$, but for brevity, we describe the general process.

We first obtain the initial feature representations of students, denoted as X_S. We employ the K-means algorithm to partition the student set S into K latent clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$. The clustering objective is to minimize the intra-cluster variance:

$$\min_{\mathcal{C}} \sum_{k=1}^{K} \sum_{s_i \in C_k} \|x_{s_i} - \mu_k\|_2^2, \tag{2}$$

where μ_k is the centroid of the k-th cluster. This step groups students with similar learning behaviors or cognitive states.

Within each cluster C_k, we identify students who lack sufficient diagnostic evidence. We define a subset $S_k^{tail} \subset C_k$ comprising the students ranked in the bottom 50% based on their interaction density or historical accuracy. To facilitate information propagation among these structurally isolated nodes, we augment the graph by adding edges between them.

Formally, the set of augmented edges $\mathcal{E}_{aug}$ is generated via the following conditional sampling process:

$$\mathcal{E}_{aug} = \bigcup_{k=1}^{K} \left\{ (u, v) \mid u \in S_k^{tail}, v \sim \text{Uniform}(S_k^{tail} \setminus \{u\}) \right\} \tag{3}$$

This formula indicates that for every tail student u, we randomly sample a peer v from the same tail subset to establish a structural link. We apply this augmentation to both subgraphs, resulting in the augmented adjacency matrices $A'_R = A_R + A^R_{aug}$ and $A'_W = A_W + A^W_{aug}$.

3.3 Dual-stream Representation Learning

We adopt LightGCN [30] as the backbone encoder to capture high-order connectivities. We establish two parallel encoding streams corresponding to the Right and Wrong views.

For each view $V \in \{R, W\}$, we propagate embeddings on both the original graph (A_V) and the augmented graph (A'_V). The graph convolution at layer l is defined as:

$$H_V^{(l)} = \left(\tilde{D}_V^{-\frac{1}{2}} \tilde{A}_V \tilde{D}_V^{-\frac{1}{2}}\right) H_V^{(l-1)}, \tag{4}$$

$$H_V'^{(l)} = \left(\tilde{D'}_V^{-\frac{1}{2}} \tilde{A'}_V \tilde{D'}_V^{-\frac{1}{2}}\right) H_V'^{(l-1)}, \tag{5}$$

where H_V and H'_V denote the embeddings learned from the original and augmented graphs, respectively. The final representations for each stream are obtained by aggregating the layer-wise outputs, e.g., $H_V = \sum_l \alpha_l H_V^{(l)}$.

To ensure the robustness of the structural augmentation, we employ Contrastive Learning to align the representations learned from the original and augmented structures. We minimize the InfoNCE loss for both Right and Wrong streams:

$$\mathcal{L}_{CL} = \sum_{V \in \{R,W\}} \sum_{u \in S} -\log \frac{\exp(\text{sim}(h_{u,V}, h'_{u,V})/\tau)}{\sum_{v \in S} \exp(\text{sim}(h_{u,V}, h'_{v,V})/\tau)} \tag{6}$$

where τ is the temperature parameter. This objective enforces the augmented graph to preserve the intrinsic semantic properties of the original graph while enriching the structure.

3.4 Fusion and Diagnosis

Constructing the final student representation requires effectively integrating the semantic information from both the Right and Wrong views. Following the fusion strategy proposed in ORCDF, we acknowledge that correct and incorrect interactions represent heterogeneous, and often contradictory, signals regarding a student's mastery level. Simple summation or concatenation may lead to information interference.

Therefore, we employ a parameterized fusion mechanism to adaptively aggregate the representations from the two subgraphs. Let H_R and H_W denote the final aggregated embeddings from the Right and Wrong streams, respectively. The fused representation H_{fuse} is computed as:

$$H_{fuse} = \sigma\left(H_R W_R + H_W W_W + b_{fuse}\right), \tag{7}$$

where $W_R, W_W \in \mathbb{R}^{d \times d}$ are trainable weight matrices that project the view-specific embeddings into a unified semantic space, b_{fuse} is the bias term, and $\sigma(\cdot)$ denotes the sigmoid function. This operation ensures that the distinct contributions of successful and failed attempts are weighted appropriately before integration.

