NeuralCD: A General Framework for Cognitive Diagnosis

Fei Wang, Qi Liu, *Member, IEEE,* Enhong Chen, *Senior Member, IEEE,* Zhenya Huang, Yu Yin, Shijin Wang, Yu Su

Abstract—Recent decades have witnessed a rapid growth of intelligent educational systems as well as an increasing demand for computer-aided educational technologies. One of the fundamental issues in intelligent education is cognitive diagnosis, which aims to discover the proficiency level of students on specific knowledge concepts. Existing approaches usually mine linear interactions of student exercising process by manually designed function (e.g., logistic function). However, the cognitive interactions between students and exercises is a complex process, and excessive simplifications would lead to under fitting and thus get inaccurate diagnostic results. Besides, the manually designed interaction functions are relatively inflexible and limits their extensibility. This consequently causes lack of consideration about useful non-numerical information in the cognitive process besides response logs. In this paper, we propose a general Neural Cognitive Diagnosis (NeuralCD) framework as well as several implemented models (a basic implementation NeuralCDM and three extensions), where we project students and exercises to factor vectors and incorporates neural networks to learn the complex exercising interactions. To ensure the interpretability of diagnostic results, which is essential for cognitive diagnosis, we apply an monotonicity assumption to our NeuralCD framework. Moreover, NeuralCD is a general framework and has good extensibility. We show the generality of NeuralCD through proving how it can cover some traditional models. Then, we demonstrate the extensibility of NeuralCD, which benefits future developments. On one hand, we demonstrate content-based extensions where we provide examples of exploring the rich contents of exercise texts (CNCD-Q and CNCD-F). On the other hand, we demonstrate a knowledge-association based extension to show that NeuralCD is flexible for structural adjustments so as to solve specific problems. For instance, we improve the diagnostic results on uncovered knowledge concepts of a student by extending NeuralCD with the knowledge associations consideration (KaNCD). Extensive experimental results on real-world datasets show the effectiveness of NeuralCD framework with both accuracy and interpretability.

Index Terms—Intelligent education, personalized learning, cognitive diagnosis, neural network.

1 Introduction

COGNITIVE diagnosis has been studied for decades, and researchers (especially from psychometrics and education) have obtained rich achievements. The purpose of cognitive diagnosis is to discover a person's cognitive state (e.g. skill proficiency) from the person's behaviors (e.g. test results). It is a necessary and fundamental task in many real-world scenarios such as games [1], clinical measurement [2], [3] and education, where users' (e.g., players, patients, students) abilities require assessments, and thus attracts wide attention. Specifically, in intelligent educational systems [4], [5], cognitive diagnosis aims to discover the states of students in the learning process, such as their proficiencies on specific knowledge concepts [6]. Figure 1 shows a toy example of cognitive diagnosis. Generally, students usually first choose to practice a set of exercises (e.g., $e_1, \cdots, e_4$) and leave their responses (e.g., right or wrong). Then, our goal is to infer their actual knowledge states on the corresponding concepts (e.g., *Equation*). In practice, these diagnostic reports are necessary as they serve as the basis of further supports, such as exercise recommendation, targeted training [7] and computerized adaptive testing [8].

In the field of psychometrics, massive efforts have been devoted for cognitive diagnosis, such as Deterministic Inputs, Noisy-And gate model (DINA) [9], Item Response Theory (IRT) [10] and Multidimensional IRT (MIRT) [11]. Despite achieving some effectiveness, these works rely on handcrafted functions that model the interaction between student and questions. The interaction functions are designed based on assumptions that are the simplification of real interaction process, and are mostly linear, such as logistic-like function in IRT [10] or inner product in matrix factorization [12]. However, the interaction between students and exercises is a complex non-linear process, and excessive simplifications would lead to under fitting of the process and thus get inaccurate diagnostic results and restrict the application scope of the models. Besides, these works were proposed mostly for scale-based tests where a set of examinees are tested with the same small set of questions, e.g., terminal examination of schools. In a scale-based test, all examinees are supposed to answer all the questions, therefore the response data is complete and usually not large. While for broader applications of cognitive diagno-

- *F. Wang, Q. Liu (corresponding author), E. Chen (corresponding author), Z. Huang and Y. Yin are with the Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China, and also with the State Key Laboratory of Cognitive Intelligence, Wangjiang West Road, Hefei, 230088, Anhui, China. Email: wf314159@mail.ustc.edu.cn, {qiliuql, cheneh, huangzhy}@ustc.edu.cn, yxonic@mail.ustc.edu.cn*
- *S. Wang is with iFLYTEK AI Research (Central China), Checheng North Road, Wuhan, 430058, Hubei, China, and also with the State Key Laboratory of Cognitive Intelligence, Wangjiang West Road, Hefei, 230088, Anhui, China. Email: sjwang3@iflytek.com*
- *Y. Su is with Hefei Normal University, Hefei, 230601, Anhui, China, and also with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Wangjiang West Road, Hefei, 230088, Anhui, China. Email: yusu@hfnu.edu.cn*

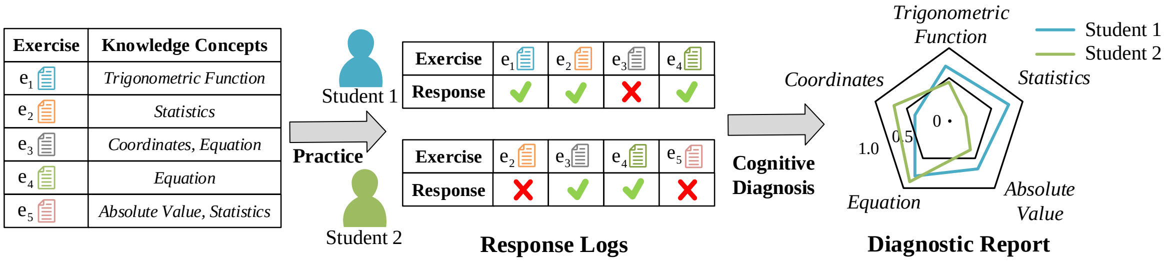


Fig. 1. A toy example of cognitive diagnosis.

k_1 : *Trigonometric Function*
k_2 : *Statistics*
k_3 : *Coordinates*
k_4 : *Equation*
k_5 : *Absolute Value*

$$\mathbf{Q} = \begin{array}{cccccc} k_1 & k_2 & k_3 & k_4 & k_5 & \\ 1 & 0 & 0 & 0 & 0 & e_1 \\ 0 & 1 & 0 & 0 & 0 & e_2 \\ 0 & 0 & 1 & 1 & 0 & e_3 \\ 0 & 0 & 0 & 1 & 0 & e_4 \\ 0 & 1 & 0 & 0 & 1 & e_5 \end{array}$$

Fig. 2. An example of Q-matrix.

sis, the data could be collected via different scenes, such as offline examinations and online self-regulated learning. Therefore, the response data could have large volume but sparse distribution, and more types of data becomes usable (e.g. texts, graph). The interaction patterns behind the data could be more elusive with simple functions, which makes it impractical to manually design the interaction functions. Fortunately on the other hand, the accumulation of data provides us an opportunity to apply data-driven methods to discover the complex interaction function [13].

In this paper, we address this issue in a principled way of proposing a Neural Cognitive Diagnosis (NeuralCD) framework by incorporating neural networks to model complex non-linear interactions. Although the capability of neural networks to approximate continuous functions has been proved in many domains, such as natural language processing [14] and recommender systems [15], it is still highly nontrivial to adapt to cognitive diagnosis due to the following domain challenges. First, the interpretability of diagnostic results, such as getting a student's mastery on certain knowledge concepts (e.g., *Equation*) is essential for cognitive diagnosis. However, the black-box nature of neural networks makes them difficult to get such explainable results. Second, the information contained in response logs is not complete for cognitive diagnosis. Extra resources such as exercise text content has valuable information (e.g. difficulty of reading comprehension) that beneficial for cognitive diagnosis. Thus it is necessary to make sure that our propose framework is extensible so as to aggregate the information from these resources. Third, a widely applicable framework should has a structure that is flexible to extend so as to meet different requirements in different situations. For example, in Figure 1, student 1 did not answer exercises related to *Absolute Value* and student 2 did not answer exercises related to *Trigonometric Function*. This is a common phenomenon especially in online exercises which are not scaled based, where the coverage of knowledge concepts in a student's response log is not complete due to the large total number of knowledge concepts and limited questions done by the student. In such situation, diagnostic models need to handle the knowledge coverage problem so as to obtain reliable diagnostic results when some knowledge concepts do not appear in a student's response history.

