



A deep cross-modal neural cognitive diagnosis framework for modeling student performance

Lingyun Song^{a,b,*}, Mengting He^{a,b}, Xuequn Shang^{a,b}, Chen Yang^{a,b}, Jun Liu^c, Mengzhen Yu^{a,b}, Yu Lu^d

^a School of Computer Science, Northwestern Polytechnical University, Xi'an, 710129, China

^b Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an, 710129, China

^c SPKLSTN Lab, Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

^d Advanced Innovation Center for Future Education, Beijing Normal University, Beijing, 100875, China

ARTICLE INFO

Keywords:

Deep neural network
Cross-modal
Cognitive diagnosis
Knowledge concept

ABSTRACT

In intelligent education systems, one fundamental task is to predict student performance on new exercises and estimate the knowledge proficiency of students on knowledge concepts. Existing prediction methods are mainly constructed based on the classical cognitive diagnosis framework *MIRT*, where student performance on exercises are modeled as the interaction results of exercises' trait vectors and students' knowledge proficiency. The trait vector learning of exercises has a big effect on the estimation results of the knowledge proficiency. However, when learning the trait vectors, existing methods cannot exploit the rich contents of cross-modal exercises that are closely related to the traits of exercises. This makes it difficult for these methods to best cope with common cross-modal exercises. Besides, existing methods overlook the intrinsic complexity of examined concepts, which actually affects exercise traits, such as exercise difficulty. To address these issues, we propose a deep Cross-Modal Neural Cognitive Diagnosis framework (CMNCD), which mainly has two appealing advantages: (i) By extending *MIRT* under the framework of deep neural networks, CMNCD can effectively explore the fine-grained semantic information in the cross-modal contents of exercises for modeling student performance. (ii) CMNCD investigates the complexity of examined concepts based on the prerequisite relationships among concepts and incorporates it into the learning of exercises' trait vectors. Extensive experiments on several real-world datasets show that our CMNCD outperforms state-of-the-art cognitive diagnosis methods.

1. Introduction

Being a crucial issue in education field, cognitive diagnosis (Hooshyar, Huang, & Yang, 2022; Wang, Liu et al., 2022) aims to infer students' knowledge proficiency (or called student trait) based on students' performance (i.e., right or wrong responses) on testing exercises, which is conducive to designing personalized instructions for student learning (Chen & Duh, 2008; Chen, Liu, & Chang, 2006). For example, a toy example of cognitive diagnosis system is shown in Fig. 1(a). Many prevalent Cognitive Diagnosis Methods (CDMs) are constructed based on the Item Response Theory (IRT) (Baylari & Montazer, 2009; Lord, 2012), which models the student performance as an interaction result of student knowledge proficiency and exercise traits, such as exercises difficulty or discrimination. Thus, exercise traits play an important role in the prediction of student performance.

One limitation of the IRT based methods lies in that they only can infer an overall knowledge proficiency level for each student, without explaining the detailed proficiency level over examined concepts, e.g., *Multiplication*. To address this issue, some researchers extended IRT to Multidimensional IRT (MIRT) (Yao & Schwarz, 2006), which represents student knowledge proficiency and exercise traits by multidimensional vectors in concept space, rather than the scalars used in IRT. During the parameter training phase, the MIRT based methods usually use probabilistic statistical methods (e.g., Maximum Likelihood Estimation (MLE)) to estimate the knowledge proficiency vectors of students and the trait vectors of exercises based on students' performance records.

Although achieving impressive performance prediction results, these MIRT based methods (De La Torre, 2009; Embretson & Yang,

* Corresponding author at: School of Computer Science, Northwestern Polytechnical University, Xi'an, 710129, China.

E-mail addresses: lysong@nwpu.edu.cn (L. Song), hmt468@mail.nwpu.edu.cn (M. He), shang@nwpu.edu.cn (X. Shang), yangchen803@mail.nwpu.edu.cn (C. Yang), liukeen@mail.xjtu.edu.cn (J. Liu), yumengzhen@mail.nwpu.edu.cn (M. Yu), luyu@bnu.edu.cn (Y. Lu).

<https://doi.org/10.1016/j.eswa.2023.120675>

Received 7 May 2022; Received in revised form 31 May 2023; Accepted 31 May 2023

Available online 3 June 2023

0957-4174/© 2023 Elsevier Ltd. All rights reserved.

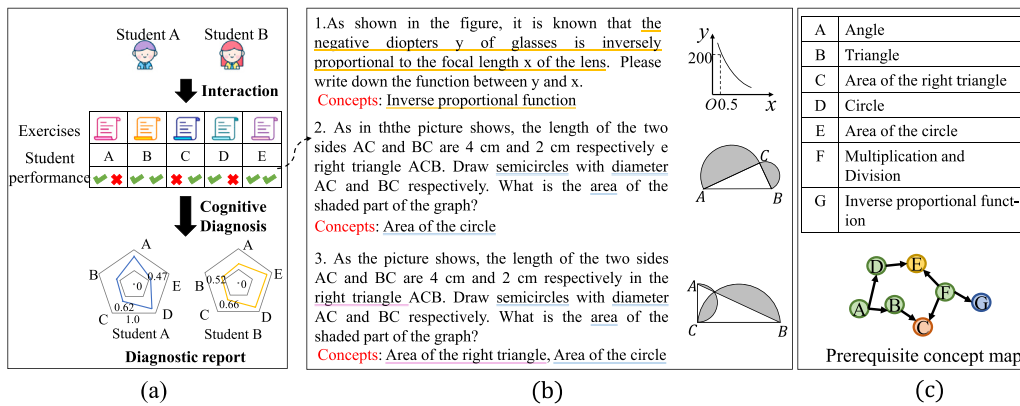


Fig. 1. (a) A toy example of cognitive diagnosis system, which exploits students' performance records on exercises for inferring student proficiency on concepts. (b) Examples of exercises in which their cross-modal contents are closely related to the examined concepts, and thus can provide valuable clues for the exercise traits based on concepts. The examined concepts and their related question texts are underlined in the same color. (c) A toy example of the prerequisite concept map that shows the prerequisite relations among the examined concepts.

2013) still have the following limitations which if addressed would improve their performance and applicability.

- (1) Existing methods cannot best cope with cross-modal exercises, which limits their performance and applicability. This is because these traditional MIRT based methods model exercises' trait vectors only based on student performance, which overlooks the rich cross-modal information in exercises that are closely related to exercise traits (Cheng & Liu, 2019; Huang et al., 2017; Tong, Zhou, & Wang, 2020), e.g., exercise difficulty. In this case, when students got the same performance on two exercises examining the same concept, e.g., Exercise 2 and 3 in Fig. 1(b), these methods tend to learn the same traits for the two exercises, even if their traits actually have significant difference. Although some recent works (Cheng & Liu, 2019; Tong et al., 2020) take into account the text contents of exercises when learning exercises' trait vectors, they still overlook the image contents of exercises that also have an impact on the trait vectors. For example, as shown in Fig. 1(b), even if Exercise 2 and 3 have the same text contents, their traits of difficulty still may vary with different images attached to them.
- (2) Existing methods overlook the effect of the cognitive complexity of concepts on exercises' trait vectors, each dimension of which indicates the traits of exercises in terms of each examined concept. The complexity of concepts refers to the difficulty involved in the learning of the concept for students, and can be characterized by the prerequisite relations (Liang, Ye, Wang, Pursel, & Giles, 2018; Pan, Li, Li, & Tang, 2017) among concepts. In general, the traits of one exercise are closely related to the complexity of examined concepts. For example, as seen from Fig. 1(b) and (c), the more difficult the exercises are (e.g., by comparing Exercise 1 and 2), the more prerequisites of the concepts examined by the exercises. Overlooking the concept complexity may result in illogical exercises' trait vectors, e.g., high exercise difficulty on the concepts of lower complexity, and thus leads to incorrect prediction results of student performance and student knowledge proficiency.

According to above discussions about the limitations, we have two hypotheses for cognitive diagnosis modeling, including: (1) Combining the different modality information from exercises' contents helps to improve the trait learning of cross-modal exercises, which is conducive to improving the performance of cognitive diagnosis based on cross-modal exercises. (2) The cognitive complexity of knowledge concepts is beneficial for learning exercises' trait vectors, which would lead to the performance improvement of cognitive diagnosis tasks.

1.1. Overview of our model

To solve these issues, we combine the framework of MIRT with Deep Neural Networks (DNN) to propose a deep Cross-Modal Neural Cognitive Diagnosis framework (CMNCD) for predicting student performance on cross-modal exercises. In CMNCD, both the cross-modal contents and the complexity of concepts are incorporated into the modeling of student performance. To effectively explore the cross-modal contents, CMNCD first exploits a new Co-Attention based Gated Cross-modal Fusion (CA-GCF) mechanism, which can adaptively fuse the information of the interrelated visual and textual information describing fine-grained semantics of exercises. Then, CMNCD selects relevant cross-modal information for estimating the trait vectors of exercises under the guidance of examined concepts. Besides, CMNCD models the complexity of concepts by the prerequisite relations among concepts and investigates its employment in the estimation of exercises' trait vectors. Finally, CMNCD predicts student performance based on cross-modal exercises by a deep MIRT module, which models the complex interactions between the trait vectors of students and exercises via a DNN.

1.2. Contributions and organization

The technical contributions of our work can be summarized as follows:

- (1) We propose a new deep neural cognitive framework for cross-modal exercises, namely CMNCD. In CMNCD, the rich visual and textual contents of cross-modal exercises are explored to learn distinguishable trait vectors of exercises, which is the key for improving the performance of cognitive diagnosis. To the best of our knowledge, CMNCD is the first cognitive diagnosis model that takes into account cross-modal contents of exercises for diagnosing the knowledge proficiency of students.
- (2) We investigate the method of how to model the complexity of knowledge concepts by the prerequisite relations among these concepts, and employ the concept complexity to enhance the learning of exercises' traits including difficulty and discrimination. Taking into account the concept complexity for modeling exercises' traits helps to learn logical traits that conform to the characteristics of human cognition. For example, the higher the complexity of examined concepts, the bigger the difficulty of exercises. The correct judgement of the difficulty traits of exercises helps to accurately diagnose students' proficiency level over concepts.

- (3) We conduct extensive experiments on several benchmark datasets to evaluate the performance of CMNCD. Experimental results show that CMNCD outperforms state-of-the-art cognitive diagnosis methods on all experimental datasets.

The rest of this paper is organized as follows: In Section 2, we review the related work. In Section 3, we introduce the cross-modal cognitive diagnosis task and briefly review the MIRT. In Section 4, we present the framework of CMNCD and introduce each of its module in detail. We introduce experimental datasets in 5 and report the experimental results in Section 6. Finally, we present the limitation our work in Section 7 and conclude the paper in Section 8.

2. Related work

In this part, we introduce the related works about *Cognitive diagnosis* and *Cross-modal representation learning*.

2.1. Cognitive diagnosis

Diagnosing students' mastery level of knowledge concepts based on their performance on exercises has been a key research topic in intelligent education systems (Gao et al., 2021; Liu, 2021; Liu, Zou et al., 2021). Prevalent CDMs are mainly constructed based on the MIRT, which models the interactions between students' proficiency vectors and exercises' trait vectors by a logistic-like function. Specifically, these methods mainly use MLE algorithms to estimate the trait vectors and student's proficiency vectors based on student performance records.