To ensure the proposed framework is model-agnostic and compatible with various backbone Cognitive Diagnosis Models (CDMs), we introduce a transformation layer. Since certain CDMs (e.g., NeuralCDM) require the input embedding dimension to strictly align with the number of knowledge concepts Z, while our LightGCN encoder

operates in a latent dimension d, we apply a linear projection:

$$H = H_{fuse}W_t + b_t, \tag{8}$$

where $W_t \in \mathbb{R}^{d\times Z}$ and $b_t \in \mathbb{R}^Z$ are the transformation weights and bias. If the backbone CDM utilizes latent vectors (e.g., KaNCD), W_t acts as an identity mapping or a dimension alignment matrix where the output dimension matches the latent size.

Finally, the transformed student representations H_{in} are fed into the interaction function of the specific CDM to predict the probability of student u correctly answering exercise e_i. The prediction process is formulated as:

$$\hat{y}_{ui} = \mathcal{M}_{CDM}\left(H_{s_i}, H_{e_j}, H_c\right), \tag{9}$$

where $\mathcal{M}_{CDM}(\cdot)$ represents the CDMs.

The entire framework is trained end-to-end by minimizing the joint objective function combining the prediction loss and the contrastive regularization term.

4 Exeperiments

In this section, we evaluate the performance of the proposed GACD framework, verify the necessity of each component, and finally present a superparametric analysis.

4.1 Setups

4.1.1 Datasets Description.

We conduct experiments on three real-world datasets: Assist2017 [31], NeurIPS2020 [32], and Junyi [33]. Assist2017 is collected from the ASSISTments online tutoring platform, containing student exercise logs in mathematics. The NeurIPS2020 dataset is released for the NeurIPS 2020 Education Challenge, covering a broader range of K-12 mathematical problems. The Junyi dataset comes from the Junyi Academy online learning platform, including interaction logs on mathematics exercises over a longer period. These datasets vary in scale, knowledge structure, and sparsity, allowing for a comprehensive evaluation of the model's effectiveness and generalization ability. Detailed statistics are summarized in Table 2. Sparsity Rate represents the density of interactions in the student-exercise interaction matrix. Average Correctness Rate (ACR) reflects the average difficulty level of items across the dataset. We divide the datasets into training, validation, and test sets in a 7:1:2 ratio.

4.1.2 Baseline Models.

To showcase the effectiveness of GACD framework, we integrate it with various CDMs, As backbone models, we selected two categories of CDMs: those based on statistical methods—including Item Response Theory (IRT) [7] and Multidimensional IRT (MIRT) [34]—and those built on neural networks, namely Neural Cognitive Diagnosis (NCD) [9], Matrix Factorization with Knowledge Consistency (MFKC) [35], and

Table 2 Dataset summary

Datasets	**Assist2017**	**Neurips2020**	**Junyi**
Students	1709	2840	10000
Exercises	3162	6000	734
Concepts	102	268	734
Response Logs	390311	214328	408057
Sparsity	0.072	0.012	0.055
ACR	0.815	0.631	0.687
Q Density	1.22	4.14	1.00

Table 3 The comparative experimental results on three datasets. Bold indicates the best result.