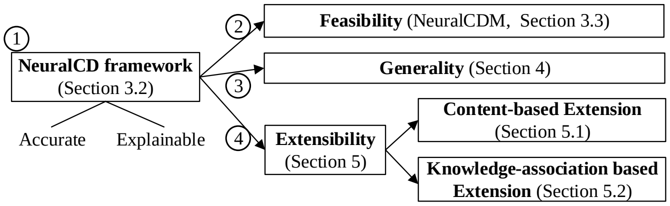

1
NeuralCD framework (Section 3.2)
Accurate
Explainable
2
Feasibility (NeuralCDM, Section 3.3)
3
Generality (Section 4)
4
Extensibility (Section 5)
Content-based Extension (Section 5.1)
Knowledge-association based Extension (Section 5.2)

Fig. 3. The organization of this work.

We propose a NeuralCD framework to address these challenges. We firstly introduce how we use NeuralCD to approximate interactions between students and exercises, for getting both accurate and explainable diagnostic results. As proposed in our preliminary work [16], in NeuralCD we projected students and exercises to factor vectors and leverage multi-layers for modeling the complex interactions in the process of exercise answering. To ensure the interpretability, without which the model would not be able to provide understandable diagnostic results and thus turns to a pure predicting model, we applied the monotonicity assumption taken from educational property [11] on the multi-layers. Then, we proposed a basic implementation of the framework called NeuralCDM, where we simply extracted exercise factor vectors from traditional Q-matrix and achieved the monotonicity property with nonnegative full connection layers, which showed feasibility of the framework. Traditional Q-matrix is an exercise-knowledge correlation matrix where $Q_{ij} = 1$ if exercise e_i contains knowledge concept k_j and 0 otherwise. The Q-matrix of the exercises in Figure 1 is shown in Figure 2.

After that, in this paper. we make further discussions and demonstrations about the two extra advantages of NeuralCD, i.e., generality and extensibility.

The generality of NeuralCD framework lies in its ability to cover some traditional models such as MF, IRT, MIRT. These models can be seen as degenerations of special cases of NeuralCD. With proper neural network structures, NeuralCD is capable of automatically learn different interaction functions that are suitable for the data.

As for the extensibility, we emphasize that a cognitive diagnosis framework should be open to extra information or better structures that benefit. Therefore we propose models that demonstrate the extensibility of NeuralCD from two aspects, i.e. content-based extension and knowledge-association based extension. In content-based extension, we demonstrate how information from exercise text can be explored with neural network to extend the framework. While in knowledge-association based extension, we provide a structural extension of NeuralCD to solve the knowledge concept coverage problem. Specifically, a representa-

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2022.3201037

tion based method is proposed to capture the associations among different knowledge concepts so as to improve the reliability of diagnostic results on uncovered knowledge concepts of a student. These two types of extensions can be combine together for better performance.

The organization of our work is presented in Figure 3. We first introduce our NeuralCD framework in Section 3.2. Then in Section 3.3, we demonstrate the feasibility of NeuralCD with an implemented model NeuralCDM. After that, generality and extensibility of NeuralCD is introduced in Section 4 and 5 respectively, and the extensibility is further discussed from two aspects, i.e. content-based extension and knowledge-association based extension.

Finally, we conduct extensive experiments on real-world datasets with basic and extended implementations, and the results show the effectiveness of NeuralCD framework with both accuracy and interpretability guarantee.

Our code is available at: https://github.com/bigdata-ustc/Neural_Cognitive_Diagnosis-NeuralCD

2 Related Work

In this section, we briefly review the related works from the following three aspects.

Cognitive Diagnosis. Existing works about student cognitive diagnosis mainly came from educational psychology area. DINA [9], [17] and IRT [10] were two of the most typical works, which model the result of a student answering an exercise as the interaction between the trait features of the student ($\boldsymbol{\theta}$) and the exercise ($\boldsymbol{\beta}$). Specifically, in DINA, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ were multi-dimensional and binary, where $\boldsymbol{\beta}$ came directly from Q-matrix (a human labeled exercise-knowledge correlation matrix). Another two exercise factors, i.e. guessing and slipping (parameterized as g and s) are also taken into consideration. The probability of student i correctly answering exercise j was modeled as $P(r_{ij} = 1|\boldsymbol{\theta}_i) = g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}}$, where $\eta_{ij} = \prod_k \theta_{ik}^{\beta_{jk}}$. On the other hand, in IRT, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ were unidimensional and continuous latent traits, indicating student ability and exercise difficulty. The interaction between the trait features was modeled in a logistic way, e.g., a simple version is $sigmoid(a(\boldsymbol{\theta} - \boldsymbol{\beta}))$, where a is the exercise discrimination parameter. Although extra parameters were added in IRT [18], [19] and latent trait was extended to multidimensional(MIRT) [11], [20], most of their item response functions were still logistic-like.

Recently, some researches from data mining perspective have demonstrated the feasibility of MF for cognitive diagnosis. Student and exercise correspond to user and item in matrix factorization (MF). For instance, Toscher et al. [21] improved SVD (Singular Value Decomposition) methods to factor the score matrix and get students and exercises' latent trait vectors. Thai-Nghe et al. [22], [23] applied some recommender system techniques including matrix factorization in the educational context, and compared it with traditional regression methods.

The interaction functions of theses traditional models are manually designed, which is based on various educational or psychometric theories or assumptions. For example, Reckase summarized the assumptions adopted by most IRT/MIRT models [11], including the independence among students, the invariance of students and exercises during a test, the monotonicity assumption, etc. In DINA model, a student is assumed to correctly answer an exercise only in two conditions: the student has mastered all the skills required by the exercise without slip, or the student does not mastered all the required skills but makes a successful guess. Due to the theories/assumptions chosen, traditional cognitive diagnosis models might perform well in some situations. However, the scope of applications are therefore restricted and excessive simplification of the cognitive process would lead to limited fitting ability. In practice, researchers need to choose suitable models from various choices ([24]) or even design their own model for specific usage, which is labor-intensive. Although the professional theories and assumptions provide valuable suggestions for cognitive diagnosis, we demand a new type of diagnosis model which requires less expert knowledge (i.e. automatically learnable), and provide accurate as well as interpretable diagnostic results that are easy to understand.

Artificial Neural Network. Techniques using artificial neural network have reached state-of-the-art in many areas, e.g., speech recognition [25], text classification [26] and image captioning [27]. There are also some educational applications such as question difficulty prediction [28], code education [29] and formula image transcribing [30]. However, using neural network for cognitive diagnosis is nontrivial as it performs poorly in parameter interpretation due to its inherent traits. To the best of our knowledge, deep knowledge tracing (DKT) [31] was the first attempt to model student learning process using recurrent neural network, followed by some variations [32], [33]. However, these knowledge tracing models paid more attention on modeling the changes of student states to predict students' scores, and did not explicitly model the effect of students' knowledge proficiencies in the learning process with an educational basis. Thus such models are unsuitable for cognitive diagnosis. Few works with neural network have high interpretability for student cognitive diagnosis. Towards this end, in this paper we propose a neural cognitive diagnosis (NeuralCD) framework which borrows concepts from educational psychology and combine them with fitting functions learned from data. NeuralCD could achieve both high accuracy and interpretation with neural network. Besides, the framework is general that can cover many tradition models, and at the same time easy for extension.

Knowledge Coverage Problem. Knowledge coverage is an important issue in cognitive diagnosis. Traditional cognitive diagnosis models mostly deal with scale-based tests where the amounts of exercises and knowledge concepts are small and the student responses are intact. The knowledge coverage is complete for each student in these conditions. However, when the amounts of exercises and knowledge concepts are large while the responses are sparse, which is the normal cases in nowadays intelligent education systems, the knowledge coverage problem becomes non-negligible. In traditional models such as IRT [10] and MIRT [11], [20], knowledge concepts are not considered. In models such as DINA [9], [17], DINO [34] and NIDA [35], knowledge concepts are considered to be independent. The diagnostic results of these models would be less reliable when knowledge coverage is incomplete. Some researches consider the relations among the proficiencies on different knowledge concepts. For example, the AHM [36] considers the hierarchical relation among knowledge concepts. De La Torre et al. [37] proposed HO-DINA which considered low-dimensional high order latent traits that affect the students' proficiencies on each knowledge concept. Liu et al. [38] proposed a FuzzyCDF model that a student's proficiencies on knowledge concepts are affected by his/her ability parameter. However, to the best of our knowledge, the knowledge coverage problem has not been explicitly studied yet in existing works.