For example, Yao and Schwarz (2006) proposed a compensatory multidimensional two-parameter partial credit model for cognitive diagnosis, which exploits Markov chain Monte Carlo methods to estimate the multi-dimensional trait vectors of both students and exercises. Junker and Sijtsma (2001) proposed the noisy input, deterministic and gate model for cognitive diagnosis, which endows concepts with different importance when inferring diagnosis results. Embretson and Yang (2013) proposed a multi-component latent trait based method for cognitive diagnosis, which divides exercises into multiple components for inferring diagnosis results. Cheng and Liu (2019) proposed a deep item response framework for cognitive diagnosis, which explores the rich semantics of exercise texts for enhancing the learning of exercises' trait vectors. Chen, Culpepper, and Liang (2020) proposed a sparse latent class model for cognitive diagnosis, which infers diagnosis results by classifying students into one of attribute profiles. Wang, Liu et al. (2022) proposed a neural cognitive diagnosis framework, which models the interactions between the trait vectors of students and exercises via a deep neural network. Tong et al. (2021) proposed an item response ranking based cognitive diagnosis method, which incorporates the monotonicity between student's proficiency and the probability of giving correct answers into parameter optimization. Gao et al. (2021) proposed a relation map driven cognitive diagnosis framework, which captures multiple different relations between the nodes representing students, exercises and concepts in hierarchical layout maps. Wang, Ma, Zhao, Li and He (2022) proposed to track knowledge proficiency of students dynamically by a calibrated Q-matrix-based cognitive diagnosis framework. Qi et al. (2023) proposed a deep cognitive diagnosis framework comprising three layers of neural networks, which exploits the interactions among concepts and the quantitative relations between exercises and concepts to improve diagnosis results. Cheng et al. (2021) proposed an importance of knowledge concept-based neural cognitive diagnosis framework, where the importance of concepts is represented by their frequency examined by exercises. Wang, Ma et al. (2022) proposed a GCN-based deep neural cognitive diagnosis model, where Q-matrix can be calibrated based on the GCN theory. Wu et al. (2023) proposed a multi-relational cognitive diagnosis model, where diverse interactions between students and exercises are employed to enhance the representation learning of students and exercises. Ma

et al. (2023) proposed a cognitive diagnosis model that measures students' cognitive status comprehensively based on the neutrosophic set theory. Wang, Yan, Zeng, Tian, Dong and et al. (2023) proposed a unified interpretable deep cognitive diagnosis framework, where multi-channel cognitive diagnosis mechanisms are constructed based on the fusion of classical cognitive models, e.g., MIRT and DINA (De La Torre, 2009). Besides, we review more recent cognitive diagnosis methods and summarize them in Table 1.

Although achieving promising performance, the above methods overlook the cross-modal contents of exercises, which are closely related to exercise traits (Cheng & Liu, 2019; Huang et al., 2017), such as exercise difficulty or discrimination. Besides, these methods overlooks the effects of the complexity of the concepts examined by exercises on exercises' traits. By contrast, the proposed CMNCD incorporates both the cross-modal contents of exercises and concept complexity into the learning of exercises' trait vectors.

2.2. Cross-modal representation learning

Due to the powerful representation learning ability, DNN have been extensively used for constructing feature representations in application scenarios involving time-series data (e.g., EEG-based recognition Shoeibi, Rezaei et al., 2022; Shoeibi, Sadeghi et al., 2021) and multiple modality data (Shoeibi, Khodatars et al., 2022, 2021), e.g., visual-language grounding (Song, Liu, Qian, & Chen, 2019). Existing DNN based cross-modal learning methods usually combine feature representations of different modality data into an unified feature representation (Bokade et al., 2021). For example, Kim et al. (2016) proposed a low-rank bilinear attention network for fusing the features of different modality data. Gao, Beijbom, Zhang, and Darrell (2016) proposed a compact bilinear pooling model for multi-modal feature fusion, which learns joint representations by an outer product of the feature vectors of two different modality data. Yu, Yu, Fan, and Tao (2017) proposed a multi-modal factorized bilinear pooling method for cross-modal feature fusion. Liu, Zhang, and Gulla (2020) proposed to fuse textual and visual information in an attentive recurrent neural network via a dynamic contextual attention mechanism. Gupta, Suman, and Ekbal (2021) proposed to fuse the features from different modalities via a hierarchical deep multi-modal network. Mou et al. (2021) proposed an attentional CNN-LSTM network for cross-modal fusion. Zhang, Li and Kong (2021) proposed a density estimation-based regression framework for cross-modal fusion, where the cross-modal interactions in multiple image locations are used to learn abundant deep features. Bakkali, Ming, Coustaty, Rusiñol, and Terrades (2023) proposed an inter-modality cross-attention mechanism for cross-modal representations based on language and visual cues, where both intra- and inter-modality relationships are used to improve the feature fusion within and across modalities. Mohammed, Omeroglu, and Oral (2023) proposed a multi-modal and multi-Layer hybrid fusion network that integrates complementary information of different modalities to learn cross-modal representations. Lin, Bas, Singh, Swaminathan, and Bhotika (2023) proposed a category-aware multimodal attention network, which learns cross-modal representations based on multi-layered category-aware visual and textual descriptions. To filter out noisy visual and textual features, these methods use visual and textual attention mechanisms to highlight important image regions and textual words. Although achieving impressive performance, these methods overlook the fine-grained interactions between the images and texts when learning visual or textual attention.

In recent works for VQA tasks, researchers addressed the above issues by learning cross-modal features via co-attention mechanisms (Guo et al., 2020; Liu, Zhang et al., 2021; Lu, Yang, Batra, & Parikh, 2016; Nam, Ha, & Kim, 2017; Yu, Xue, Jiang, An and Li, 2021). Specifically, these methods first connect visual attention and textual attention by an affinity matrix, which represents the similarity between visual and textual modalities. Then, they jointly reason about visual

Table 1

Summary of related cognitive diagnosis methods. The standard preprocessing includes the initialization of the trait vectors for students and exercises, and the encoding of concepts.

Works	Dataset	Preprocessing	DL	Validation Criteria	Performance
IK-NeuralCD (Cheng et al., 2021)	Math2 (Liu et al., 2018)	Standard procedure	MLP	ACC, RMSE, AUC	ACC = 0.701, RMSE = 0.447, AUC = 0.771
GKT-CD (Zhang, Mo, Chen and He, 2021)	Math (Zhang, Mo et al., 2021), ASSIST ^a , KDD_Cup2010 ^b	Standard procedure	GNN	Accuracy, RMSE, AUC	ACC = 0.74, 0.73, 0.76, for ASSIST, KDD and Maths
CD-PG (Xu, Li, Liu, Lv, & Yu, 2021)	MOOC (Xu et al., 2021)	Standard procedure	N/A	RMSE	RMSE = 1.63
DeepCDF (Gao, Zhao, Li, Zhao, & Zeng, 2022)	Math1, Math2 (Liu et al., 2018)	Standard procedure	MLP	RMSE, MAE	RMSE = 0.409, 0.442 for Math1, Math2, MAE = 0.345, 0.372 for Math1, Math2
DIRT (Cheng et al., 2019)	Zhixue ^c (Cheng et al., 2019)	Standard procedure, Student filtering	LSTM, MLP	RMSE, MAE, AUC, ACC	RMSE = 0.406, MAE = 0.238, AUC = 0.750, ACC = 0.761
HGA_CDM (Bu et al., 2022)	FrcSub (DeCarlo, 2011), Math1, Math2 (Liu et al., 2018), CSEDm_spring ^d	Standard procedure	N/A	AUC, ACC, Recall, F1	AUC = 0.894, 0.795, 0.837, 0.74 F1 = 0.887, 0.774, 0.827, 0.659; for FrcSub, Math1, Math2, CSEDm_spring
CQKT (Wang, Ma et al., 2022)	ASSIST0910, Intellilence18 (Wang, Ma et al., 2022), BridgeAlgebra2006 ^e	Student and Concept Filtering	MLP, LSTM	AUC, RMSE	AUC = 0.763, 0.812, 0.783; RMSE = 0.352, 0.270, 0.284; for ASSISTment09, Intellilence18, BridgeAlgebra2006
LDM-HMI (Wang, Yan et al., 2023)	CL21 (Wang, Yan et al., 2023), Math1 (Wu et al., 2015), Synthetic-5 (Wang, Yan et al., 2023)	Standard procedure	CNN, MLP	AUC, RMSE	AUC = 0.859, 0.819, 0.902; RMSE = 0.412, 0.416, 0.353; for CL21, Math1, Synthetic-5
KIEDLKD (Gan, Sun, & Sun, 2022)	Algebra0506, Statics2011, Assist0910, Assist1213, Bridge2Algebra (Gan et al., 2022)	Standard procedure	MLP, Memory network	AUC, ACC	AUC = 0.851, 0.827, 0.831, 0.797, 0.734 for Algebra0506, Statics2011, Assist0910, Assist1213, Bridge2Algebra
Evidential-CDM (Liu, Hou, Zhang, Liu and He, 2023)	ASSIST (Feng, Heffernan, & Koedinger, 2009), UncASSIST (Liu, Hou et al., 2023)	Standard procedure	MLP	AUC, RMSE	AUC = 0.735, 0.737; RMSE = 0.443, 0.440 for UncASSIST and ASSIST
MRCD (Wu et al., 2023)	ASSIST0910, EdNet ^f	Exercise and Student filtering	GCN, MLP	AUC, ACC, RMSE	AUC = 0.782, 0.770; ACC = 0.741, 0.737; RMSE = 0.419, 0.420; for ASSIST0910, EdNet
QRCDM (Yang et al., 2022)	FrcSub, Math1, Math2, ASSIST0910, ASSIST17 ^g	Exercise and Student filtering	MLP	ACC, AUC, RMSE	ACC = 0.84, 0.70, 0.72, 0.74, 0.73; AUC = 0.91, 0.79, 0.80, 0.78, 0.80; RMSE = 0.35, 0.43, 0.43, 0.42, 0.42; for FrcSub, Math1, Math2, ASSIST0910, ASSIST17
HierCDF (Li, Wang et al., 2022)	JunYi, MATH2021	Exercise and Student Filtering	MLP	AUC, ACC, F1, RMSE	AUC = 0.784, 0.736; ACC = 0.749, 0.701; F1 = 0.825, 0.809; RMSE = 0.415, 0.434; for JunYi, MATH2021
RCD (Gao et al., 2021)	ASSIST0910, JunYi	Exercise and Student filtering	GCN, MLP	ACC, AUC, RMSE	ACC = 0.736, 0.772; AUC = 0.772, 0.826; RMSE = 0.421, 0.396; for ASSIST, JunYi

(continued on next page)

attention and text attention based on the learned affinity matrix. For example, Guo et al. (2020) proposed a cross attention mechanism for VQA, which focuses on the interactions between each question word

and each visual object in the image. Song et al. (2022) proposed a new cross-media grouping co-attention mechanism, which capture fine-grained semantics by the dense interactions between image regions and

Table 1 (continued).

Works	Dataset	Preprocessing	DL	Validation Criteria	Performance
ICD (Qi et al., 2023)	ASSIST0910 ^h , ASSIST2017 ⁱ , JunYi (Chang, Hsu, & Chen, 2015), MathEC ^j	Student and Answer filtering	MLP	ACC, AUC, RMSE	ACC = 0.746, 0.729, 0.820, 0.741; AUC = 0.784, 0.802, 0.828, 0.802; RMSE = 0.415, 0.423, 0.359, 0.415; for ASSIST0910, ASSIST2017, JunYi, MathCE
IRR (Tong et al., 2021)	ASSIST0910, MATH	Student filtering	MLP	AUC, Precision, Recall, F1	AUC = 0.839, 0.608; Precision = 0.734, 0.737; Recall = 0.579, 0.719; F1 = 0.626, 0.724; for ASSIST, MATH
NeuralCD (Wang, Liu et al., 2022)	Math (Wang, Liu et al., 2022), ASSIST0910	Standard procedure, Student filtering	MLP	ACC, AUC, RMSE	ACC = 0.792, 0.719; AUC = 0.820, 0.749; RMSE = 0.378, 0.439; for Math, ASSIST
CWNCD (Wang, Fu et al., 2022)	ASSIST0910	Standard procedure, Duplicated data elimination	MLP	ACC, AUC, RMSE	ACC = 0.889, AUC = 0.762, RMSE = 0.302
Graph-EKLN (Liu, Shao et al., 2021)	ASISST0910, KDDcup ^k	Standard procedure, Exercise filtering	GCN	ACC, AUC, RMSE	ACC = 0.778, 0.827; AUC = 0.830, 0.829; RMSE = 0.394, 0.359; for ASSIST, KDDcup
IncreCD (Tong et al., 2022)	ASSIST0910, MATH ^l	Standard procedure, Student filtering	MLP	ACC, AUC	ACC = 0.724, 0.797; AUC = 0.758, 0.792; for ASSIST, MATH
CDMFKC (Li, He et al., 2022)	ASSIST0910, Math1, Math2 (Wu et al., 2015)	Standard procedure	MLP	ACC, AUC, RMSE	ACC = 0.762, 0.740, 0.732; AUC = 0.750, 0.797, 0.778; RMSE = 0.431, 0.437, 0.442; for ASSIST, Math1, Math2
DCD (Wang, Huang et al., 2023)	ASSIST2009, ASSIST2012 ^m , KDDCup ⁿ	Exercise and Student filtering, Student division	Memory network, RNN, MLP	AUC, ACC	AUC = 0.784, 0.778, 0.811; ACC = 0.747, 0.760, 0.853; for ASSIST2009, ASSIST2012, KDDCup
SPP-NCD (Ma et al., 2023)	FrcSub, Math1 and Math2 ^o , ASSIST0910	Student filtering	N/A	ACC, RMSE	ACC = 0.882, 0.679, 0.691, 0.696; RMSE = 0.311, 0.373, 0.378, 0.464; for FrcSub, Math1, Math2, ASSIST
KSCD (Ma et al., 2022)	JunYi, e-Math (Ma et al., 2022)	Standard procedure, Student filtering	MLP	ACC, AUC, RMSE	ACC = 0.778, 0.714; AUC = 0.822, 0.769; RMSE = 0.391, 0.429; for JunYi, e-Math
NeuralNCD (Li, Hu et al., 2022)	FrcSub, ASSIST0910	Standard procedure, Student filtering	MLP	ACC, AUC, RMSE	ACC = 0.836, 0.734; AUC = 0.905, 0.776; RMSE = 0.364, 0.425; for FrcSub, ASSIST
ECD (Zhou et al., 2021)	Asia, Europe, America (Zhou et al., 2021)	Standard procedure, Student filtering	MLP	ACC, AUC, RMSE	ACC = 0.677, 0.700, 0.699; AUC = 0.745, 0.770, 0.764; RMSE = 0.468, 0.443, 0.445; for Asia, Europe, America