Dataset	Assist2017				Neurips2020				Junyi			
Method	ACC	AUC	F1	DOA	ACC	AUC	F1	DOA	ACC	AUC	F1	DOA
IRT	85.56	88.27	91.66	-	69.95	75.22	76.85	-	76.27	80.37	83.49	-
LightGCN-IRT	86.75	89.43	92.11	-	71.43	76.56	78.63	-	77.46	81.74	84.31	-
ORCDF-IRT	86.83	89.43	92.13	-	71.45	76.68	78.24	-	77.57	81.68	84.42	-
GACD-IRT	**86.84**	**89.46**	**92.22**	-	**71.65**	**84.95**	**78.78**	-	**77.63**	**81.84**	**84.57**	-
MIRT	86.38	89.32	91.98	-	70.09	74.51	77.84	-	77.09	80.95	84.25	-
LightGCN-MIRT	86.07	88.58	91.89	-	70.49	75.47	78.25	-	77.29	81.17	84.28	-
ORCDF-MIRT	86.76	89.52	92.11	-	71.54	76.58	78.35	-	77.51	81.33	84.25	-
GACD-MIRT	**87.25**	**89.86**	**92.45**	-	**71.74**	**77.25**	**78.95**	-	**77.61**	**81.85**	**84.26**	-
NCDM	82.22	79.56	89.39	55.54	69.73	74.19	76.8	66.55	74.27	78.07	82.59	50.41
LightGCN-NCDM	83.29	82.94	89.87	56.94	70.55	75.58	77.00	67.57	75.00	74.8	82.43	55.44
ORCDF-NCDM	85.63	87.59	91.27	60.27	71.36	76.68	77.72	69.25	76.86	80.89	83.78	59.43
GACD-NCDM	**86.69**	**89.35**	**92.08**	**64.73**	**71.54**	**76.79**	**77.95**	**70.34**	**77.21**	**81.25**	**83.85**	**60.41**
CDMFKC	83.01	84.34	89.93	55.98	70.86	76.05	77.21	68.35	74.87	78.43	82.51	49.24
LightGCN-CDMFKC	83.09	83.46	89.70	57.20	71.83	76.82	78.42	69.90	76.19	80.83	82.47	59.33
ORCDF-CDMFKC	86.94	89.58	92.08	62.70	71.66	76.72	78.55	**72.15**	77.26	81.05	84.37	60.64
GACD-CDMFKC	**87.17**	**89.81**	**92.20**	**65.99**	**71.86**	**77.11**	**78.62**	72.14	**77.48**	**81.39**	**84.62**	**60.78**
KaNCD	85.03	86.45	90.98	57.90	71.01	75.97	77.91	68.67	75.27	75.21	83.32	53.86
LightGCN-KaNCD	85.49	87.05	91.25	58.38	71.32	76.59	77.33	69.23	76.61	79.89	83.6	58.88
ORCDF-KaNCD	87.14	89.73	92.27	62.42	71.59	76.92	78.18	70.09	77.54	81.43	84.35	60.58
GACD-KaNCD	**87.56**	**90.17**	**92.46**	**67.69**	**71.91**	**77.56**	**78.42**	**71.14**	**77.80**	**81.84**	**84.66**	**61.22**
KaNCD	85.03	86.45	90.98	57.90	71.01	75.97	77.91	68.67	75.27	75.21	83.32	53.86
LightGCN-KaNCD	85.49	87.05	91.25	58.38	71.32	76.59	77.33	69.23	76.61	79.89	83.6	58.88
ORCDF-KaNCD	87.14	89.73	92.27	62.42	71.59	76.92	78.18	70.09	77.54	81.43	84.35	60.58
GACD-KaNCD	**87.66**	**90.19**	**92.62**	**68.07**	**71.96**	**77.53**	**78.56**	**70.75**	**77.78**	**82.13**	**84.66**	**61.37**

Knowledge-aware Neural Cognitive Diagnosis (KaNCD) [36]. For comparative evaluation, we adopted two representative frameworks: LightGCN (LGCN) [30] and the ORCDF framework [12].

4.1.3 Evaluation Metrics and Parameters Setting.

To evaluate classification performance, we used three metrics: Accuracy (ACC), AUC, and F1-score. For interpretability assessment, we utilized the Degree of Agreement (DOA) [12]. All models were optimized with Adam, with parameters initialized via Xavier. For fair comparison, we set the embedding size d to 32 for MIRT and KaNCD, and to the number of knowledge concepts C for NCDM and MFKC. The batch size was fixed at 4096 across all datasets. The hyperparameters were configured as follows: learning rate $= 0.04$, K with the range $\{1, 2, 3, 4, 5, 6\}$, τ within the range $\{0.1, 0.5, 1.0, 3.0, 5.0\}$. All reported results are averaged over five runs with different random seeds.