3 Neural Cognitive Diagnosis

We first formally introduce cognitive diagnosis task. Then we describe the details of NeuralCD framework. After that, we design a specific diagnostic network NeuralCDM with traditional Q-matrix to show the feasibility of the framework. In the next two sections, we will introduce the superiority of NeuralCD framework in two aspects, i.e. generality and extensibility.

3.1 Task Overview

Suppose there are N Students, M Exercises and K Knowledge concepts at a learning system, which can be represented as $S = \{s_1, s_2, \ldots, s_N\}, E = \{e_1, e_2, \ldots, e_M\}$ and $K_n = \{k_1, k_2, \ldots, k_K\}$ respectively. Each student will choose some exercises for practice, and the response logs R are denoted as set of triplet (s, e, r) where $s \in S, e \in E$ and r is the score (transferred to percentage) that student s got on exercise e. In addition, we have Q-matrix (usually labeled by experts) $\mathbf{Q} = \{Q_{ij}\}_{M \times K}$, where $Q_{ij} = 1$ if exercise e_i relates to knowledge concept (abbreviated as KC) k_j and $Q_{ij} = 0$ otherwise.

Problem Definition *Given students' response logs R and the Q-matrix $\mathbf{Q}$, the goal of our cognitive diagnosis task is to mine students' proficiency on knowledge concepts through the student performance prediction process.*

3.2 Neural Cognitive Diagnosis Framework

Generally, cognitive diagnosis models are designed to simulate the results of students' exercise answering process where students use their cognition (e.g. knowledge, skills) to overcome the obstacles set in exercises. Thus for a cognitive diagnostic system, there are basically three elements need to be considered: *student factors*, *exercise factors* and the *interaction function* among them [39]. In this paper, we propose a general NeuralCD framework to address them by using multi-layer neural network modeling, which is shown in Figure 4. Specifically, for each response log, we use one-hot vectors of the corresponding student and exercise as input and obtain the diagnostic factors of the student and exercise. Then the interactive layers learn the interaction function among the factors and output the probability of correctly answering the exercise. After training, we get students' proficiency vectors as diagnostic results. Details are introduced as bellow.

Student Factors. Student factors characterize the traits of students, which would affect the students' response to exercises. As our goal is to mine students' proficiency on knowledge concepts, we do not use the latent trait vectors as in IRT and MIRT [11], which is not explainable enough to guide students' self-assessment. Instead, we design the student factors as explainable vectors similar to DINA, but has a major difference that they are continuous. Specifically, We use a vector F^s to characterize a student, namely *proficiency vector*. Each entry of F^s is continuous ([0,1]), which indicates the student's proficiency on a knowledge concept. For example, $F^s = [0.9, 0.2]$ indicates a high mastery on the first knowledge concept but low mastery on the second. F^s is got through the parameter estimation process.

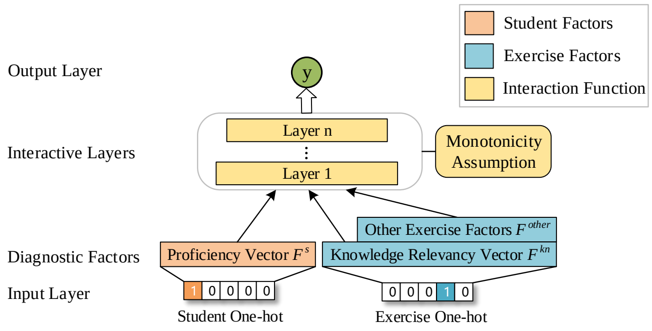


Fig. 4. Structure of NeuralCD framework.

Exercise Factors. Exercise factors are designed to characterize the traits of exercises. We divide exercise factors into two categories. The first indicates the relationship between exercises and knowledge concepts, which is fundamental as we need it to make each entry of F^s correspond to a specific knowledge concept for our diagnosis goal. We call it *knowledge relevancy vector* and denote it as F^{kn}. F^{kn} has the same dimension as F^s, with the ith entry indicating the relevancy between the exercise and the knowledge concept k_i. Each entry of F^{kn} is non-negative. F^{kn} is previously given (e.g., obtained from Q-matrix). Other factors are of the second type and are optional. Factors from IRT [19] and DINA [9] such as knowledge difficulty, exercise difficulty and discrimination can be incorporated if reasonable.

Interaction Function. Interaction function simulates how student factors interact with exercise factors to get the response results (e.g. right or wrong). We use artificial neural network to obtain the interaction function for the following reasons. First, the neural network has been proven to be capable of approximating any continuous function [40]. The strong fitting ability of neural network makes it competent for capturing relationships among student and exercise factors. Second, with neural network, the interaction function can be learned from data with few assumptions (that behind traditional models). This makes NeuralCD more general and can be applied in broad areas. Third, the framework can be highly extendable with neural network. For instance, extra information such as exercise texts can be integrated in with neural network (We will discuss its extendability in the following subsections.). Mathematically, we formulate the output of NeuralCD framework as:

$$y = \varphi_n(\ldots \varphi_1(F^s, F^{kn}, F^{other}, \theta_f)), \quad (1)$$

where φ_i denotes the mapping function of the ith MLP layer; F^{other} denotes factors other than F^s and F^{kn} (e.g., difficulty); and θ_f denotes model parameters of all the interactive layers.

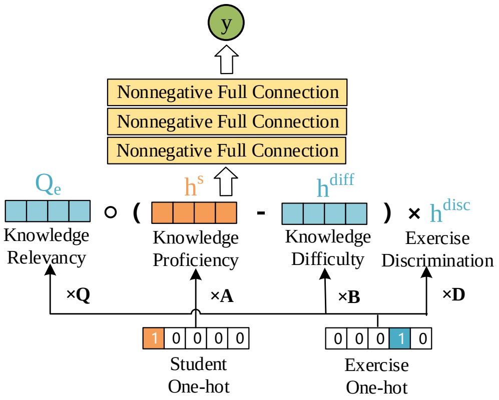


Fig. 5. Neural cognitive diagnosis model (NeuralCDM).

Interpretability Guarantee. Obtaining interpretable results is indispensable for cognitive diagnosis, as the diagnostic results are the basis for evaluating students' learning states and providing further personalized supports. However, due to some intrinsic characteristics, neural networks usually have poor performance on interpretation [41]. To solve this problem, We take two steps to ensure that the factors are explainable. The first step is to have the term $F^s \circ F^{kn}$ in the input layer in order to align each dimension of F^s to an knowledge concept specified by the corresponding dimension of F^{kn}. The second step is to utilize the monotonicity assumption, which is used in some IRT and MIRT models [11], to make the values in F^s variate in the same direction with y. Monotonicity assumption is general and reasonable in almost all circumstance, thus it has less influence on the generality of NeuralCD framework. The assumption is defined as follows:

Monotonicity Assumption *The probability of correct response to the exercise is monotonically increasing at any dimension of the student's knowledge proficiency.*

This assumption should be converted as a property of the interaction function. Intuitively, we assume student s to answer exercise e correctly. During training, the optimization algorithm should increase (or at least not decrease) the student's proficiency if the model output a wrong prediction (i.e., a value below 0.5). The increment of each knowledge proficiency is otherwise controlled by F^{kn} (step 1).

After introducing the structure of NeuralCD framework, we will next show some specific implementations. We first implement a basic model based on NeuralCD where knowledge relevancy vectors are directly get from pre-given Q-matrix to show the feasibility of NeuralCD (section 3.3). Then we discuss the generality of NeuralCD by showing that some traditional models can be regarded as its special cases (section 4). Further, we show the extendability of NeuralCD from content aspect (section 5.1) and structure aspect (section 5.2).

3.3 Neural Cognitive Diagnosis Model

Here we introduce a specific neural cognitive diagnosis model (NeuralCDM) under NeuralCD framework. Figure 5 illustrates the structure of NeuralCDM.

Student Factors. In NeuralCDM, each student is represented with a knowledge proficiency vector. The student factor F^s aforementioned is $\boldsymbol{h}^s$ here, and $\boldsymbol{h}^s$ is obtained by multiplying the student's one-hot representation vector $\boldsymbol{x}^s$ with a trainable matrix $\mathbf{A}$. That is,

$$\boldsymbol{h}^s = \text{sigmoid}(\boldsymbol{x}^s \times \mathbf{A}), \tag{2}$$

in which $\boldsymbol{h}^s \in (0,1)^{1\times K}, \boldsymbol{x}^s \in \{0,1\}^{1\times N}, \mathbf{A} \in \mathbb{R}^{N\times K}$.