(continued on next page)

textual words. By various attention mechanisms, these methods can learn cross-modal features representing fine-grained semantics of cross-modal information. However, most of these co-attention mechanisms are designed for VQA tasks, which estimate the correctness of candidate answers by evaluating the relationships between answers and the cross-modal contents of visual questions. As answers usually are only related

to partial contents of visual questions, these works reason the co-attention based on an affinity matrix between local image regions and words.

However, our CMNCD needs to estimate the difficulty of exercises on each of the concepts being examined. In general, the exercise difficulty requires the consideration of the overall semantic contents

Table 1 (continued).

Works	Dataset	Preprocessing	DL	Validation Criteria	Performance
CMNCD	CMTD, CBTD, JunYi	Standard Procedure, Exercise filtering	CNN, MLP	ACC,AUC, RMSE,F1, Sensitivity, Specificity, Precision	ACC = 0.790,0.769, 0.780; AUC = 0.844,0.841, 0.838; RMSE = 0.394,0.395, 0.401; for CBTD,CMTD,JunYi.

^a<https://sites.google.com/site/assistentmsdata/home/assistent-2009-2010-data/skill-builder-data-2009-2010>.

^b<https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp>.

^c<http://www.zhixue.com>.

^d<https://pslcdatashop.web.cmu.edu/Files?datasetId=3458>.

^e<https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp>.

^f<http://ednet-leaderboard.s3-website-ap-northeast-1.amazonaws.com/>.

^g<https://sites.google.com/view/assistentmsdatamining/data-mining-competition-2017>.

^h<https://sites.google.com/site/assistentmsdata/home/2009-2010-assistent-data/skill-builder-data-2009-2010>.

ⁱ<https://sites.google.com/view/assistentmsdatamining/data-mining-competition-2017>.

^j<https://eedi.com/projects/neurips-education-challenge>.

^k<https://pslcdatashop.web.cmu.edu/KDDCup/>.

^l<https://www.zhixue.com/>.

^m<https://sites.google.com/site/assistentmsdata/datasets/2012-13-school-data-with-affect>.

ⁿ<https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp>.

^o<http://staff.ustc.edu.cn/~qiliuql/data/math2015.rar>.

of image-question pairs, and thus should focus on diverse global interactions between questions and images. Therefore, our CMNCD first learns multiple semantic representations for both images and questions, and then computes the affinity matrix by evaluating the similarities between these representations of different modalities. Besides, our CMNCD can adaptively select relevant cross-modal information specific to the concepts examined by exercises, which helps to enhance the learning of exercises' trait vectors.

3. Preliminary

In this part, we introduce the task of cognitive diagnosis based on cross-modal exercises in this paper and briefly review the classic cognitive diagnosis model MIRT.

3.1. Notation and problem definition

Notation: Suppose there are N students, J cross-modal exercises and K concepts. The response of the n th ($1 \leq n \leq N$) student got on the j th ($1 \leq j \leq J$) exercise is denoted by r_{nj} . The j th cross-modal exercise can be represented by (T_j, I_j) , which consists of an image I_j and the accompanied question text T_j . In classic CDMs, the relations between the J exercises and the K concepts are usually manually labeled and denoted by a binary Q-matrix $(q_{jk})_{J \times K}$, where $q_{jk} = 1$ if the j th exercise examine the k th concept and $q_{jk} = 0$ otherwise.

Problem Definition: Given student performance records $R = \{r_{nj}\}$ on cross-modal exercises, the goal of cognitive diagnosis is to infer the proficiency vector $\theta_n \in \mathbb{R}^K$ of each student on all the K examined concepts, through the student performance prediction process.

Note that the groundtruth for diagnosis results are unavailable in general. Previous works (Wu et al., 2015) demonstrate that the more accurate the prediction of student performance, the better the diagnosis results. Thus, following conventional cognitive diagnosis methods (Cheng & Liu, 2019; Wang, Liu et al., 2022), we estimate the accuracy of cognitive diagnosis results by the task of student performance prediction.

3.2. MIRT

MIRT extends the scalar traits of students and exercises used in IRT to multi-dimensional latent vectors. With these multidimensional trait

vectors, MIRT predicts the probability that the n th student answers the question correctly by the following logistic-like formula:

$$p(\theta_n, \mathbf{a}, \mathbf{b}) = \frac{e^{\mathbf{a}^T(\theta_n - \mathbf{b})}}{1 + e^{\mathbf{a}^T(\theta_n - \mathbf{b})}}, \quad (1)$$

where θ_n denotes the trait vector of the n th student, and each of its dimension indicates the degree the n th student masters one latent knowledge item. The \mathbf{b} and \mathbf{a} denote the trait vectors of exercise difficulty and discrimination, respectively.

4. Deep cross-modal neural cognitive diagnosis

In this part, we introduce a deep neural cognitive diagnosis framework for cross-modal exercises. The framework of the proposed CMNCD is shown in Fig. 2, which takes cross-modal exercises and Q-matrix as inputs, and student performance as outputs.

Specifically, given a cross-modal exercise, Module I incorporates both the cross-modal contents of exercises and the complexity of concepts output by Module II into the difficulty trait learning of exercises. In Module II, the complexity of concepts can be derived from the prerequisite relations among concepts. Module III aims to learn discrimination trait vectors of exercises, and Module IV aims to learn students' trait vector, i.e., the knowledge proficiency vector. At last, the trait vectors of exercises and students are fed into a deep MIRT module for predicting student performance. The algorithm and all the parameters of our CMNCD is summarized in Tables 2 and 3, respectively

4.1. Exercise trait: Difficulty and discrimination

In this part, we introduce how to estimate the trait vectors of exercises based on the cross-modal contents of exercises and the complexity of concepts.

4.1.1. Module I: Cross-modal content based difficulty trait learning

In this module, we first learn cross-modal features of exercises by a CA-GCF mechanism, and then estimate exercise traits of difficulty based on the learned cross-modal features.

Step 1: CA-GCF: In Fig. 3, we show the framework of the Co-Attention based Gated Cross-modal Fusion (CA-GCF) mechanism. The CA-GCF aims to learn cross-modal exercise features by a gated cross-modal feature fusion strategy, where both visual attention and textual attention are jointly reasoned via a co-attention mechanism. Specifically, given a cross-modal exercise, CA-GCF first jointly learns visual

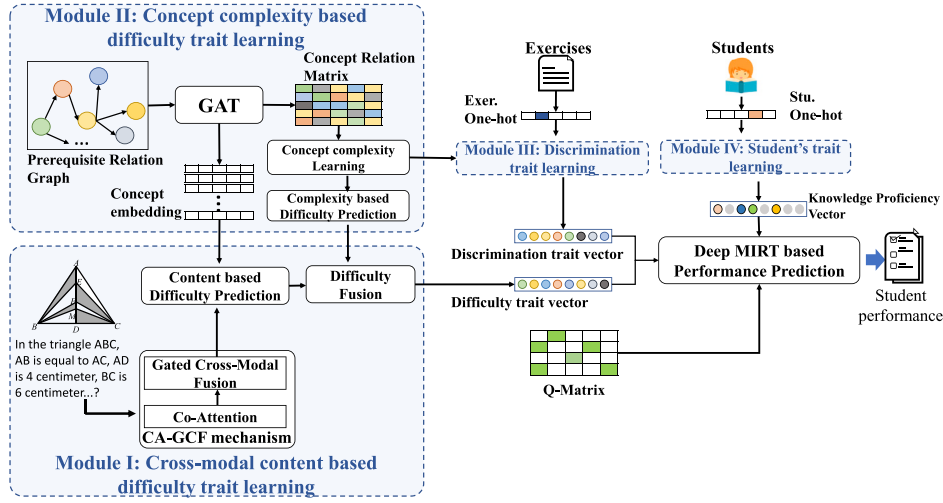


Fig. 2. The framework of the proposed CMNCD, which takes cross-modal exercises and relevant Q-matrix as inputs, and the performance of students on the exercises as outputs.

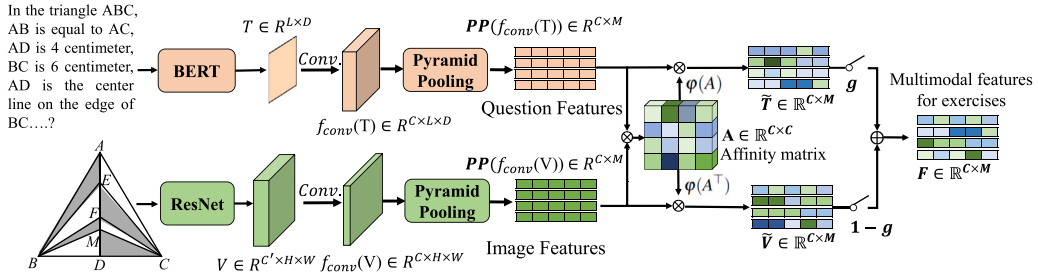


Fig. 3. The framework of CA-GCF.

Table 2

Algorithm description for our proposed CMNCD.

Algorithm 1 CMNCD

Input: Cross-modal exercise corpus $\{(V_j, T_j, y_{n_j})\}_{j=1}^J$, $Q \in \mathbb{R}^{J \times K}$, learnable knowledge proficiency θ_n , prerequisite concept graph G_{pcg} , one-hot representations for exercises x_j , learning rate, epoch number E .

Output: Student performance predictions $\{p_{n_j}\}_{n=1}^N$, θ_n .

- 1: Initialize the parameters θ of CMNCD and θ_n ;
- 2: **for** Epoch $e = 1$ to E **do**
- 3: **for** Cross-modal exercises $j = 1$ to J **do**
- 4: $f_j^c = \text{Crossmodal_Content_based_Difficulty_Trait_Learning}(V_j, T_j)$;
- 5: $(f_j^p, h) = \text{Concept_Complexity_based_Difficulty_Trait_Learning}(G_{pcg})$;
- 6: $f_j^{diff} = \text{Exercise_Difficulty_Trait_Fusion}(f_j^c, f_j^p)$;
- 7: $f_j^{disc} = \text{Exercises_Discrimination_Trait_Learning}(Q, x_j, h)$;
- 8: $\theta_n = \text{Student_Trait_Learning}(x_n)$;
- 9: $p_{n_j} = \text{Deep_MIRT_based_Performance_Prediction}(Q, x_j, f_j^{diff}, f_j^{disc}, \theta_n)$;
- 10: **end for**;
- 11: **end for**.

and textual features via a co-attention mechanism. Then, a cross-modal exercise representation is learned by fusing the textual and visual features via a gated cross-modal fusion strategy.