4.2 Comparative Experiments

The comparative experimental results on the three datasets are presented in Table 3. As shown, GACD consistently improves all baseline CDMs across every dataset. This demonstrates that GACD effectively mitigates the impact of imbalanced response data distribution and achieves more accurate cognitive diagnosis. Among the methods, MIRT demonstrates enhanced diagnostic capabilities by modeling students' multidimensional latent abilities, overcoming the limitation of traditional IRT's unidimensional ability assumption. KaNCD achieves superior diagnostic outcomes by implicitly learning inter-knowledge relationships through latent vectors, leveraging mastered knowledge components to infer proficiency in uncovered ones. LightGCN significantly improves cognitive diagnosis models by using graph neural networks to model heterogeneous interaction graphs while integrating existing models as response prediction modules. The ORCDF framework effectively distinguishes correct and erroneous behavior patterns by decomposing the heterogeneous graph into right-response and wrong-response graphs. However, its random response reversal strategy for data augmentation during training may unintentionally reinforce erroneous information in the original data. GACD employs a group-homogeneity enhancement strategy, avoiding random or uncertainty-based operations. This allows it to more reliably supplement semantically relevant interaction information for data-scarce student nodes. At the same time, a contrastive learning alignment strategy ensures that the enhanced graph representations remain consistent with the original semantics, effectively preventing the information interference or loss risks.

Table 4 Ablation study on three datasets.

Method	Assist2017				NeurIPS2020				Junyi			
	ACC	AUC	F1	DOA	ACC	AUC	F1	DOA	ACC	AUC	F1	DOA
GACD-KaNCD	**87.56**	**90.17**	**92.46**	67.69	**71.91**	**77.56**	**78.42**	71.14	**77.80**	**81.84**	**84.66**	**61.22**
w/o CSA	84.39	87.26	91.13	**68.62**	69.22	73.42	77.44	**73.68**	75.39	79.66	83.05	60.65
w/o RGD	85.72	88.60	91.62	60.84	66.78	69.09	74.96	59.35	73.33	79.94	83.31	55.94

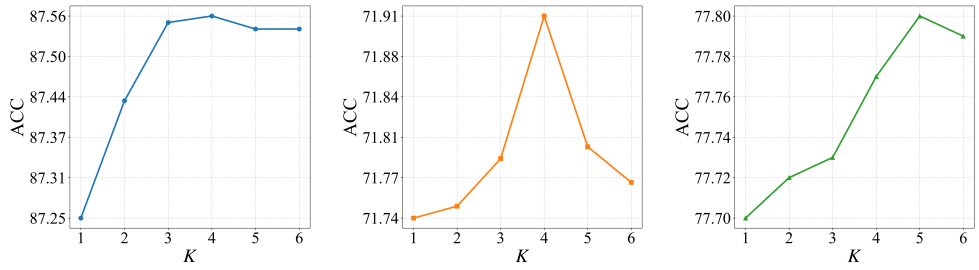
4.3 Ablation Study

To diagnose the relative importance of each key component within the GACD framework, we perform an ablation study and evaluating their impacts through controlled experiments. the ablation analysis is conducted using the following two versions:

- w/o CSA: Remove Clustering-aware structural augmentation from the GACD framework without clustering students and increasing intra class connectivity.
- w/o RGD: Remove response graph decomposition from the GCAD framework and replace it with perform convolution on the entire response graph.

Based on the findings presented in Table 3, we selected KaNCD as the backbone for ablation study. The experimental findings on the three datasets are presented in Table 4. The results indicate that eliminating either module leads to a significant decline in the ACC, AUC, and F1-score. The DOA outcomes for w/o RGD are

markedly lower than those for GACD-KaNCD and w/o CSA, suggesting that differentiation of response signals plays a pivotal role in enhancing the interpretability of the model. The clustering enhancement might cause individual representations to converge towards their respective cluster centroids, consequently diminishing the nuanced differences among students within the same group. This could explain why GACD-KaNCD exhibits a lower DOA compared to w/o CSA.



(a) Assist2017 (b) NeurIPS2020 (c) Junyi

Fig. 2 Effect of K (Number of groups) on three datasets. Based on KaNCD.