Exercise Factors. As for each exercise, the aforementioned exercise factor F^{kn} is $\boldsymbol{Q}_e$ here, which directly comes from the pre-given Q-matrix:

$$\boldsymbol{Q}_e = \boldsymbol{x}^e \times \mathbf{Q}, \tag{3}$$

where $\boldsymbol{Q}_e \in \{0,1\}^{1\times K}$, $\boldsymbol{x}^e \in \{0,1\}^{1\times M}$ is the one-hot representation of the exercise. In order to make a more precise diagnosis, we adopt other two exercise factors: knowledge difficulty $\boldsymbol{h}^{diff}$ and exercise discrimination h^{disc}. $\boldsymbol{h}^{diff} \in (0,1)^{1\times K}$, indicates the difficulty of each knowledge concept examined by the exercise, which is extended from exercise difficulty used in IRT. $h^{disc} \in (0,1)$, used in some IRT and MIRT models, indicates the capability of the exercise to differentiate between those students whose knowledge mastery is high from those with low knowledge mastery. They can be obtained by:

$$\boldsymbol{h}^{diff} = \text{sigmoid}(\boldsymbol{x}^e \times \mathbf{B}), \mathbf{B} \in \mathbb{R}^{M\times K} \tag{4}$$

$$h^{disc} = \text{sigmoid}(\boldsymbol{x}^e \times \mathbf{D}), \mathbf{D} \in \mathbb{R}^{M\times 1} \tag{5}$$

where $\mathbf{B}$ and $\mathbf{D}$ are trainable matrices.

Interaction Function. The first layer of the interaction layers is inspired by MIRT models. We formulate it as:

$$\boldsymbol{x} = \boldsymbol{Q}_e \circ (\boldsymbol{h}^s - \boldsymbol{h}^{diff}) \times h^{disc}, \tag{6}$$

where $\circ$ is element-wise product. Following are two full connection layers and an output layer:

$$\boldsymbol{f}_1 = \phi(\mathbf{W}_1 \times \boldsymbol{x}^T + \boldsymbol{b}_1), \tag{7}$$

$$\boldsymbol{f}_2 = \phi(\mathbf{W}_2 \times \boldsymbol{f}_1 + \boldsymbol{b}_2), \tag{8}$$

$$y = \phi(\mathbf{W}_3 \times \boldsymbol{f}_2 + b_3), \tag{9}$$

where ϕ is the activation function. Here we use Sigmoid.

Different methods can be used to satisfy the monotonicity assumption. We adopt a simple strategy: restrict each element of $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ to be nonnegative. It can be easily proved that $\frac{\partial y}{\partial h^s_k}$ is nonnegative for each entry h^s_k in $\boldsymbol{h}^s$. (Please refer to Appendix A for detailed proof.) Thus monotonicity assumption is always satisfied during training.

The loss function of NeuralCDM is cross entropy between output y and true label r:

$$loss_{CDM} = -\sum_i (r_i \log y_i + (1 - r_i)\log(1 - y_i)). \tag{10}$$

After training, the value of $\boldsymbol{h}^s$ is what we get as diagnosis result, which denotes the student's knowledge proficiency.

4 Generality of NeuralCD

In this section we show that NeuralCD is a general framework which can cover many traditional cognitive diagnostic models. Using Eq. (6) as the first layer, we now show the close relationship between NeuralCD and traditional models, including MF, IRT and MIRT. Figure 6 gives an intuitive comparison between NeuralCD and these models.

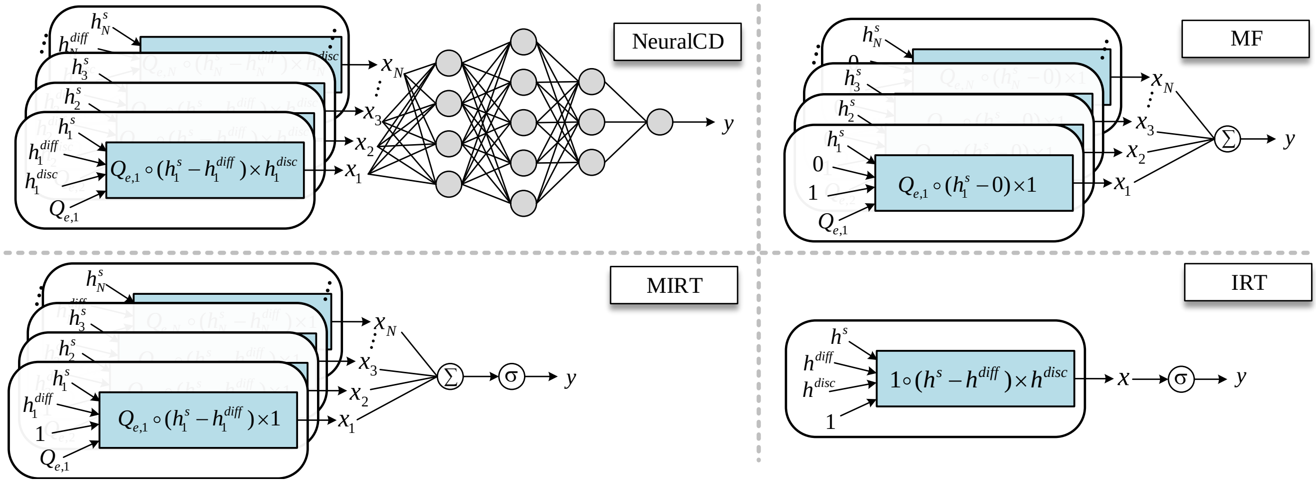


Fig. 6. Relation between NeuralCD and some traditional models.

MF [22]. In MF there are student and exercise latent vectors ($\boldsymbol{h}^s$ and $\boldsymbol{Q}_e$), and we take the basic formation of interaction function $\boldsymbol{Q}_e \cdot \boldsymbol{h}^s$ as an example. It should be noted the $\boldsymbol{Q}_e$ in MF is a learnable exercise parameter and cannot indicate the knowledge relevancy. By setting $\boldsymbol{h}^{diff} \equiv \mathbf{0}$ and $h^{disc} \equiv 1$, the output of the first layer is $\boldsymbol{x} = \boldsymbol{Q}_e \circ \boldsymbol{h}^s$. Then in order to work like MF, all the rest of layers need to do is to sum up ($\sum$) the values of each entry in $\boldsymbol{x}$, which is easy to achieve. Monotonicity assumption is not applied in MF approaches.

IRT [10]. Take the typical formation of IRT $y = \mathrm{Sigmoid}((h^s - h^{diff}) \times h^{disc})$ as example. First, set $Q_e \equiv 1$, and let $\boldsymbol{h}^s$ and $\boldsymbol{h}^{diff}$ be unidimensional, the output of the first layer is $x = (h^s - h^{diff}) \times h^{disc}$. Second, The multi-layer neural network in NeuralCD degenerates to a single Sigmoid activation function (σ). Monotonicity assumption could be achieved by limiting h^{disc} to be positive. Other variations of IRT (e.g., $y' = C + (1-C)y$ where C is guessing parameter) can be realized with a few changes.

MIRT [11]. One direct extension from IRT to MIRT is to use multidimensional trait vectors of exercises and students. Here we take the typical formation proposed in [20] as an example:

$$y = \frac{\mathrm{e}^{\mathbf{Q}_e \cdot \boldsymbol{h}^s - d_e}}{1 + \mathrm{e}^{\mathbf{Q}_e \cdot \boldsymbol{h}^s - d_e}}, \tag{11}$$

where $\mathbf{Q}_e$ is usually a low dimensional parameter learned from response data instead of the previously given knowledge relevancy vector. First, let $h^{disc} \equiv 1$, the output of the first layer given by Eq. (6) is $\boldsymbol{x} = \mathbf{Q}_e \circ (\boldsymbol{h}^s - \boldsymbol{h}^{diff})$. Second, make the multi-layers in NeuralCD degenerate to a summation ($\sum$) followed by a Sigmoid function (σ). Specifically, by Setting $\mathbf{W}_1 = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}, \boldsymbol{b}_1 = \mathbf{0}$ and $\phi(x) = \mathrm{Sigmoid}(x)$ in Eq. (7), we have $f_1 = \mathrm{Sigmoid}(\mathbf{Q}_e \cdot \boldsymbol{h}^s - d_e) = y$ (where $d_e = \boldsymbol{Q}_e \cdot \boldsymbol{h}^{diff}$). f_1 could be output without any more layers. Monotonicity assumption is not compulsory but can be realized if each entry of $\boldsymbol{Q}_e$ is restricted to be nonnegative.