Specifically, as shown in the figure, we first separately feed the question text and the image in the exercise into the BERT (Devlin, Chang, Lee, & Toutanova, 2019) and ResNet (He, Zhang, Ren, & Sun, 2016) to learn the word embedding matrix $T \in \mathbb{R}^{L \times D}$ for the question and visual feature map $V \in \mathbb{R}^{C \times H \times W}$ for the image. L denotes the length of the question and D denotes the dimension of word embeddings. Then, the learned T and V are separately fed into a convolutional layer followed by one pyramid pooling layer (He, Zhang, Ren, & Sun, 2015) (denoted by $PP(f_{conv}(T))$ and $PP(f_{conv}(V))$), which aims to learn textual and visual features for cross-modal exercises. Finally, with the outputs of

the pyramid pooling layers, we learn the co-attention between images and texts based on their affinities, which can be computed by

$$A = PP(f_{conv}(V)) \cdot PP(f_{conv}(T))^T, \quad (2)$$

where $A \in \mathbb{R}^{C \times C}$ denotes the text-image affinity matrix and its elements denote the affinities between textual and visual semantic representations. The $f_{conv}(\cdot)$ denotes the operation of applying C convolutions with kernel size of 1×1 to image feature map V or word embedding matrix T . Specifically, the operation $f_{conv}(V) \in \mathbb{R}^{C \times H \times W}$ shrinks the channels of image feature maps to reduce redundant visual information, where each channel conveys specific visual semantics (Zeiler & Fergus, 2014). Similar to the multi-head projection (Vaswani et al., 2017), $f_{conv}(T) \in \mathbb{R}^{C \times L \times D}$ performs feature linear projection multiple

Table 3

A recap of all the parameters in our CMNCD.

Parameters	Definitions
N	Number of students.
J	Number of cross-modal exercises.
K	Number of knowledge concepts.
$(q_{jk})_{J \times K}$	A Binary Q-matrix indicating the relations between J exercises and K knowledge concepts.
A	The affinity matrix between the image and the question.
V	Image feature map.
T	The feature matrix consisting of question word encodings.
D	The dimension of word encoding.
L	Length of the question text.
C'	The initial channel number of V
$f_{conv}(\cdot)$	The convolution operations applied to V and T .
C	The channel number of the V processed by $f_{conv}(\cdot)$.
$PP(\cdot)$	The pyramid pooling operation.
M	The number of the spatial bins in $PP(\cdot)$.
A_{nr}	The Attention weight matrix of question features over image features.
A_{nc}	The Attention weight matrix of image features over question features.
\tilde{V}	The updated image feature map under the guidance of questions.
\tilde{T}	The updated question feature matrix under the guidance of images.
g	The cross-modal gate.
F_j	The feature matrix of the j th exercise obtained by the cross-modal feature fusion.
f_j^c	The difficulty trait vector learned from the contents of the j th exercise.
f_{jk}^c	The element of f_j^c that denotes the difficulty of the j th on the k th concept.
K_j	The number of the concepts examined by the j th exercise.
U_j	The cross-modal feature matrix for the j th exercise related to knowledge concepts.
W_j	The relation matrix between knowledge concepts and cross-modal features.
p	The concept encoding.
σ	The sigmoid function.
h	The concept complexity vector.
$s(n, D)$	The importance of the concept n for learning the concept D .
$r(i, j)$	The relevance score between the concepts i and j .
f^p	The exercises' difficulty trait vector learned from the concept complexity.
f_j^{diff}	The difficulty trait vector for the j th exercise obtained by fusing f_j^c and f^p .
f_j^{disc}	The discrimination trait vector for the j th exercise.
x_j	The one-hot representation for the j th exercise.
x_n	The one-hot representation for the n th student.
θ_n	The knowledge proficiency vector.
p_{nj}	The predicted performance score for the n th student on the j th exercise.
y_{nj}	The groundtruth score for the n th student on the j th exercise.
$W_g, W_1, W_2, W_p, W_a, W_b, W_c, W_h, S, W_s$	Learnable parameter matrices.

times by different parameter matrices, where each projection masks out different textual information and maintains specific semantics. That is, each channel of the feature maps (i.e., $f_{conv}(V)$ or $f_{conv}(T)$) can be viewed as features representing specific visual or textual semantics.

The pyramid pooling (denoted by $PP(\cdot)$) aims to transform $f_{conv}(V)$ and $f_{conv}(T)$ into the same dimensional space before evaluate the affinities between each image-question pair. The pyramid pooling is a spatial feature sampling strategy that pools the semantic statistics of feature maps along the channel dimension using spatial bins, dividing feature maps into fixed-size bins. In each spatial bin, $PP(\cdot)$ takes the strategy of max pooling on the visual or textual information within the bin. Specifically, in Eq. (2), the size of feature matrix output by the $PP(\cdot)$ is $C \times M$, where C and M denote the numbers of the channels and spatial bins, respectively. That is, we can obtain C visual features of M dimensions, and C question textual features of M dimensions by the $PP(\cdot)$.

With the learned affinity matrix A , we learn text-image co-attention by applying row-wise and column-wise normalization to A , which can be formulated by

$$A_{nr} = \varphi(A), \quad (3)$$

$$A_{nc} = \varphi(A^T), \quad (4)$$

where φ denotes the softmax function. The $A_{nr} \in \mathbb{R}^{C \times C}$ denotes the attention weights of each textual feature vector on the C visual feature vectors. The $A_{nc} \in \mathbb{R}^{C \times C}$ denotes the attention weights of each visual feature vector on the C textual feature vectors. With A_{nr} and A_{nc} , we can update the image and textual feature maps by

$$\tilde{V} = A_{nr} \cdot PP(f_{conv}(V)) \in \mathbb{R}^{C \times M}, \quad (5)$$

$$\tilde{T} = A_{nc} \cdot PP(f_{conv}(T)) \in \mathbb{R}^{C \times M}, \quad (6)$$

where \tilde{V} denotes the question-guided image feature map and \tilde{T} denotes the image-guided textual feature matrix.

Gated cross-modal fusion: The learned \tilde{V} and \tilde{T} are fused to learn cross-modal representations for exercises by a gated cross-modal fusion strategy, which can be formulated by

$$g = \sigma(W_g(f_i(\tilde{V}) \oplus f_i(\tilde{T}))), \quad (7)$$

$$F = g \cdot \tilde{V} + (1 - g) \cdot \tilde{T}, \quad (8)$$

where the scalar g denotes the cross-modal gate, W_g denotes a learnable parameter matrix, the function $f_i(\cdot)$ flattens a matrix into a vector, and

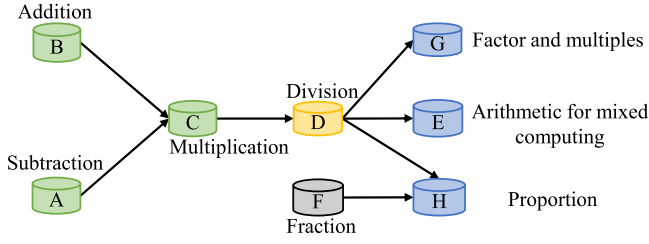


Fig. 4. Examples of the prerequisite concept maps. The prerequisite relations between concepts indicate the learning dependencies between concepts.

the symbol \oplus denotes element-wise summation of two vectors. The $F \in \mathbb{R}^{C \times M}$ denotes a multi-modal feature matrix, each row vector of which is a multi-modal feature representing specific implicit semantics.

Step 2: Content based difficulty prediction: In this step, we use the learned F to estimate the difficulty trait vector of the j th cross-modal exercise, denoted by $f_j^c = (f_{j1}^c, f_{j2}^c, \dots, f_{jK_j}^c)$. The $f_{jk}^c \in (0, 1)$ denotes the difficulty on the k th concept, and K_j denotes the number of the concepts examined by the j th exercise. To compute f_j^c , we first learn concept-specific cross-modal features by

$$U_j = W_j F_j, \quad (9)$$

where $U_j = (u_{jk})_{k=1}^{K_j} \in \mathbb{R}^{K_j \times M}$ denotes a multi-modal feature matrix and each of its M -dimension row vector (i.e., u_{jk}) represents the multi-modal exercise feature specific to the k th concept. The $F_j \in \mathbb{R}^{C \times M}$ denotes the multi-modal feature matrix of the j th cross-modal exercise. The $W_j \in \mathbb{R}^{K_j \times C}$ denotes a relation matrix representing the relations between the K_j concepts and the C cross-modal features of F_j . The computation of W_j can be formulated by

$$W_j = P_j W_p F_j^T, \quad (10)$$

where $P_j \in \mathbb{R}^{K_j \times D}$ denotes the feature matrix of the K_j concepts, and D denotes the dimensionality of concept encodings. The $W_p \in \mathbb{R}^{D \times M}$ is a learnable parameter matrix that projects concept encodings into a M -dimension space, which has the same dimensionality as that of the multi-modal features of F_j .

Finally, the cross-modal feature of exercises specific to the k th concept, denoted by u_{jk} , is fed into a Multi-Layer Perceptron (MLP) network for learning exercises' difficulty trait by

$$f_{jk}^c = \sigma(W_2^T(W_1 u_{jk}^T)), \quad (11)$$

where f_{jk}^c denotes the exercises' difficulty trait with respect to the k th examined concept, and σ represents a sigmoid function. The $W_1 \in \mathbb{R}^{64 \times M}$ and $W_2 \in \mathbb{R}^{64 \times 1}$ are two parameter matrices of the MLP. Besides these examined concepts, the difficulty of the j th exercise on the remained concepts is set to 0. At last, we obtain the estimated difficulty trait vector of the j th exercise by cross-modal contents, denoted by $f_j^c = (f_{j1}^c, f_{j2}^c, \dots, f_{jK_j}^c) \in (0, 1)^{K_j}$.

4.1.2. Module II: Concept complexity based difficulty trait learning

In addition to the cross-modal contents of exercises, the complexity of the concepts being examined also have an impact on exercises' trait vectors. For example, the more complex of the concepts examined by exercises, the more difficult the exercises. In this part, we describe how to model the complexity of concepts based on the prerequisite relations among concepts.

Specifically, given the prerequisite relation graph in Fig. 4, we denote the complexity of the concept D as h_D , which can be learned by

$$h_D = \frac{\sum_{n \in V_D^-} s(n, D)}{\sum_{n \in V_D^-} s(n, D) + \sum_{n \in V_D^+} s(D, n)}, \quad (12)$$

where $V_D^- = \{A, B, C\}$ denotes the set of the prerequisite concepts of the concept D , and $V_D^+ = \{G, E, H\}$ denotes the set of the concepts that taking the concept D as their prerequisite concept.

The score $s(n, D)$ denotes the importance of the concept n for learning the concept D . Obviously, the farther the distance between the concepts n and D , the smaller the importance of the concept n for learning the concept D . Considering both the semantic information of concepts and the prerequisite structure among concepts, we learn $s(n, D)$ by

$$s(n, D) = \prod_{i, j \in \mathcal{V}_{p(n, D)}, i \neq j} r(i, j), \quad (13)$$

where $p(n, D)$ denotes the shortest directed acyclic path from n to D , $\mathcal{V}_{p(n, D)}$ denotes the set of the concepts on the path $p(n, D)$, and j denotes the adjacent concept of i that takes i as its prerequisite concept on the $p(n, D)$. The $r(i, j)$ represents the relevance score between the concepts i and j .

For example, if the concept A in Fig. 4 is selected as the concept n , then $\mathcal{V}_{p(A, D)} = \{A, C, D\}$. In this case, we have $i = A, j = C$ and $i = C, j = D$. Therefore, $s(A, D)$ can be computed by

$$s(A, D) = r(A, C) \times r(C, D). \quad (14)$$

Specifically, $r(i, j)$ is learned by the masked-attention mechanism used in the Graph Attention neTwork (GAT) (Veličković et al., 2018), which only computes the relevance score between two linked nodes in a graph. The computation of $r(i, j)$ can be formulated by

$$r(i, j) = f_r(W_a(W_c p_i \parallel W_c p_j)), \quad (15)$$

where $f_r(\cdot)$ denotes the LeakyReLU function, \parallel is the concatenation operation, $p_i \in \mathbb{R}^F$ and $p_j \in \mathbb{R}^F$ denote the embeddings of the concepts i and j , respectively. The $W_c \in \mathbb{R}^{F' \times F}$ and $W_a \in \mathbb{R}^{2F'}$ are two parameter matrices.