4.4 Hyperparameter Analysis

To analyze the selection of the hyperparameter K in the K-means clustering method described in 3.2, we conduct experiments using ACC as the evaluation metric. ACC offers an intuitive measure of model performance under different hyperparameter settings. The results on the three datasets are summarized in Fig 2. The optimal clustering grouped students into 4 clusters for Assist2017 and Neurips2020 datasets, and 5 clusters for Junyi dataset. To investigate the reasons, we statistically analyzed the degree information of students in each dataset presented in Table 1. Combined with Table 2, the Junyi dataset contains more student nodes, yet exhibits a significantly lower and more dispersed average node degree compared to the Assist2017 and NeurIPS2020 datasets. This explains why the Junyi dataset requires a greater number of student clusters: when most student nodes suffer from severe information scarcity. If only a few clusters are used, each cluster will contain too many students with diverse characteristics. Consequently, the edges added between them may not accurately capture true student similarities, leading to inefficient information propagation. By increasing the number of clusters, students within each group become more homogeneous. This results in higher-quality added edges and more precise information dissemination, thereby better compensating for the original structural sparsity. It follows that on the premise of avoiding over classification, GACD needs a larger K when applied to Sparse datasets of student interaction.

5 Conclusion

This paper proposes the Group-Aware Cognitive Diagnosis (GACD) framework to address the challenge of data sparsity and imbalance in student response data, which leads to inaccurate cognitive assessment. The core solution involves grouping students with similar patterns and artificially augmenting connections for data-sparse students within their groups. Specifically, the framework first decomposes the interaction graph into separate subgraphs for correct and wrong responses, then performs clustering-aware structural augmentation to enhance connectivity for tail students, followed by dual-stream representation learning with a contrastive alignment strategy to maintain semantic consistency, and finally fuses the multi-view representations for diagnosis. Experiments on three real-world datasets demonstrate that GACD consistently improves the performance of various backbone cognitive diagnosis models, validating its effectiveness in achieving more reliable assessment by mitigating data sparsity through group-aware graph augmentation and self-supervised learning. These findings demonstrate the effectiveness of applying latent class analysis to cognitive diagnosis and offer insight for the design of future diagnostic frameworks.

Declarations

Funding

This work was supported by the Anhui Provincial Higher Education Quality Engineering Project(NO.2024xjzlts064), the Anhui Provincial Teaching Reform and Research Project (NO.2023jyxm0865), the Anhui Provincial University Excellent Young Teachers Training Project (NO.YQYB2025056), the Scientific research Project of the Education Department of Anhui Province (NO.2025AHGXZK40255), and the Doctoral Research Start-up Project (NO.BYKQ202518).

Competing interests

The authors declare no competing interests.

Clinical Trial Registration

Not applicable.

Ethics, Consent to Participate, and Consent to Publish declarations

Not applicable.

References

[1] Deho, O.B., Liu, L., Li, J., Liu, J., Zhan, C., Joksimovic, S.: When the past != the future: Assessing the impact of dataset drift on the fairness of learning analytics models. IEEE Trans. Learn. Technol. **17**, 1007–1020 (2024) https://doi.org/10.1109/TLT.2024.3351352

[2] Yue, M., Jong, M.S.Y., Ng, D.T.K.: Understanding k–12 teachers' technological pedagogical content knowledge readiness and attitudes toward artificial intelligence education. Educ. Inf. Technol. **29**(15), 19505–19536 (2024) https://doi.org/10.1007/s10639-024-12621-2

[3] Deho, O.B., Joksimovic, S., Liu, L., Li, J., Zhan, C., Liu, J.: Assessing the fairness of course success prediction models in the face of (un)equal demographic group distribution. In: Proc. 10th ACM Conf. Learn. @ Scale. L@S '23, pp. 48–58. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3573051.3593381

[4] Liu, Y., Zhang, T., Wang, X., Wu, J., Chen, M., He, X., Li, X.: New development of cognitive diagnosis models. Front. Comput. Sci. **17**(1), 171604 (2023) https://doi.org/10.1007/s11704-022-1128-3

[5] Ma, H., Huang, Z., Tang, W., Zhang, X.: Exercise recommendation based on cognitive diagnosis and neutrosophic set. In: Proc. 25th IEEE Int. Conf. Comput. Support. Coop. Work Des., pp. 1467–1472 (2022). https://doi.org/10.1109/CSCWD54268.2022.9776197