5 Extensibility of NeuralCD

In this section we show that NeuralCD is an open framework that is easily extendable. We demonstrate two types of extensions, i.e., utilize extra content information and explore knowledge associations.

5.1 Content-based Extension

Due to the limitations of manually designed interaction functions, traditional cognitive diagnosis models mostly concentrate on numerical data such as student IDs, exercise IDs, response results (right or wrong) and the knowledge concepts of exercises in students' response logs. However, these information is not enough to characterize students' cognitive process which is quite complex. Many other information, such as the time duration of when the student answered exercises and the text content of exercises which have been proved to be highly related to some exercise features (e.g., difficulty, relevant knowledge concepts [28], [42]), are also relevant to the students' responses. Thus an extendable framework should be able to aggregate these extra information for better diagnostic results. Here we choose two typical types of information in exercise text, i.e., knowledge concepts and extra text-related factor, and demonstrate their utilizations.

5.1.1 Knowledge Extraction From Text Content

The first demonstration is to extract relevant knowledge concepts from exercise contents. In NeuralCDM we use manually-labeled Q-matrix to represent the knowledge relevancies of each exercise (a common practice in traditional works). However, manually-labeled Q-matrix may be deficient because of inevitable errors and subjective bias [39], [43]. For example, in Q-matrix, maybe only *'Equation'* is labeled for an equation solving exercise. However, *'Division'* is also required if we discover the existence of '÷' in the text. It is quite common that only target knowledge concepts are marked in Q-matrix while other relevant knowledge concepts are neglected. An optional strategy to solve this problem is to leverage the text content to refine the Q-matrix by discovering ignored knowledge concepts of exercises, which is feasible with the advantage of neural network. We denote this extended model as content enhanced NeuralCD with Q-matrix refinement (CNCD-Q), and present its structure in Figure 7.

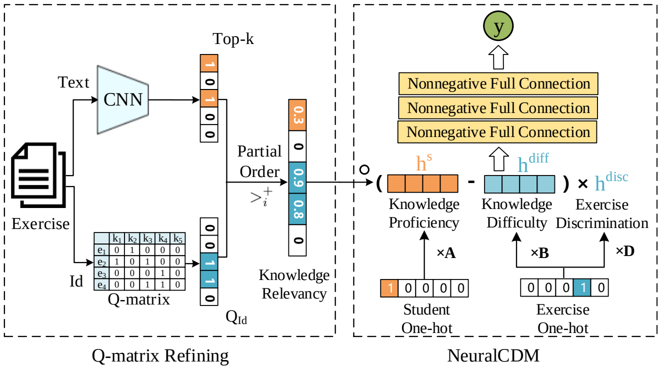


Fig. 7. CNCD-Q: Content enhanced NeuralCD with Q-matrix refinement.

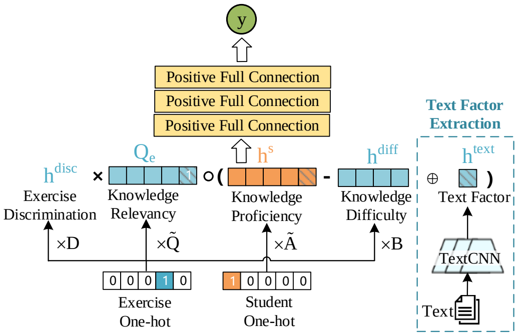


Fig. 8. CNCD-F: Content enhanced NeuralCD with text factor.

Specifically, we first pre-train a model that predicts knowledge concepts related to the input exercise. Lots of models that suitable for text processing can be used for this task [44]. As CNN (convolutional neural network) has advantage of extracting local information in text processing, it's able to capture important words from texts (e.g., words that are highly relative to certain knowledge concepts). Thus CNN is practically sufficient for our goal. Comparing or designing more advanced KC prediction models is beyond this work and we leave it for future research. The CNN network we use takes concatenated word2vec embedding of words in texts as input, and output the relevancy of each predefined knowledge concept (that has occurred in data) to the exercise (more details in section 6.2). Human-labeled Q-matrix is used as label for training. We define $V_i^k = \{V_{ij_1}, V_{ij_2}, \ldots, V_{ij_k}\}$ as the set of top-k knowledge concepts of exercise e_i outputted by the CNN.

Then we combine V_i^k with Q-matrix. Although there are defects in human-labeled Q-matrix, it still has high confidence. Thus we consider knowledge concepts labeled by Q-matrix are more relative than $\{k_j | k_j \in V_i^k \; and \; Q_{ij} = 0\}$. To achieve this, we adopt a pairwise Bayesian method as follows. For convenience, we define partial order $>_i^+$ as:

$$a >_i^+ b, \text{ if } Q_{ia} = 1 \text{ and } Q_{ib} = 0 \text{ and } b \in V_i^k, \quad (12)$$

and define the partial order relationship set as $D_V = \{(i, a, b) | a >_i^+ b, i = 1, 2, \ldots, M\}$. Following traditional Bayesian treatment, we assume $\tilde{\mathbf{Q}}$ follows a zero mean Gaussian prior with standard deviation σ of each dimension. To give Q-matrix labels higher confidence, we define $p(a >_i^+ b|\tilde{\mathbf{Q}}_i)$ with a pairwise logistic-like function:

$$p(a >_i^+ b|\tilde{\boldsymbol{Q}}_i) = \frac{1}{1 + e^{-\lambda(\tilde{Q}_{ia} - \tilde{Q}_{ib})}}. \quad (13)$$

The parameter λ controls the discrimination of relevance values between labeled and unlabeled knowledge concepts. The log posterior distribution over D_V on $\tilde{\mathbf{Q}}$ is finally formulated as:

$$\begin{aligned} \ln p(\tilde{\mathbf{Q}}|D_V) &= \ln \prod_{(i,a,b)\in D_V} p(a >_i^+ b|\tilde{\boldsymbol{Q}}_i) p(\tilde{\boldsymbol{Q}}_i) \\ &= \sum_{i=1}^{M}\sum_{a=1}^{K}\sum_{b=1}^{K} I(a >_i^+ b) \ln \frac{1}{1 + e^{-\lambda(\tilde{Q}_{ia} - \tilde{Q}_{ib})}} \quad (14) \\ &+ C - \sum_{i=1}^{M}\sum_{j=1}^{K} \frac{\tilde{Q}_{ij}^2}{2\sigma^2}, \end{aligned}$$

where C is a constant that can be ignored during optimization. Before using $\tilde{\mathbf{Q}}$ in NeuralCDM, we need to restrict its elements to the range $(0, 1)$, and set elements of concepts unlabeled or not predicted to 0. Thus, $Sigmoid(\tilde{\mathbf{Q}}) \circ \mathbf{M}$ is used to replace $\mathbf{Q}$ in NeuralCDM, where $\mathbf{M} \in \{0, 1\}^{M\times K}$ is a mask matrix, and $M_{ij} = 1$ if $j \in V_i^k$ or $Q_{ij} = 1; M_{ij} = 0$ otherwise. $\tilde{\mathbf{Q}}$ is trained together with the cognitive diagnostic model, thus the loss function is:

$$loss = -\ln p(\tilde{\mathbf{Q}}|D_V) + loss_{CDM}. \quad (15)$$

5.1.2 Factors Extraction From Text Content

Extracting relevant knowledge concepts is not the only way to make use of exercise text contents. Some other factors such as guess, slip [9], [17] and gaming [45] are also considered as reasons that influence students' performances. Similarly, other cognitively relevant information is contained in exercise texts. For example, the understanding of the text contents is the first stage of solving an exercise. Sometimes the expression used in the text can be confusing, although the knowledge concepts examined by the exercise might not be difficult. Here are two examples:

- E1: $2 \times 10 + 3 =$ ___.
- E2: Alice's speed is 1 m/s. John is twice as fast as Alice. John starts at 3m from the starting point of a straight runway and move forward. How far is John from the starting point after 10s?

Both E1 and E2 examine the mastery of *Addition* and *Multiplication*. The text of E1 is straightforward. However, E2 assumes a practical scenario that the student needs to understand and switch to the expression like E1 at first. although the difficulty of knowledge concept examined by E1 and E2 are close, the possibilities of correctly solving the exercises are influenced by the student's understanding of the text contents.