By the Eqs. (12), (13) and (15), we can obtain the complexity of all the K concepts, denoted by $h \in (0, 1)^K$. Then, we use the learned concept complexity to estimate the difficulty trait vectors of exercises by

$$f^p = \sigma(W_b(W_h \times h)), \quad (16)$$

where σ denotes the sigmoid function, and $f^p \in (0, 1)^K$ denotes the concept complexity based difficulty trait vector. The W_h and W_b are two parameter matrices. For each element of W_h and W_b , we restrict their values to be positive, which makes exercise trait of difficulty to monotonically increase with an increment of the complexity of the concepts being examined.

Difficulty fusion: For the j th cross-modal exercise, we learn its difficulty trait vector over all the concepts by fusing f^p with f_j^c , which can be formulated by

$$f_j^{diff} = \sigma(f_j^c + f^p) \in (0, 1)^K, \quad (17)$$

where the function $\sigma(\cdot)$ represents the operation of applying the sigmoid to each dimension of the input vector.

4.1.3. Module III: Exercises' discrimination trait learning

Exercises' discrimination trait refers to the ability of exercises in differentiating the students with different proficiency on concepts. In general, the more prerequisites of a concept has, the number of the students who master the concept is smaller. That is, the discrimination ability of exercises is related to the prerequisite information of the concepts being examined.

Therefore, in this part we estimate the discrimination trait vectors of exercises based on the complexity of examined concepts, which can be formulated by

$$f_j^{disc} = \sigma(W_s((x_j \times Q) \circ h)^T), \quad (18)$$

where $f_j^{disc} \in (0, 1)^K$ denotes the estimated discrimination trait vector for the j th cross-modal exercise, $\sigma(\cdot)$ denotes the sigmoid function,

and $\mathbf{h} \in \mathbb{R}^K$ represents the complexity vector of concepts that can be learned by Module II. The $\mathbf{W}_s \in \mathbb{R}^{K \times K}$ denotes a parameter matrix and its elements are also restricted to be positive. $\mathbf{x}_j \in \mathbb{R}^{1 \times J}$ denotes the one-hot representation of the j th exercise. The $\mathbf{Q} \in \mathbb{R}^{J \times K}$ denotes the Q-matrix, which indicates the relationships between J exercises and K concepts. The symbol \circ denotes the operation of element-wise production.

4.2. Module IV: Student trait learning

In CMNCN, the trait vector of the n th student, i.e., knowledge proficiency vector $\theta_n = \{\theta_{n1}, \theta_{n2}, \dots, \theta_{nK}\}$, can be obtained by

$$\theta_n = \sigma(\mathbf{x}_n \times \mathbf{S}), \quad (19)$$

where the function $\sigma(y)$ represents the operation of applying sigmoid on each dimension of the vector y , and θ_{nk} denotes proficiency level that the student masters the k th concept. The $\mathbf{x}_n \in \mathbb{R}^{1 \times N}$ represents student's one-hot representation vector and $\mathbf{S} \in \mathbb{R}^{N \times K}$ represents a trainable parameter matrix.

4.3. Deep MIRT based performance prediction

In this part, we predict the student performance on each cross-modal exercise by a deep MIRT module, which preserves the characteristics of the interactions between students' trait vectors and exercises' trait vectors in MIRT. Specifically, our deep MIRT predicts the probability of the n th student correctly answering the j th exercise by

$$p_{nj} = MLP(\mathbf{Q}_j \circ \mathbf{f}_j^{disc} \circ (\theta_n - \mathbf{f}_j^{diff})), 1 \leq n \leq N, \quad (20)$$

where p_{nj} denotes the predicted probability, $\mathbf{Q}_j = \mathbf{x}_j \times \mathbf{Q}$ indicates the concepts examined by the j th exercise. The employed MLP is a 4-layer neural network with K units on the input layer, 256 units on the first hidden layer, 64 units on the second hidden layer, and 1 unit on the output layer. All hidden layers use ReLUs and the output layer uses sigmoid as their activation functions. To regularize the network, a dropout of 0.4 is used with the hidden layers. Following Wang, Liu et al. (2022), we restrict the parameters of the MLP to be positive, which aims to ensure the p_{nj} is monotonically increasing at any dimension of θ_n .

The loss function of CMNCD is sigmoid cross entropy between the p_{nj} and the corresponding groundtruth y_{nj} , which can be represented by

$$loss = - \sum_n \sum_j (y_{nj} \log p_{nj} + (1 - y_{nj}) \log(1 - p_{nj})) \quad (21)$$

By directly minimizing the loss using Adam optimization, the trait vectors of students and exercises, as well as the parameters of the neural networks in CMNCD are optimized. Finally, after training, the vector θ_n is what we get as the diagnosis result, denoting student knowledge proficiency over all examined concepts.

5. Datasets and evaluation criteria

5.1. Datasets

To the best of our knowledge, there is no public available cross-modal exercise datasets for cognitive diagnosis. To validate the performance of CMNCD, we constructed the Cross-modal Math Test Data (CMTD) and Cross-modal Biology Test Data (CBTD) for cognitive diagnosis. All the exercise data and students' performance records are collected from the Advanced Innovation Center for Future Education.¹ In addition, we also validate the effectiveness of CMNCD on a popular cognitive diagnosis dataset, i.e., Junyi (Chang et al., 2015). We summarize the statistics of the above three real-world datasets in Table 4 and present their detailed descriptions as follows.

Table 4

The statistics of three cross-modal exercise datasets for cognitive diagnosis.

Dataset	CMTD	CBTD	Junyi
Students	2601	513	10,000
Exercises	336	194	835
Knowledge concepts	37	14	835
Response records	27,262	6990	353,835
Average concepts per exercise	1	1	1
Average Response records per student	10.48	13.62	35.38
Prerequisite relations between concepts	52	17	988

CMTD dataset: This dataset contains 27,262 performance records of 2601 students on 336 cross-modal exercises for primary school math test, where each exercise is comprised of an image-question pair. Solving these exercises requires students to master 37 concepts. There are 52 prerequisite relations between these concepts. The Q-matrix that indicates the relationships between exercises and concepts are labeled by domain experts.

CBTD dataset: This dataset contains 6990 performance records of 513 students on 194 cross-modal exercises for junior high school biology test. Solving these exercises requires students to master 14 concepts. Like CMTD, the prerequisite relations and the Q-matrix are also labeled by domain experts.

Junyi dataset²: This dataset contains 353,835 performance records of 10,000 students on 835 exercises, where each exercise only contains one textual question. These exercises are crawled from the math test database of a Chinese e-learning platform. To answer these exercises, students are required to master 835 concepts. There are 988 prerequisite relations between these concepts.

5.2. Evaluation metric

As we cannot obtain the true knowledge proficiency of students, it is difficult to evaluate the performance of Cognitive Diagnosis Models (CDM) directly. Following previous cognitive diagnosis works, we validate and compare the performance of different models indirectly through the results of students' performance prediction from two perspectives: regression and classification. Specifically, from the regression perspective, we follow previous CDM (Cheng et al., 2019; Wang, Liu et al., 2022) to quantify the distance between the predicted results and groundtruth by the Root Mean Square Error (RMSE) (Pei, Yang, Liu, & Dong, 2018). The smaller the RMSE values are, the better the prediction results are.

Treating the performance prediction as a classification task, where a student answer record with score 1 (0) indicates a correct (wrong) answer, we compare the prediction results of different cognitive diagnosis methods by the ACCuracy (ACC) and the AUC (Bradley, 1997). The AUC is the Area Under an ROC Curve, which shows the tradeoff between sensitivity and specificity for all possible thresholds rather than just the one that was chosen. This is, the ROC curve helps to validate the performance of our model independent of the choice of a threshold.

The AUC and ACC have been extensively used as the classification metrics in previous cognitive diagnosis models (Cheng et al., 2019; Wang, Liu et al., 2022), and the larger their values are, the better the results are. Besides, we also further validate the performance of our model by measuring some statistical metrics, including *Precision*, *Sensitivity*, *Specificity* and *F1 score*. To be more specific, *Precision* indicates the classifier's ability to not mark a negative sample as positive. *Sensitivity (Recall)* indicates classifier's ability to classify all positive samples correctly (i.e., true positive rate). *Specificity* indicates the classifier's ability to correctly identify all negative samples (i.e., true negative rate). *F1-score* is the harmonic mean of *Precision* and *Recall*.

¹ <https://aic-fe.bnu.edu.cn/en/>.

² <https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198>.

6. Experiments

To validate the performance of our CMNCD, we conduct extensive experiments to address the following Research Questions (RQ), including:

- RQ1: Could CMNCD achieve state-of-the-art performance on the task of cognitive diagnosis based on cross-modal exercises containing images and texts?
- RQ2: Is it necessary to exploit the information of the images or texts in exercises for learning the trait vectors of exercises?
- RQ3: How about the effectiveness of the Co-Attention based Gated Cross-modal Fusion (CA-GCF) mechanism?
- RQ4: How about the effect of concept complexity on result prediction?
- RQ5: How about the effectiveness of the deep MIRT module on the prediction of student performance?

6.1. Implementation details

In CMDCD, for each image, we extract its visual feature map $V \in \mathbb{R}^{H \times W \times C}$ from the fourth basic convolutional block in ResNet152 (He et al., 2016), where $H = W = 14$ and $C = 1024$. We truncate the texts in exercises with a maximum of 80 words. For each word, its 768 dimensional embedding vector is learned by BERT-base (Turc, Chang, Lee, & Toutanova, 2019) and thus the feature matrix of texts $T \in \mathbb{R}^{1 \times 80 \times 768}$. Before the pyramid pooling, we apply $C = 512$ convolutions with kernel size of 1×1 to image feature V and text feature T . In the pyramid pooling, the pyramid is $\{1 \times 1, 2 \times 2, 3 \times 3, 6 \times 6\}$, which totally contains 50 bins, i.e., $M = 50$. Thus, after the pyramid pooling, V and T are both transformed into the space of 512×50 . For the prerequisite relation graph, the embedding of each concept node is learned by the GAT, which outputs a 2048-D vector for each node, i.e., concept embeddings $c \in \mathbb{R}^{2048 \times 1}$.

During the training process, the parameters of CMNCD are initialized by Xavier (Glorot & Bengio, 2010) and an Adam optimizer is used to train the parameters. The batch sizes on CMTD, CBTD and Junyi are set to 64, 32 and 128, respectively. Our method is implemented by PyTorch using Python, and all experiments are run on a Linux server with four 3.1 GHz Intel 6254 CPUs and 4 RTX 2080Ti GPU.

6.2. Baseline methods

To demonstrate the effectiveness of CMNCD, we compare it with following classic and state-of-the-art cognitive diagnosis methods.

IRT (Lord, 2012): a classical cognitive diagnosis method that uses item response function to simulate the interaction between students and exercises.

MIRT (Yao & Schwarz, 2006): a multidimensional cognitive diagnosis method that can diagnosis multiple knowledge proficiency of students and the latent items of exercises.

PMF (Mnih & Salakhutdinov, 2008): a Probabilistic matrix factorization (PMF) method that factors score matrix to get students' and exercises' latent trait vectors.

NeuralCD (Wang, Liu et al., 2022): a deep learning based cognitive diagnosis method that incorporates deep neural networks to learn the traits of students and exercises, as well as the complex exercising interactions between students and exercises.

RCD (Gao et al., 2021): a Relation map driven Cognitive Diagnosis (RCD) framework that improves the diagnosis results using a multi-layer student-exercise-concept relation map. In RCD, the prerequisite relations between concepts also incorporated into cognitive diagnosis.

Table 5

Experimental results of different methods on the task of student performance prediction. The best results are marked in bold.