[6] Wang, C., Zhu, H., Zhu, C., Zhang, X., Chen, E., Xiong, H.: Personalized employee training course recommendation with career development awareness. In: Proc. Web Conf. 2020. WWW '20, pp. 1648–1659. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3366423.3380236

[7] Haberman, S.J.: Identifiability of parameters in item response models with unconstrained ability distributions. ETS Res. Rep. Ser. **2005**(2), 22 (2005) https://doi.org/10.1002/j.2333-8504.2005.tb02001.x

[8] Torre, J.: Dina model and parameter estimation: A didactic. J. Educ. Behav. Stat. **34**(1), 115–130 (2009) https://doi.org/10.3102/1076998607309474

[9] Wang, F., Liu, Q., Chen, E., Huang, Z., Yin, Y., *et al.*: Neural cognitive diagnosis for intelligent education systems. Proc. AAAI Conf. Artif. Intell. **34**(4), 6153–6161 (2020) https://doi.org/10.1609/aaai.v34i04.6080

[10] Ma, H., Li, M., Wu, L., Zhang, H., Cao, Y., Zhang, X., Zhao, X.: Knowledge-sensed cognitive diagnosis for intelligent education platforms. In: Proc. 31st ACM Int. Conf. Inf. Knowl. Manag. CIKM '22, pp. 1451–1460. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3511808.3557372

[11] Gao, W., Liu, Q., Huang, Z., Yin, Y., Bi, H., Wang, M., Ma, J., Wang, S., Su, Y.: Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In: Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. SIGIR '21, pp. 501–510. Association for Computing Machinery, New York, NY, USA (2021).

https://doi.org/10.1145/3404835.3462932

[12] Qian, H., Liu, S., Li, M., Li, B., Liu, Z., Zhou, A.: Orcdf: An oversmoothing-resistant cognitive diagnosis framework for student learning in online education systems. In: Proc. 30th ACM SIGKDD Conf. Knowl. Discov. Data Min. KDD '24, pp. 2455–2466. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3637528.3671988

[13] Liu, S., Shen, J., Qian, H., Zhou, A.: Inductive cognitive diagnosis for fast student learning in web-based intelligent education systems. In: Proc. ACM Web Conf. 2024. WWW '24, pp. 4260–4271. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589334.3645589

[14] Zhang, G., Yuan, G., Xu, Z., Zhang, Y., Ren, J., Deng, Z., Cheng, D.: Harnessing llm for noise-robust cognitive diagnosis in web-based intelligent education systems. In: Proc. ACM Web Conf. 2026. WWW '26, pp. 4093–4104. Association for Computing Machinery, New York, NY, USA (2026). https://doi.org/10.1145/3774904.3792493

[15] Urhahne, D., Wijnia, L.: Theories of motivation in education: An integrative framework. Educ. Psychol. Rev. **35**(2), 45 (2023) https://doi.org/10.1007/s10648-023-09767-9

[16] Zhang, G., Zhang, S., Yuan, G.: Bayesian graph local extrema convolution with long-tail strategy for misinformation detection. ACM Trans. Knowl. Discov. Data **18**(4), 89–18921 (2024) https://doi.org/10.1145/3639408

[17] Cho, S., Cohen, A.S.: A multilevel mixture irt model with an application to dif. J. Educ. Behav. Stat. **35**, 336–370 (2010)

[18] Witkin, H.A., Moore, C.A., Goodenough, D.R., Cox, P.W.: Field-dependent and field-independent cognitive styles and their educational implications. Rev. Educ. Res. **47**, 1–64 (1977)

[19] Alloway, T.P., Alloway, R.G.: Investigating the predictive roles of working memory and iq in academic attainment. J. Exp. Child Psychol. **106**(1), 20–29 (2010) https://doi.org/10.1016/j.jecp.2009.11.003

[20] Mayer, R.E., Massa, L.J.: Three facets of visual and verbal learners: Cognitive ability, cognitive style, and learning preference. J. Educ. Psychol. **95**(4), 833–846 (2003) https://doi.org/10.1037/0022-0663.95.4.833