Traditional cognitive diagnosis models are difficult to aggregate this type of content information due to their limited extensibility. Here we show an example of considering the extra type of exercise factor, i.e. text factor, to extend our NeuralCD framework (denoted as CNCD-F). The architecture of CNCD-F is presented in Figure 8. Inside the dashed box is the text factor extraction process where given the exercise text, we first use a TextCNN [46] (which is efficient to process NLP texts) to get the text embedding $\boldsymbol{e} \in \mathbb{R}^{d_0}$. Then we translate it to text factor vector $\boldsymbol{h}^{text} \in \mathbb{R}^{d_1}$ through:

$$\boldsymbol{h}^{text} = \mathbf{W}_t \times \boldsymbol{e} + \boldsymbol{b}_t, \tag{16}$$

where $\mathbf{W}_t \in \mathbb{R}^{d_0 \times d_1}$ and $\boldsymbol{b}_t \in \mathbb{R}^{d_1}$ are trainable parameters, and we set $d_1 = 1$. Correspondingly, we extend the matrix $\mathbf{A}$ and $\mathbf{Q}$ to $\tilde{\mathbf{A}}(\in \mathbb{R}^{N \times (K+d_1)})$ and $\tilde{\mathbf{Q}}(\in \mathbb{R}^{M \times (K+d_1)})$ respectively. The extended dimensions represent the skills corresponding to the text factors (e.g. reading comprehension). $\widetilde{Q}_{\cdot,(K+d_1)} = 1$. The extended dimensions/factors are exhibited in Figure 8 using squares with oblique lines.

The overall modeling process is similar to NeuralCDM except that:

1) $\mathbf{Q}$ in Eq. (3) and $\mathbf{A}$ in Eq. (2) are replaced with $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{A}}$ respectively.
2) Eq. (6) is changed to:

$$\boldsymbol{x} = \boldsymbol{Q}_e \circ (\boldsymbol{h}^s - (\boldsymbol{h}^{diff} \oplus \boldsymbol{h}^{text})) \times h^{disc}. \tag{17}$$

After training, the values of the first K dimensions in $\boldsymbol{h}^s$ are the diagnosed proficiencies on KCs.

5.2 Knowledge-association Based Extension

In this subsection we demonstrate that the structure of NeuralCD is flexible to extend so as to adapt to different situations. Here we propose a knowledge-association based extension, in order to address the knowledge coverage problem in cognitive diagnosis. Normally in an intelligent educational system, there can be numerous knowledge concepts for a single subject. However, due to the limited questions answered (in a test or a short time span for diagnosis) by a student, this student's coverage of knowledge concept is usually quite low (analyses on real-world datasets are provided in Fig. 9 and Table 2). For convenience, we call these untested KCs of a student as *weak-KC* and those questions of which more than half of the contained KCs are weak-KCs as *weak-question*. In our previous models, the proficiency values of these weak-KCs are not reliable. For example, if student s never answered a question related to *Statistics*, the corresponding dimension in the student's knowledge proficiency vector ($\boldsymbol{h}^s$) would never be trained as the relevancy is always 0. To make the diagnostic model more robust, the ability of approximating the proficiencies on the weak-KCs of a student is necessary.

In this work we address this problem by considering the relations among knowledge concepts. Existing researches has revealed that knowledge concepts are not independent [36], [47]. Knowledge proficiencies are associated with each other, as well as knowledge difficulties of exercises. We formulated the proficiency of student s_i on knowledge k_j as $Prof(s_i, k_j) = \Phi(s_i, k_j, Rel^k)$, where the Rel^k denotes the knowledge relations, e.g. knowledge hierarchy [36], knowledge concept graph [48]. Considering that these explicit knowledge relations require expert knowledge and is not always available, we here provide an example of modeling implicit knowledge relations purely from response logs. When explicit knowledge relations are absent, a normal practice is to calculate pairwise constraints between each pair of item, which requires large scale of parameters, especially when the constraints are student specific or exercise specific. Instead, we adopt a representation based method to implicitly model the knowledge association, and the extended framework is called KaNCD.

Specifically, we do not directly learn the matrix $\mathbf{A}$ (Eq. (2)). Each student (s_i) and each KC (k_j) are represented with a d-dimensional ($d < K$) latent vector respectively ($\boldsymbol{l}_i^s$ and $\boldsymbol{l}_j^k$). Each element in $\mathbf{A}$ is the result of operation between the corresponding student and KC vectors. Here we regard the d dimensions as higher order skills behind the pre-defined knowledge concepts (inspired by [37]). The values of each dimension in $\boldsymbol{l}_j^k$ denotes its preference for each high order skill, thus $\boldsymbol{l}_i^s$ is filtered by:

$$\boldsymbol{a}_1 = \boldsymbol{l}_i^s \circ \boldsymbol{l}_j^k, . \tag{18}$$

Then the proficiency of s_i on k_j is calculated as the weighted sum of the filtered latent traits in $\boldsymbol{a}_1$ with Sigmoid activation (in Eq. (2), to limit the proficiency to (0, 1)):

$$A_{i,j} = \boldsymbol{W}_{a2} \times \boldsymbol{a}_1 + b_{a2}, , \tag{19}$$

where $\boldsymbol{W}_{a2} \in \mathbb{R}^{d \times 1}$ and $\boldsymbol{b}_{a2} \in \mathbb{R}$ are trainable parameters.

Following this way, we apply the same process to the knowledge difficulty matrix $\mathbf{B}$ with the consideration of knowledge associations. Each exercise (e_i) is represented with a d-dimensional latent vector ($\boldsymbol{l}_i^e$). The difficulty of e_i on KC k_j is calculated as:

$$\boldsymbol{b}_1 = \boldsymbol{l}_i^e \circ \boldsymbol{l}_j^k, \tag{20}$$

$$B_{i,j} = \boldsymbol{W}_{b2} \times \boldsymbol{b}_1 + b_{b2}, , \tag{21}$$

where $\boldsymbol{W}_{b2} \in \mathbb{R}^{d \times 1}$ and $b_{b2} \in \mathbb{R}$ are trainable parameters.

Overall, the process of KaNCD is as follows. We compute $A_{i,j} (i = 1, \ldots, N, j = 1, \ldots, K)$ and $B_{i,j} (i = 1, \ldots, M, j = 1, \ldots, K)$ to get $\mathbf{A}$ and $\mathbf{B}$. Then we feed the training data with Eq. (2)˜ (9) and train the parameters (including latent vectors of students, exercises and KCs) with the same loss function as Eq. (10). After training, the student proficiencies on KCs can be inferred with Eq. (18)˜ (19) and Eq. (2). In Appendix C, we provide comparisons between KaNCD and some existing relevant models including AHM [36], HO-DINA [37] and FuzzyCDF [38].

5.3 Discussion

We have introduced the details of NeuralCD framework and showed special cases of it. NeuralCD is a general framework that could get accurate and explainable diagnostic results. Meanwhile, the framework has better extensibility than traditional models, such as aggregating extra information (e.g. text content) and improving framework structure to solve specific problems (e.g. knowledge coverage problem). 1) It's necessary to point out that the student's proficiency vector F^s and exercise's knowledge relevancy vector F^{kn} are basic factors needed in NeuralCD framework. Additional factors such as exercise discrimination can be integrated into if reasonable. 2) The formation of the first interactive layer is not limited, but it's better to contain the term $F^s \circ F^{kn}$ to ensure that each dimension of F^s corresponds to a specific knowledge concept. 3) The nonnegative full connection is only one of the strategies that implement monotonicity assumption. More sophisticated network structures can be designed as the interaction layers. For example, recurrent neural network or memory network may be used to capture the time characteristics of the student's learning process. 4) As for the model output, we focus on objective exercises where responses are correct (1) or incorrect (0) in this paper, therefore the outputs of NeuralCD models are the probabilities that the students would correctly answer the exercises. In fact, NeuralCD models can also handle exercises with non-dichotomous responses. For example, for exercises with continuous response labels (e.g., scoring rates in range (0,1)), the model outputs are predicted scores; for exercise with polytomous possible scores, the output layer can be changed to output a classification vector which indicates the predicted class (i.e., score). Better measures could be considered into the modeling, such as multiple independent components or multiple sequential steps in an exercise [11], and we leave it for future research.

6 Experiments

In this section, we conducted extensive experiments to demonstrate the effectiveness of our NeuralCD models from various aspects: (1) the student performance prediction task against baselines; (2) the model analysis about the interpretation of diagnostic results; (3) the visualization of learned embeddings of knowledge concepts, exercises and students.