Method	CBTD			CMTD			Junyi		
	ACC	AUC	RMSE	ACC	AUC	RMSE	ACC	AUC	RMSE
IRT	72.22	78.62	42.49	71.00	78.54	66.06	67.60	77.50	42.68
MIRT	72.34	76.19	44.79	71.77	78.25	45.92	75.13	79.89	41.17
PMF	70.24	72.66	47.46	74.08	81.44	58.58	68.34	76.44	43.73
NeuralCD	75.62	81.83	41.91	74.12	80.79	42.39	74.43	79.09	41.72
RCD	76.05	82.34	40.55	75.96	82.95	40.98	77.16	82.62	39.63
CMNCD	79.02	84.38	39.45	76.89	84.08	39.51	78.04	83.82	40.12

Table 6

Results of statistical classification metrics for proposed CMNCD on CBTD, CMTD and JunYi datasets (%).

Dataset	ACC	Sensitivity	Specificity	Precision	AUC	F1-Score
CBTD	79.02	83.27	72.44	82.89	84.38	83.08
CMTD	76.89	79.91	73.92	82.96	84.08	81.04
JunYi	78.04	82.46	71.88	81.98	83.82	82.22

6.3. Performance evaluation

In this section, we first show the loss and ROC curves for our model on three experimental datasets in Figs. 5 and 6, respectively. As seen from Fig. 5, the loss values of our model decrease rapidly and reach a relatively stable range after 60 epoches. This indicates our model can converge efficiently. Then, we compare CMNCD with several state-of-the-art cognitive diagnosis methods on benchmark datasets (RQ1). Then, we conduct an ablation study to validate the effectiveness of each module in CMNCD (RQ2-RQ5).

6.3.1. Comparisons with state-of-the-art methods (RQ1)

In this part, we compare our model with several classical and state-of-the-art cognitive diagnosis models to address RQ1. The experimental results of different methods on the benchmark datasets are shown in Table 5. As seen from the table, on both cross-modal and single-modal datasets, CMNCD consistently outperforms the baseline methods on all the metrics. For example, CMNCD achieves the highest Accuracy and AUC values on the CBTD dataset, i.e., 78.9% and 84.4%, showing an improvement against previous methods by at least 2.85% and 2.06%, respectively. This demonstrates the effectiveness of CMNCD on the task of student performance prediction, which can answer RQ1. We also further validate the performance of our model by some statistical metrics and show the experimental results in Table 6. The superior performance of CMNCD over the baselines can be attributed to two reasons:

- (1) CMNCD explores the rich semantic information in cross-modal contents of exercises for student performance prediction, where the cross-modal contents are important for learning distinguishable exercises' trait vectors. By contrast, with no considering the cross-modal contents, these baseline methods only can exploit student performance to estimate exercises' trait vectors under the framework of MIRT. This impairs the ability of these methods to distinguish the traits of different cross-modal exercises, and thus impedes the improvement of their results on the task of student performance prediction.
- (2) CMNCD takes into account the complexity of the concepts being examined when modeling student performance, which endows the concepts of different complexity with different effects on the trait learning exercises. This is conducive to avoid learning illogical exercises' trait vectors, where the trait (e.g., difficulty) value on one concept not matching the complexity of the concept. By contrast, the baselines overlook the effect of the concept complexity on exercises' trait vectors.

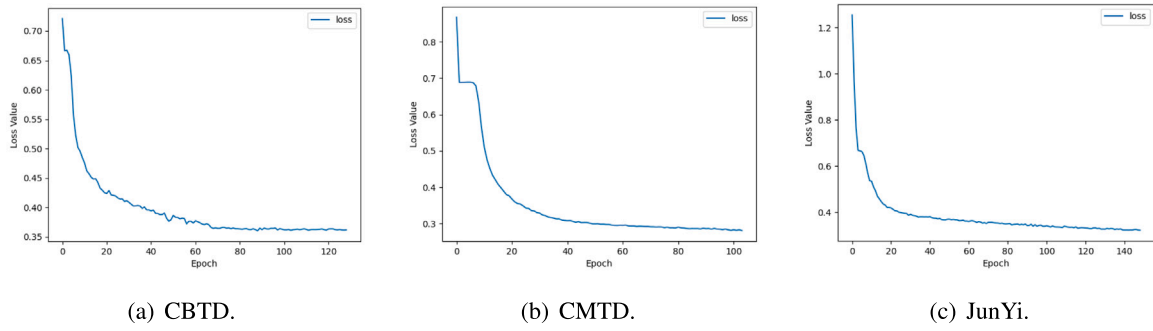


Fig. 5. The loss curves for our model on (a) CBTD, (b) CMTD and (c) JunYi datasets.

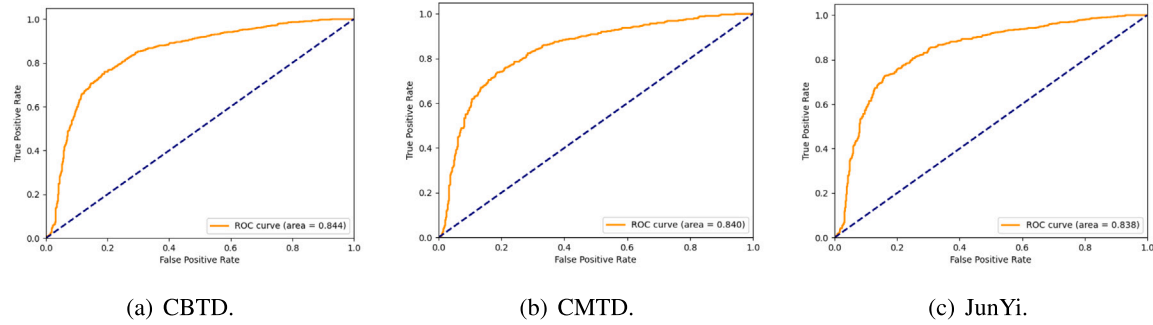


Fig. 6. The ROC curves for our model on (a) CBTD, (b) CMTD and (c) JunYi datasets.

6.3.2. Ablation study (RQ2 ~ RQ5)

In this part, we test the performance of CMNCD with following different configurations to address RQ2 ~ RQ5. Specifically, the variants *CMNCD w/o Text* and *CMNCD w/o Image* are used to address RQ2. The variants *CMNCD w/o CA-GCF*, *CMNCD w/o Concept complexity* and *CMNCD w/o Deep MIRT* are used to address RQ3, RQ4 and RQ5, respectively. Besides, we also test the performance of our CMNCD without using any exercises' contents by the variant *CMNCD w/o Exercises' contents*. The details of each variant model are represented as follows.

(1) *CMNCD w/o Text*: it is obtained by replacing the component *Cross-modal feature fusion* in Module I with a *ResNet* followed by a pyramid pooling layer. That is, this variant only use image information for learning exercises' traits.

(2) *CMNCD w/o Image*: it is obtained by replacing the component *Cross-modal feature fusion* in Module I with a bert followed by a pyramid pooling layer. Thus, when learning exercises' trait vectors, this variant model only considers the information in exercises' question texts.

(3) *CMNCD w/o CA-GCF*: it is obtained by replacing the cross-modal feature fusion strategy introduced in Module I, i.e., the CA-GCF mechanism, with an existing well-known cross-modal fusion strategy, namely UFSCAN (Zhang, Chen et al., 2021).

(4) *CMNCD w/o Concept complexity*: it is obtained by removing the Module II. With no considering the complexity of examined concepts derived from the prerequisite relations among concepts, this variant model learns the trait vectors of exercises only based on the cross-modal contents of exercises.

(5) *CMNCD w/o Deep MIRT*: it is obtained by replacing the module *Deep MIRT* with the classic MIRT model, which aims to demonstrate the effectiveness of our *Deep MIRT* module in predicting students' performance.

(6) *CMNCD w/o Exercises' contents*: it is obtained by removing the Module I. That is, this variant only incorporates the prerequisite relations among concepts into the learning of exercise traits.

The experiment results of different variant models on two cross-modal datasets are reported in Table 7. From the table, we have the following five observations.

Table 7

Experimental results of the ablation study on CBTD and CMTD.

Method	CBTD			CMTD		
	ACC	AUC	RMSE	ACC	AUC	RMSE
CMNCD(full model)	79.02	84.38	39.45	76.89	84.08	39.51
-w/o Text	77.11	81.88	41.03	75.10	80.81	41.75
-w/o Image	77.04	81.90	40.97	74.93	79.97	42.99
-w/o CA-GCF	77.03	82.75	40.55	75.29	82.23	41.68
-w/o Concept complexity	74.32	80.44	42.43	73.52	78.51	42.74
-w/o Deep MIRT	65.95	73.56	48.31	69.05	72.46	45.38
-w/o Exercises' contents	75.10	78.18	43.14	72.49	76.50	45.38

- (1) Without exploiting the text or image contents of exercises, CMNCD witnesses a performance drop on both datasets. For example, the AUC of the variant *CMNCD w/o Image* has 2.5% and 4.1% drops on CBTD and CMTD, respectively. This demonstrates that either textual or visual information of cross-modal exercises is important for modeling student performance, which can answer RQ2. This is because some examined concepts only occur in one modality data (e.g., images or texts), and some other concepts are described by both modalities. Therefore, *CMNCD w/o Image* and *CMNCD w/o Text* cannot completely get the information about examined concepts to learn exercises' trait vectors, which leads to the performance degeneration. The necessity of exploring the cross-modal contents of exercises for modeling student performance can also be observed by the comparisons between the full model and *CMNCD w/o Exercises' contents*.
- (2) After replacing the CA-GCF mechanism, *CMNCD w/o CA-GCF* also suffers from a performance drop, e.g., from 84.4% to 82.75% in terms of AUC on CMTD. This validates the effectiveness of the CA-GCF on performance prediction, which can answer RQ3. With the CA-GCF mechanism, *CMNCD* can adaptively select relevant information from the text and image contents of exercises when learning exercise's trait vectors, suppress the noisy information in each modality. This improves the trait

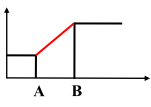



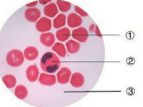
<p>Q1. Urea levels change as blood flows through the nephron. If the vertical axis in the following figure represents the change of urea content, then where is the AB segment on the abscissa?</p>  <p>Real response: 1 Predicted score: 0.9746 Prediction result is right</p>	<p>Q2. Describe the relationship among producer, consumer and decomposer according to the diagram below.</p>  <p>Real response: 1 Predicted score: 0.9971 Prediction result is right</p>	<p>Q3. The picture below shows an eco-bottle containing lake water, water plants and a few small fish. How long would you predict the creatures in the bottle would live? State your reasons.</p>  <p>Real response: 1 Predicted score: 0.9893 Prediction result is right</p>	<p>Q4. Please use daphnia in the picture as biological material to design an experiment to explore the influence of external factors such as coke, alcohol and tobacco on its heart rate.</p>  <p>Real response: 0 Predicted score: 0.2906 Prediction result is right</p>	<p>Q5. Below is a part of the blood smear observed by a student under amicroscope. Which of the smears can stop bleeding and accelerate clotting?</p>  <p>Real response: 1 Predicted score: 0.9666 Prediction result is right</p>
--	---	--	---	--

Fig. 7. Qualitative results of the proposed CMNCD. Green means the forecast is correct.

vectors of exercises and thus helps to improve the predicted results of student performance.