[21] Veenman, M.V.J., Van Hout-Wolters, B.H.A.M., Afflerbach, P.: Metacognition and learning: conceptual and methodological considerations. Metacogn. Learn. **1**(1), 3–14 (2006) https://doi.org/10.1007/s11409-006-6893-0

[22] Zhang, G., Yuan, G., Cheng, D., Liu, L., Li, J., Zhang, S.: Mitigating propensity

bias of large language models for recommender systems. ACM Trans. Inf. Syst. **43**(6), 150–115026 (2025) https://doi.org/10.1145/3736404

[23] Zhang, G., Yuan, G., Cheng, D., Liu, L., Li, J., Zhang, S.: Towards fair graph representation learning by overcoming social homophily. ACM Trans. Intell. Syst. Technol. **17**(2), 36–13625 (2026) https://doi.org/10.1145/3785503

[24] Clogg, C.C.: Latent class models. In: Arminger, G., Clogg, C.C., Sobel, M.E. (eds.) Handbook of Statistical Modeling for the Social and Behavioral Sciences, pp. 311–359. Springer, Boston, MA, USA (1995). https://doi.org/10.1007/978-1-4899-1292-3_6

[25] Rost, J.: Rasch models in latent classes: An integration of two approaches to item analysis. Appl. Psychol. Meas. **14**(3), 271–282 (1990) https://doi.org/10.1177/014662169001400305

[26] Smit, J.A., Kelderman, H., Flier, H.: Collateral information and mixed rasch models. Methods Psychol. Res. Online **4**(3), 1–13 (1999)

[27] Vermunt, J.K.: Latent class and finite mixture models for multilevel data sets. Stat. Methods Med. Res. **17**(1), 33–51 (2008) https://doi.org/10.1177/0962280207081238

[28] Jong, M.G., Steenkamp, J.B.E.M.: Finite mixture multilevel multidimensional ordinal irt models for large scale cross-cultural research. Psychometrika **75**(1), 3–32 (2010) https://doi.org/10.1007/s11336-009-9134-z

[29] Wang, X., Tan, G., Wang, X., Zhang, M., Luo, C.: The mixture item response theory models and its application traces. Adv. Psychol. Sci. **22**(3), 540–548 (2014) https://doi.org/10.3724/SP.J.1042.2014.00540

[30] Wang, R., Lou, J., Jiang, Y.: Lightgcan: A lightweight graph convolutional attention network for user preference modeling and personalized recommendation. Expert Syst. Appl. **232**, 120741 (2023) https://doi.org/10.1016/j.eswa.2023.120741

[31] Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User Model. User-Adapt. Interact. **19**(3), 243–266 (2009) https://doi.org/10.1007/s11257-009-9063-7

[32] Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernandez-Lobato, J.M., Turner, R.E., Baraniuk, R., Barton, C., Jones, S.L.P., Woodhead, S., Zhang, C.: Results and insights from diagnostic questions: The neurips 2020 education challenge. In: Proc. Adv. Neural Inf. Process. Syst., vol. 34 (2021)

[33] Chang, H., Hsu, H., Chen, K.: Modeling exercise relationships in e-learning: A unified approach. In: Proc. 8th Int. Conf. Educ. Data Min. (2015)

[34] Chalmers, R.P.: mirt: A multidimensional item response theory package for the r environment. J. Stat. Softw. **48**(6), 1–29 (2012) https://doi.org/10.18637/jss.v048.i06

[35] Li, S., Guan, Q., Fang, L., Xiao, F., He, Z., He, Y., Luo, W.: Cognitive diagnosis focusing on knowledge concepts. In: Proc. 31st ACM Int. Conf. Inf. Knowl. Manag. CIKM '22, pp. 3272–3281. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3511808.3557096

[36] Wang, F., Liu, Q., Chen, E., Huang, Z., Yin, Y., Wang, S., Su, Y.: Neuralcd: A general framework for cognitive diagnosis. IEEE Trans. Knowl. Data Eng. **35**, 8312–8327 (2023) https://doi.org/10.1109/TKDE.2022.3229871