6.1 Dataset Description

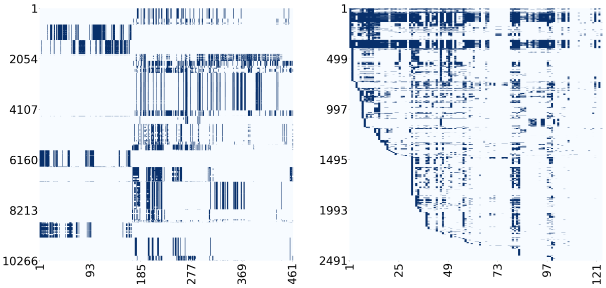
We used two real-world datasets in the experiments, i.e., Math and ASSIST. Math dataset supplied by iFLYTEK Co., Ltd. was collected from the widely-used online learning system Zhixue[1], which contains mathematical exercises and logs of high school examinations. ASSIST (ASSISTments 2009-2010 ”skill builder”) is an open dataset collected by the ASSISTments online tutoring systems [49], which only provides student response logs and knowledge concepts[2]. We chose the public corrected version that eliminates the duplicated data issue pointed out by previous work [50]. Table 1 summarizes basic statistics of the datasets.

Preprocess. We filtered out students with less than 30 and 15 response logs for Math and ASSIST respectively to guarantee that each student has enough data for diagnosis. Therefore for dataset Math, we got 10,268 students, 2,507 exercises with 497 knowledge concepts for diagnostic network, and the remaining exercises with knowledge concepts not appearing in logs were used for the Q-matrix refining part of CNCD-Q; for dataset ASSIST, we got 2,493 students, 17,671 exercises and 123 knowledge concepts. We performed a 80%/20% train/test split of each student's response log. As for ASSIST, we divided the response logs in the same way with Math, but CNCD-Q and CNCD-F were not evaluated on this dataset as exercise text was not provided. All models were evaluated with 5-fold cross validation.

Verification of static knowledge proficiency. Students' knowledge proficiencies are stable in Math as the dataset is composed of logs from examinations. However, a student's proficiency on a knowledge concept in ASSIST may change as he will be continually given exercises of that concept until meeting certain criterion (e.g., answering 3 relevant exercises correctly in a row). To analyze whether static models (e.g., NeuralCD and static traditional models) are suitable to apply on ASSIST, we compare two metrics between Math and ASSIST. The first metric is the average amount of logs that each student toke for each knowledge concept:

TABLE 1
Dataset summary.

Dataset	Math	ASSIST
#Students	10,268	4,163
#Exercises	917,495	17,746
#Knowledge concepts	1,488	123
#Response logs	864,722	324,572
#Knowledge concepts per exercise	1.53	1.19
AVG$_{\#log}$	2.28	8.05
STD$_{\#log>1}$	0.305	0.316



(a) Math (b) ASSIST

Fig. 9. Knowledge coverage of student response logs.

$$\text{AVG}_{\#log} = \frac{\sum_i^N \sum_j^K Log(i,j)}{\sum_i^N \sum_j^K I(Log(i,j) > 0)}, \tag{22}$$

where $Log(i,j)$ is the amount of exercises student s_i answered that related to knowledge concept k_j. Further, another metric is the mean standard deviation of scores r_{ij} that $Log(i,j) > 1$ as:

$$\text{STD}_{\#log>1} = \underset{s_i \in S}{\text{mean}}(\underset{\substack{k_j \in K_n, \\ Log(i,j)>1}}{\text{mean}} (std_{ij})), \tag{23}$$

where std_{ij} is the standard deviation of scores that student s_i got for exercises related to knowledge concept k_j. As listed in Table 1, although ASSIST has a much larger AVG$_{\#log}$ than Math, their STD$_{\#log>1}$ are close. Therefore, it is reasonable to assume that the knowledge states of students in ASSIST are also stable, and our static NeuralCD models and baselines are applicable for both datasets. There will be more discussions in Model Interpretation.

Knowledge coverage visualization. To illustrate the knowledge coverage of each student's full response logs (before splitting the data into training and testing sets), we draw heat maps for the two datasets. In Figure 9, the horizontal axis and vertical axis denote the KC ID and student ID respectively. The blue color means that the corresponding student has response logs that related to the corresponding KC (i.e has answered relevant exercises), and the color would be white otherwise. We could observe that the knowledge coverage of student response logs are quite low on both datasets. This confirms the statement we proposed

[1] https://www.zhixue.com
[2] https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data/skill-builder-data-2009-2010

before that the knowledge concept coverage problem is a common phenomenon and needs attention.

6.2 Experimental Setup

The dimensions of the full connection layers (Eq. (7) $\sim$ (9)) were 512, 256, 1 respectively, and Sigmoid was used as activation function for all of the layers. We set hyperparameters $\lambda = 0.1$ (Eq. (13)) and $\sigma = 1$ (Eq. (14)). For k in top-k knowledge concepts selecting, we used the value that make the predicting network reach 0.85 recall. That is, in our experiment, $k = 20$. We initialized the parameters with *Xavier* initialization [51], which fill the weights with random values sampled from $\mathcal{N}(0, std^2)$, where $std = \sqrt{\frac{2}{n_{in}+n_{out}}}$. n_{in} is the number of neurons feeding into the weights, and n_{out} is the number of neurons the results is fed to. As for the monotonicity assumption, the implementation is not limited to certain method. In our experiments, after each parameter updating using a batch of data, we clipped all the elements into full connection weights (i.e., $\boldsymbol{W}_1, \boldsymbol{W}_2$ and $\boldsymbol{W}_3$ in Eq. (7)$\sim$(9)) to $[0, +\infty)$.

In CNCD-Q, the CNN contained 3 convolutional layers followed by a full connection output layer. MaxPooling was used after 1st and 3rd convolutional layers. The channels of convolutional layers were 400, 200, 100, and kernel sizes were set to 3, 4, 5 respectively. We adopted ReLu activation function for convolution layers and Sigmoid for the output layer. Multi-label binary cross entropy was used as loss function for training the CNN.

In CNCD-F, the TextCNN architecture was basically the same with [46]. We set 150, 150 and 150 filters with kernel size 3, 4 and 5 respectively. Average pooling was used after each filter. We adopted ReLu activation function for convolution layers and Sigmoid for the output layer. The dropout of the output layer was set to 0.5.

To evaluate the performance of our NeuralCD models, we compare them with previous approaches, i.e., DINA, IRT, MIRT and PMF. All models are implemented by PyTorch using Python, and all experiments are run on a Linux server with four 2.0GHz Intel Xeon E5-2620 CPUs and a Tesla K20m GPU.

6.3 Student Performance Prediction

The performance of a cognitive diagnosis model is difficult to evaluate as we cannot obtain the true knowledge proficiency of students. As diagnostic result is usually acquired through predicting students' performance in most works, performance on these prediction tasks can indirectly evaluate the model from one aspect [38]. In order to sufficiently evaluate the models, we applied two methods to split the datasets. The first is a random split where 80% of each student's response logs were randomly chosen as training set, which is the normal practice in student performance prediction task. The other is a weak-coverage split which is designed in the purpose of better comparing model performances in case when the KC coverages of students' training data are low. Therefore, we designed a split algorithm which assign more weak-questions into the test set while keep the train/test ratio (80%/20%) unchanged. The details are showed in Appendix A. We design this algorithm in order to control that the changes of model performances result from the differences of weak response proportions in test sets instead of the differences of data size for training. The proportions of responses to weak-questions in test sets are showed in Table 2, where we can observe that the weak response proportions of weak coverage split are significantly higher than random split.

TABLE 2
Proportion of response to weak-questions in test sets.

	Random Split		Weak Coverage Split	
	Math	ASSIST	Math	ASSIST
Weak response proportion	0.315	0.040	0.876	0.857

Considering that all the exercises we used in our data are objective exercises, we use evaluation metrics from both classification aspect and regression aspect, including accuracy, RMSE (root mean square error) [52] and AUC (area under the curve) [53]. The experimental results of normal scenario and low-knowledge-coverage scenario are shown in Table 3 and Table 4 respectively, where the error bars after '±' are the standard deviations of 5-fold cross validation runs for each model.

Normal Scenario. From Table 3, we have the following observations. First, the NeuralCD models outperform almost all the other baselines on both datasets, indicating the effectiveness of our framework. Second, the better performance of content based extensions (CNCD-Q and CNCD-F) over NeuralCDM proves that extra information more than response logs, such as exercise text contents, is beneficial to cognitive diagnosis. Moreover, the Q-matrix refining method we propose is effective, and also demonstrates the importance of fine estimated knowledge relevancy vectors for cognitive diagnosis. The results of CNCD-F show that text factors indeed plays an import role in the cognitive process. Third, comparing KaNCD and NeuralCDM, we could observe significant improvements, which is benefited from the modeling of knowledge associations. The improvements of these extended models proves the future potential of NeuralCD framework.