- (3) The performance of *CMNCD w/o Concept complexity* obviously degenerates on both datasets. Specifically, the AUC value of CMNCD is decreased from 84.38% to 80.44% on CBTD. That is, the performance of CMNCD shows a sharp drop when the concept complexity is overlooked for learning the trait vectors of exercises. The performance drop indicates the effectiveness of the concept complexity on student performance prediction, and thus can answer RQ4 and demonstrate the correctness of our second hypotheses. The reason for the performance degeneration mainly lies in that this variant model may generate exercises' trait vectors running counter to the complexity of concepts. For example, exercises exhibit higher difficulty trait on the concepts of low complexity than the concepts of high complexity. These illogical exercises' trait vectors can result in inaccurate student knowledge proficiency vectors, leading to incorrect results for student performance prediction.
- (4) Without the module of deep MIRT, the performance of CMNCD has a drastic drop on all the datasets. For example, the AUC values on CBTD and CMTD witness 10.84% and 11.61% drops, respectively. The drastic performance drop demonstrates the effectiveness of our Deep MIRT module on the prediction of student performance, which can answer RQ5. Without the deep MIRT module, this variant model uses a simple logistic-like interaction function, which is not sufficient for capturing the complex interactions between students and exercises (Wang, Liu et al., 2022), and thus results in the performance drop.
- (5) The performance of *CMNCD* witnesses a significantly drop when the trait vectors of exercises are learned without using cross-modal contents of exercises, while the performance of *CMNCD* is improved significantly with one or both of two modality information in exercises. For example, the comparison between *CMNCD w/o Exercises' contents* and *CMNCD w/o Image* show that the introduction of image contents of exercises lead to 3.72% and 3.47% improvements in AUC values for CMNCD on CBTD and CMTD datasets, respectively. Besides, the performance of CMNCD can be significantly improved by combining the image and text information of exercises. This can be demonstrated by comparing *CMNCD w/o CA-GCF* with variant models that do not use exercise's contents or use only one modality information, where *CMNCD w/o CA-GCF* combines different modality information by an existing cross-modal fusion strategy. Taking the results on CBTD as an example, *CMNCD w/o CA-GCF* achieves a 4.57% improvement in AUC values over *CMNCD w/o Exercises' contents*, 0.85% improvement over *CMNCD w/o Image* and 0.87% improvement over *CMNCD w/o Text*. The comparison results demonstrate the correctness of our first hypotheses introduced in Section 1.

6.4. Case study

To illustrate the effectiveness of CMNCD, here we show the cognitive diagnosis results on some cross-modal exercises from the CBTD dataset. The selected exercises are shown in Fig. 7 and the results are shown in Fig. 8. Specifically, we use a radar chart on the left to show the proficiency of a student on concepts diagnosed by MIRT, NeuralCD and CMNCD. The performance prediction results of the student are shown in the table on the right.

From the table, we observe that CMNCD can provide more accurate prediction results than MIRT and NeuralCD. Specifically, MIRT and NeuralCD give wrong predictions for the fourth and the third exercises, respectively. By comparing student proficiency and exercise trait of difficulty, we see that student performance tends to be positive when students' proficiency satisfies exercises' requirement, and otherwise negative.

For example, given the third exercise, the proficiency of one student on the concept (i.e., k_3) estimated by NeuralCD is 0.7751, which is lower than the estimated difficulty on this concept, i.e., 0.8076. Thus, NeuralCD gives a negative result, however, which goes against with the real response record. By contrast, the student proficiency on the concept k_3 estimated by our CMNCD is 0.9981, which is higher than the estimated difficulty trait, i.e., 0.6823, leading to a correct prediction. This indicates that CMNCD can give more accurate diagnosis results than the baselines.

7. Limitation of study

In our work, the complexity of the concepts examined by exercises is estimated to enhance the exercise difficulty learning, where the complexity estimation is based on the *Prerequisite Concept Graph*. In CMNCD, the structure of the *Prerequisite Concept Graph* is assumed to be fixed and the prerequisite relations between concepts require manual annotation. However, new concepts may emerge with growing exercise records. In this case, CMNCD needs to be retrained based on the prerequisite relation annotations for new concepts. This is because the structure of *Prerequisite Concept Graph* changes with the addition of new concepts, which requires the complexity of all the concepts to be re-estimated based on new graph structure. Therefore, CMNCD cannot dynamically estimate the complexity of concept nodes in dynamic *Prerequisite Concept Graphs*, and thus cannot learn the complexity-based exercise difficulty traits online. This makes it difficult for CMNCD to handle exercises examining new concepts that have never been seen before, which limits its applicability to emerging student performance records.

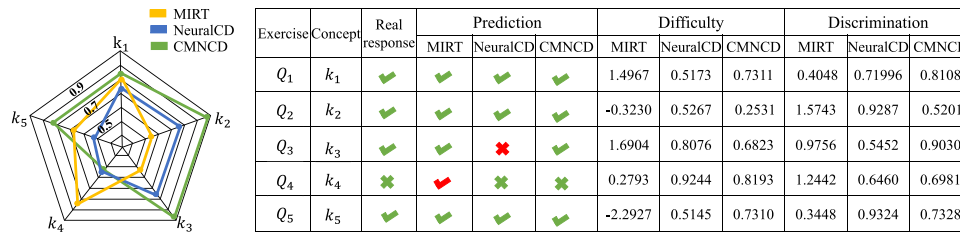


Fig. 8. Diagnosis examples of a student on the cross-modal exercises from the CBTD dataset. The concept k_1 denotes *Urinary system*, k_2 and k_3 both denote the *Ecosystem structure and function*, k_4 denotes the *Circulatory system* and k_5 denotes the *Cell structure and function*.

8. Conclusions and future work

In this paper, we study how to explore the contents of cross-modal exercises and concept complexity to improve the results of cognitive diagnosis. In Section 2, we summarize the studies on the task of cognitive diagnosis and compare our CMNCD with related works. Different from previous studies that mostly overlook exercises' contents, our CMNCD explores both visual and textual contents of cross-modal exercises for enhancing the learning of exercises' trait vectors. Specifically, the fine-grained semantics of cross-modal exercises are explored by a CA-GCF mechanism, where cross-modal information is adaptively selected and combined to estimate the trait vectors of exercises under the guidance of examined concepts. In addition, we also model the concept complexity by the prerequisite relations between concepts, and incorporate the complexity into the learning of exercises' trait vectors. The superior performance of CMNCD on both cross-modal and single-modal benchmark datasets demonstrate its effectiveness on the task of cognitive diagnosis.

Our work also have some limitations. For example, one limitation of the proposed CMNCD is that it cannot learn the prerequisite relations between newly incoming and existing concepts automatically. To address this issue, one of the solutions is to endow CMNCD with the ability to construct the *Prerequisite Concept Graph* automatically. Previous works (Chen, Lu, Zheng, Chen, & Li, 2018; Yu, Li, Mao and Cai, 2021) have studied the methods for knowledge graph construction, where new concepts and the relations between concepts can be extracted and inferred automatically. Inspired by these works, in the future, we can investigate how to accurately construct the *Prerequisite Concept Graph* automatically, where new concepts can be extracted from education texts (e.g., exercises) and knowledge bases by the named entity recognition techniques, e.g., K12EduKG (Chen et al., 2018). The prerequisite relations between concepts can be inferred from online educational resources based on latent representations of concepts, such as the methods used in Pan et al. (2017) and Roy, Madhyastha, Lawrence, and Rajan (2019). Furthermore, online knowledge graph construction methods also can be used to build and update the *Prerequisite Concept Graph* dynamically based on streaming data, e.g., Stream2Graph (Barry et al., 2022).

The second limitation of our work lies in the dynamic learning of exercises' trait based on concept complexity. Previous works have investigated how to learn and track non-linear functions over the nodes of the graphs with dynamic structures, e.g., online MultiKernel Learning framework (MKL) (Shen, Leus, & Giannakis, 2019). In the future, we would follow the methods like MKL to extend the module of concept complexity based exercise difficulty estimation (i.e., Module II) to an online learning method, where the nonlinear mapping functions from the complexity of the concept nodes of the *Prerequisite Concept Graph* to exercises' difficulty traits. By addressing the above two issues, CMNCD is expected to develop into an online deep neural cognitive diagnosis framework.

Besides, in recent years many deep learning efforts (Chen et al., 2022; Liu, Gao, Li, Fu and Ding, 2023; Liu, Zhang et al., 2023; Mercea, Riesch, Koepke, & Akata, 2022) have been proposed to learn

visual, textual and cross-modal features, including attention-based and transformer-based methods, etc. Many attention-based methods have been developed to solve different problems, such as Contour-enhanced attention CNN for image segmentation (Karthik, Menaka, Hariharan, & Won, 2022) and attention-Bi-LSTM for text classification (Zheng, Gao, Shen, & Zhai, 2022). In the future, inspired by existing attention mechanisms, we would further study the attention mechanism for identifying key visual features for the images of cross-modal exercises under the framework of zero-shot learning. The transformer-based methods introduce the attention mechanisms into encoder-decoder frameworks to enhance representation learning for various language and vision tasks, e.g., vision Transformers (Dosovitskiy et al., 2021) and linguistic Transformers (Devlin et al., 2019; Lin, Wang, Liu, & Qiu, 2022). Future works can use Transformer-like architectures to improve the representation learning of cross-modal exercises.

CRedit authorship contribution statement

Lingyun Song: Conceptualization, Methodology, Supervision, Writing – original draft, Funding acquisition. **Mengting He:** Conceptualization, Methodology, Software. **Xuequn Shang:** Conceptualization, Validation, Methodology, Supervision, Writing – review & editing. **Chen Yang:** Investigation, Data curation, Writing – review & editing. **Jun Liu:** Conceptualization, Methodology, Writing – review & editing. **Mengzhen Yu:** Investigation, Data curation, Formal analysis. **Yu Lu:** Data curation, Conceptualization, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The research was supported in part by National Nature Science Foundation of China under Grant Nos. 62102321, and Fundamental Research Funds for the Central Universities, China under Grant No. D5000230095, D5000200146.

References

- Bakkali, S., Ming, Z., Coustaty, M., Rusiñol, M., & Terrades, O. R. (2023). VI-cdoc: Vision-language contrastive pre-training model for cross-modal document classification. *Pattern Recognition*, 139, Article 109419.
- Barry, M., Bifet, A., Chiky, R., El Jaouhari, S., Montiel, J., El Ouafi, A., et al. (2022). Stream2Graph: Dynamic knowledge graph for online learning applied in large-scale network. In *2022 IEEE international conference on big data (Big Data)* (pp. 2190–2197).

- Baylari, A., & Montazer, G. A. (2009). Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications*, 36(4), 8013–8021.
- Bokade, R., Navato, A., Ouyang, R., Jin, X., Chou, C.-A., Ostadabbas, S., et al. (2021). A cross-disciplinary comparison of multimodal data fusion approaches and applications: Accelerating learning through trans-disciplinary information sharing. *Expert Systems with Applications*, 165, Article 113885.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159.
- Bu, C., Liu, F., Cao, Z., Li, L., Zhang, Y., Hu, X., et al. (2022). Cognitive diagnostic model made more practical by genetic algorithm. *IEEE Transactions on Emerging Topics in Computational Intelligence*.
- Chang, H.-S., Hsu, H.-J., & Chen, K.-T. (2015). Modeling exercise relationships in E-learning: A unified approach. In *Proceedings of the 8th international conference on educational data mining* (pp. 532–535).
- Chen, Y., Culpepper, S., & Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika*, 1–33.
- Chen, C.-M., & Duh, L.-J. (2008). Personalized web-based tutoring system based on fuzzy item response theory. *Expert Systems with Applications*, 34(4), 2298–2315.
- Chen, S., Hong, Z., Liu, Y., Xie, G.-S., Sun, B., Li, H., et al. (2022). Transzero: Attribute-guided transformer for zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 330–338).
- Chen, C.-M., Liu, C.-Y., & Chang, M.-H. (2006). Personalized curriculum sequencing utilizing modified item response theory for web-based instruction. *Expert Systems with Applications*, 30(2), 378–396.
- Chen, P., Lu, Y., Zheng, V. W., Chen, X., & Li, X. (2018). An automatic knowledge graph construction system for K-12 education. In *Proceedings of the fifth annual ACM conference on learning at scale* (pp. 1–4).
- Cheng, Y., Li, M., Chen, H., Cai, Y., Sun, H., Wu, G., et al. (2021). Neural cognitive modeling based on the importance of knowledge point for student performance prediction. In *Proceedings of the 16th international conference on computer science & education* (pp. 495–499).
- Cheng, S., & Liu, Q. (2019). Enhancing item response theory for cognitive diagnosis. In *Proceedings of the conference on information and knowledge management*.
- Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., et al. (2019). DIRT: Deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 2397–2400).
- De La Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130.
- DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, 35(1), 8–26.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American chapter of the association for computational linguistics*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations*.
- Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika*, 78(1), 14–36.
- Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19, 243–266.
- Gan, W., Sun, Y., & Sun, Y. (2022). Knowledge interaction enhanced sequential modeling for interpretable learner knowledge diagnosis in intelligent education systems. *Neurocomputing*.
- Gao, Y., Beijbom, O., Zhang, N., & Darrell, T. (2016). Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 317–326).
- Gao, W., Liu, Q., Huang, Z., Yin, Y., Bi, H., Wang, M.-C., et al. (2021). RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international conference on research and development in information retrieval* (pp. 501–510).
- Gao, L., Zhao, Z., Li, C., Zhao, J., & Zeng, Q. (2022). Deep cognitive diagnosis model for predicting students' performance. *Future Generation Computer Systems*, 126, 252–262.
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256).
- Guo, W., Zhang, Y., Wu, X., Yang, J., Cai, X., & Yuan, X. (2020). Re-attention for visual question answering. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 91–98).
- Gupta, D., Suman, S., & Ekbal, A. (2021). Hierarchical deep multi-modal network for medical visual question answering. *Expert Systems with Applications*, 164, Article 113993.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), 1904–1916.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Hooshyar, D., Huang, Y.-M., & Yang, Y. (2022). GameDKT: Deep knowledge tracing in educational games. *Expert Systems with Applications*, 196, Article 116670.
- Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., et al. (2017). Question difficulty prediction for READING problems in standard tests. In *Thirty-First AAAI conference on artificial intelligence*.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272.
- Karthik, R., Menaka, R., Hariharan, M., & Won, D. (2022). Contour-enhanced attention CNN for CT-based COVID-19 segmentation. *Pattern Recognition*, 125, Article 108538.
- Kim, J.-H., On, K.-W., Lim, W., Kim, J., Ha, J.-W., & Zhang, B.-T. (2016). Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.
- Li, S., He, Z., Guan, Q., He, Y., Fang, L., Luo, W., et al. (2022). Cognitive diagnosis focusing on knowledge components. In *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners' and doctoral consortium: 23rd International conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part II* (pp. 311–314).
- Li, G., Hu, Y., Shuai, J., Yang, T., Zhang, Y., Dai, S., et al. (2022). NeuralNCD: A neural network cognitive diagnosis model based on multi-dimensional features. *Applied Sciences*, 12(19), 9806.
- Li, J., Wang, F., Liu, Q., Zhu, M., Huang, W., Huang, Z., et al. (2022). HierCDF: A Bayesian network-based hierarchical cognitive diagnosis framework. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 904–913).
- Liang, C., Ye, J., Wang, S., Pursel, B., & Giles, C. L. (2018). Investigating active learning for concept prerequisite learning. In *Thirty-second AAAI conference on artificial intelligence*.
- Lin, Z., Bas, E., Singh, K. Y., Swaminathan, G., & Bhotika, R. (2023). Relaxing contrastiveness in multimodal representation learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2227–2236).
- Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*.
- Liu, Q. (2021). Towards a new generation of cognitive diagnosis. In *IJCAI* (pp. 4961–4964).
- Liu, S., Gao, P., Li, Y., Fu, W., & Ding, W. (2023). Multi-modal fusion network with complementarity and importance for emotion recognition. *Information Sciences*, 619, 679–694.
- Liu, J., Hou, J., Zhang, N., Liu, Z., & He, W. (2023). Learning evidential cognitive diagnosis networks robust to response bias. In *The proceedings of the AAAI international conference on artificial intelligence* (pp. 171–181).
- Liu, M., Shao, P., & Zhang, K. (2021). Graph-based exercise-and knowledge-aware learning network for student performance prediction. In *Proceedings of the AAAI international conference on artificial intelligence* (pp. 27–38).
- Liu, Q., Wu, R., Chen, E., Xu, G., Su, Y., Chen, Z., et al. (2018). Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology*, 9(4), 1–26.
- Liu, P., Zhang, L., & Gulla, J. A. (2020). Dynamic attention-based explainable recommendation with textual and visual fusion. *Information Processing & Management*, 57(6), Article 102099.
- Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., et al. (2023). A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*.
- Liu, Y., Zhang, X., Zhang, Q., Li, C., Huang, F., Tang, X., et al. (2021). Dual self-attention with co-attention networks for visual question answering. *Pattern Recognition*, 117, Article 107956.
- Liu, S., Zou, R., Sun, J., Zhang, K., Jiang, L., Zhou, D., et al. (2021). A hierarchical memory network for knowledge tracing. *Expert Systems with Applications*, 177, Article 114935.
- Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.
- Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems*, 29, 289–297.
- Ma, H., Huang, Z., Tang, W., Zhu, H., Zhang, H., & Li, J. (2023). Predicting student performance in future exams via neutrosophic cognitive diagnosis in personalized E-learning environment. *IEEE Transactions on Learning Technologies*.
- Ma, H., Li, M., Wu, L., Zhang, H., Cao, Y., Zhang, X., et al. (2022). Knowledge-sensed cognitive diagnosis for intelligent education platforms. In *Proceedings of the 31st ACM international conference on information & knowledge management* (pp. 1451–1460).
- Mercea, O.-B., Riesch, L., Koepke, A., & Akata, Z. (2022). Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10553–10563).
- Mnih, A., & Salakhutdinov, R. R. (2008). Probabilistic matrix factorization. In *Advances in neural information processing systems* (pp. 1257–1264).
- Mohammed, H. M., Omeroglu, A. N., & Oral, E. A. (2023). MMHFNet: Multi-modal and multi-layer hybrid fusion network for voice pathology detection. *Expert Systems with Applications*, Article 119790.

- Mou, L., Zhou, C., Zhao, P., Nakisa, B., Rastgoo, M. N., Jain, R., et al. (2021). Driver stress detection via multimodal fusion using attention-based CNN-LSTM. *Expert Systems with Applications*, 173, Article 114693.
- Nam, H., Ha, J. W., & Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 299–307).
- Pan, L., Li, C., Li, J., & Tang, J. (2017). Prerequisite relation learning for concepts in moocs. In *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 1447–1456).
- Pei, H., Yang, B., Liu, J., & Dong, L. (2018). Group sparse bayesian learning for active surveillance on epidemic dynamics. In *Thirty-second AAAI conference on artificial intelligence*.
- Qi, T., Ren, M., Guo, L., Li, X., Li, J., & Zhang, L. (2023). ICD: A new interpretable cognitive diagnosis model for intelligent tutor systems. *Expert Systems with Applications*, 215, Article 119309.
- Roy, S., Madhyastha, M., Lawrence, S., & Rajan, V. (2019). Inferring concept prerequisite relations from online educational resources. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 9589–9594).
- Shen, Y., Leus, G., & Giannakis, G. B. (2019). Online graph-adaptive learning with scalability and privacy. *Sport Psychologist*, 67(9), 2471–2483.
- Shoeibi, A., Khodatars, M., Jafari, M., Ghassemi, N., Moridian, P., Alizadesani, R., et al. (2022). Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review. *Information Fusion*.
- Shoeibi, A., Khodatars, M., Jafari, M., Moridian, P., Rezaei, M., Alizadehsani, R., et al. (2021). Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Computers in Biology and Medicine*, 136, Article 104697.
- Shoeibi, A., Rezaei, M., Ghassemi, N., Namadchian, Z., Zare, A., & Gorriz, J. M. (2022). Automatic diagnosis of schizophrenia in EEG signals using functional connectivity features and CNN-LSTM model. In *Artificial intelligence in neuroscience: Affective analysis and health applications: 9th International work-conference on the interplay between natural and artificial computation* (pp. 63–73).
- Shoeibi, A., Sadeghi, D., Moridian, P., Ghassemi, N., Heras, J., Alizadehsani, R., et al. (2021). Automatic diagnosis of schizophrenia in EEG signals using CNN-LSTM models. *Frontiers in Neuroinformatics*, 58.
- Song, L., Liu, J., Qian, B., & Chen, Y. (2019). Connecting language to images: A progressive attention-guided network for simultaneous image captioning and language grounding. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8885–8892).
- Song, L., Yu, M., Shang, X., Lu, Y., Liu, J., Zhang, Y., et al. (2022). A deep grouping fusion neural network for multimedia content understanding. *IET Image Processing*, 16(9), 2398–2411.
- Tong, S., Liu, J., Hong, Y., Huang, Z., Wu, L., Liu, Q., et al. (2022). Incremental cognitive diagnosis for intelligent education. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 1760–1770).
- Tong, S., Liu, Q., Yu, R., Huang, W., Huang, Z., Pardos, Z., et al. (2021). Item response ranking for cognitive diagnosis. In *IJCAI*.
- Tong, H., Zhou, Y., & Wang, Z. (2020). Exercise hierarchical feature enhanced knowledge tracing. In *International conference on artificial intelligence in education* (pp. 324–328).
- Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv: 1908.08962*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. In *ICLR*.
- Wang, S., Fu, P., Fu, M., Li, B., Zhang, B., Chen, Z., et al. (2022). Continuous weighted neural cognitive diagnosis method for online education. In *Proceedings of the 8th international conference on artificial intelligence and security* (pp. 142–150).
- Wang, F., Huang, Z., Liu, Q., Chen, E., Yin, Y., Ma, J., et al. (2023). Dynamic cognitive diagnosis: An educational priors-enhanced deep knowledge tracing perspective. *IEEE Transactions on Learning Technologies*.
- Wang, F., Liu, Q., Chen, E., Huang, Z., Yin, Y., Wang, S., et al. (2022). NeuralCD: A general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*.
- Wang, W., Ma, H., Zhao, Y., Li, Z., & He, X. (2022). Tracking knowledge proficiency of students with calibrated Q-matrix. *Expert Systems with Applications*, 192, Article 116454.
- Wang, Z., Yan, W., Zeng, C., Tian, Y., Dong, S., et al. (2023). A unified interpretable intelligent learning diagnosis framework for learning performance prediction in intelligent tutoring systems. *International Journal of Intelligent Systems*, 2023.
- Wu, R., Liu, Q., Liu, Y., Chen, E., Su, Y., Chen, Z., et al. (2015). Cognitive modelling for predicting examinee performance. In *The proceedings of the twenty-fourth international joint conference on artificial intelligence*.
- Wu, K., Yang, Y., Zhang, K., Wu, L., Liu, J., & Li, X. (2023). Multi-relational cognitive diagnosis for intelligent education. In *The proceedings of the CAAI international conference on artificial intelligence* (pp. 425–437).
- Xu, J., Li, Q., Liu, J., Lv, P., & Yu, G. (2021). Leveraging cognitive diagnosis to improve peer assessment in moocs. *IEEE Access*, 9, 50466–50484.
- Yang, H., Qi, T., Li, J., Guo, L., Ren, M., Zhang, L., et al. (2022). A novel quantitative relationship neural network for explainable cognitive diagnosis model. *Knowledge-Based Systems*, 250, Article 109156.
- Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, 30(6), 469–492.
- Yu, H., Li, H., Mao, D., & Cai, Q. (2021). A domain knowledge graph construction method based on Wikipedia. *Journal of Information Science*, 47(6), 783–793.
- Yu, C., Xue, H., Jiang, Y., An, L., & Li, G. (2021). A simple and efficient text matching model based on deep interaction. *Information Processing & Management*, 58(6), Article 102738.
- Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 1821–1830).
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the 13th European conference on computer vision* (pp. 818–833).
- Zhang, S., Chen, M., Chen, J., Zou, F., Li, Y.-F., & Lu, P. (2021). Multimodal feature-wise co-attention method for visual question answering. *Information Fusion*, 73, 1–10.
- Zhang, S., Li, H., & Kong, W. (2021). A cross-modal fusion based approach with scale-aware deep representation for RGB-D crowd counting and density estimation. *Expert Systems with Applications*, 180, Article 115071.
- Zhang, J., Mo, Y., Chen, C., & He, X. (2021). GKT-CD: Make cognitive diagnosis model enhanced by graph-based knowledge tracing. In *2021 International joint conference on neural networks (IJCNN)* (pp. 1–8).
- Zheng, Y.-f., Gao, Z.-h., Shen, J., & Zhai, X.-s. (2022). Optimising automatic text classification approach in adaptive online collaborative discussion-A perspective of attention mechanism-based Bi-LSTM. *IEEE Transactions on Learning Technologies*.
- Zhou, Y., Liu, Q., Wu, J., Wang, F., Huang, Z., Tong, W., et al. (2021). Modeling context-aware features for cognitive diagnosis in student learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 2420–2428).