Weak-Knowledge-coverage Scenario. From Table 4, we have the following observations. First, the model performances drop significantly compared to normal scenario, indicating that the low coverage problem of KCs has considerable negative effects on the diagnostic results. Second, the NeuralCD models still perform better than baselines, which demonstrates the superiority of our NeuralCD framework. Third, the improvements of extended NeuralCD models over NeuralCDM are more than those in normal scenario. In other words, the falls in the model performances are smaller than NeuralCDM, which proves that the extension methods increase the tolerance of NeuralCD to the problem of low KC coverage. The better performances of CNCD-Q and CNCD-F than KaNCD indicate that extra information (e.g. exercise text) has greater positive effect to cognitive diagnosis than barely improving the model structure.

6.4 Interpretability of Diagnostic Results

The student performance prediction task is not sufficient to evaluate the cognitive diagnosis models, as the inter-

TABLE 3
Experimental results of student performance prediction with random split.

	Math (Random Split)			ASSIST (Random Split)		
Model	Accuracy	RMSE	AUC	Accuracy	RMSE	AUC
DINA	0.593±.001	0.487±.001	0.686±.001	0.650±.001	0.467±.001	0.676±.002
IRT	0.782±.002	0.387±.001	0.795±.001	0.674±.002	0.464±.002	0.685±.001
MIRT	0.793±.001	0.378±.002	0.813±.002	0.701±.002	0.461±.001	0.719±.001
PMF	0.763±.001	0.407±.001	0.792±.002	0.661±.002	0.476±.001	0.732±.001
NeuralCDM	0.792±.002	0.378±.001	0.820±.001	0.719±.008	0.439±.002	0.749±.001
KaNCD	**0.805±.001**	**0.368±.002**	0.836±.001	**0.732±.001**	**0.424±0.001**	**0.767±0.001**
CNCD-Q	0.804±.001	0.371±.002	0.835±.002	-	-	-
CNCD-F	0.802±.001	0.370±.002	**0.840±.002**	-	-	-

TABLE 4
Experimental results of student performance prediction with weak-coverage split.

	Math (Weak-coverage Split)			ASSIST (Weak-coverage Split)		
Model	Accuracy	RMSE	AUC	Accuracy	RMSE	AUC
DINA	0.223±.001	0.502±.002	0.560±.001	0.471±.001	0.490±.001	0.588±.002
IRT	0.624±.001	0.467±.001	0.638±.002	0.657±.001	0.464±.001	0.633±.002
MIRT	0.620±.001	0.583±.001	0.572±.001	0.637±.001	0.505±.001	0.612±.001
PMF	0.596±.001	0.585±.002	0.625±.001	0.625±.001	0.478±.002	0.730±.003
NeuralCDM	0.735±.002	0.432±.002	0.649±.001	0.710±.003	0.455±.001	0.633±.002
KaNCD	0.736±.001	0.430±.001	0.691±.001	**0.720±.001**	**0.435±.001**	**0.732±.002**
CNCD-Q	**0.748±.001**	**0.418±.001**	0.725±.001	-	-	-
CNCD-F	0.741±.001	0.419±.001	**0.732±.001**	-	-	-

pretability is an essential part of cognitive diagnostic results. Specifically, we adopt Degree of Agreement (DOA) [54] as the evaluation metric for the diagnosed student states ($\boldsymbol{h}^s$). This metric is based on the intuition that if student a has a better mastery on knowledge concept k than student b, then a is more likely to answer exercises related to k correctly than b [55]. For knowledge concept k, $DOA(k)$ is formulated as[3]:

$$DOA(k) = \frac{1}{Z_1}\sum_{a=1}^{N}\sum_{b=1}^{N} I(F_{ak}^s > F_{bk}^s)\frac{\sum_{j=1}^{M} I(Q_{jk}=1) \wedge J(j,a,b) \wedge I(r_{aj} > r_{bj})}{Z_0}, \tag{24}$$

$$Z_0 = \sum_{j=1}^{M} I(Q_{jk}=1) \wedge J(j,a,b) \wedge I(r_{aj} \neq r_{bj}), \tag{25}$$

$$Z_1 = \sum_{a=1}^{N}\sum_{b=1}^{N} I(F_{ak}^s > F_{bk}^s) I(Z_0 > 0), \tag{26}$$

where F_{ak}^s is the proficiency of student a on knowledge concept k. $I(Statement) = 1$ if $Statement$ is true and $I(Statement) = 0$ otherwise. $J(j,a,b) = 1$ if both student a and b did exercise j and $J(j,a,b) = 0$ otherwise. It should be noted that if $Z_0 = 0$, the corresponding (a,b,k) triplet is excluded from the calculation of DOA. We average $DOA(k)$ on all knowledge concepts to evaluate the quality of diagnostic result (i.e., knowledge proficiency acquired by models). It should be noted that although the DOA we define in Eq. (24) ignores the synergism when an exercise contains multiple KCs, it does reflect an interpretable cognitive phenomenon to some extent.

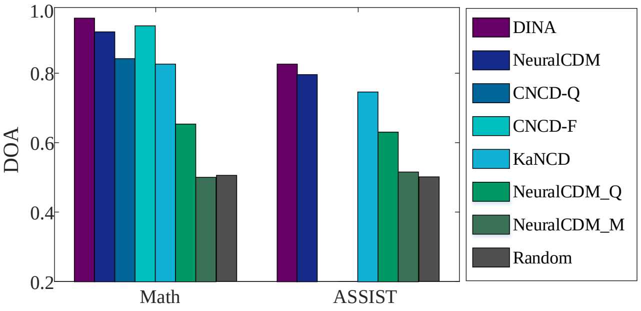


Fig. 10. DOA results of models on two datasets.

Among traditional models, we only compare with DINA, since for IRT, MIRT and PMF, there are no clear correspondence between their latent features and knowledge concepts. Besides, we conduct experiments on two reduced NeuralCDM models. In the first reduced model (denoted as NeuralCDM_Q), knowledge relevancy vectors are estimated during unsupervised training instead of getting from Q-matrix. While in another reduced model (denoted as NeuralCDM_M), monotonicity assumption is removed by eliminating the nonnegative restriction on the full connection layers. These two reduced models are used to demonstrate the importance of fine-estimated knowledge relevancy vector and monotonicity assumption respectively. Furthermore, we conduct an extra experiment in which students' knowledge proficiencies are randomly estimated, and compute the DOA for comparison.

Figure 10 presents the experimental results, from which we can observe several conclusions. First, the DINA model, which is indeed a highly explainable cognitive diagnosis

[3] This formula corrects the mistake in [16].

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2022.3201037

model, gets the highest DOAs on two datasets. This is because the type of interpretation of cognitive diagnosis results that DOA measures is highly agree with the intuition behind DINA. However, considering the results of student performance prediction task, the cognitive diagnosis results of DINA are still not suitable as diagnosis report and for further learning assist. Second, DOAs of NeuralCD models are in the order of CNCD-F >NeuralCDM >CNCD-Q >KaNCD. Although lower than DINA, their DOAs are still significantly higher than Random, which reveals their strong interpretability. The reason that the DOA of CNCD-Q is lower than NeuralCDM is that there are more relevant KCs of each exercise in CNCD-Q, while the increasing synergism of KCs is not measured by DOA. The lower DOA of KaNCD is the result of knowledge associations modeled. The DOA of CNCD-F is even higher than NeuralCDM, which shows that text factor is a good supplement of the cognitive model. For example, when student A, who has higher knowledge proficiency than student B, get lower scores on some exercises than B, the reason might be that A misunderstood the texts of these exercises. Since the influence of exercise texts is transferred to text factor, CNCD-F gets higher DOA. Third, comparing NeuralCDM_Q and NeuralCDM_M with NeuralCDM, there are noticeable drops of DOA, which indicates that both information from Q-matrix and monotonicity assumption are important for getting interpretable diagnosis results (knowledge proficiency vectors). Besides, NeuralCDM and KaNCD perform much better on Math than on ASSIST. This is mainly due to the contradictions in logs, i.e., a student may answer some exercises containing knowledge concept k_j correctly while others containing k_j wrong (reasons may be the change of knowledge proficiency, or other knowledge concepts contained by the exercises). As showed in Table 1, ASSIST has much larger $\mathrm{AVG}_{\#log}$ and slightly higher $\mathrm{STD}_{\#log>1}$ than Math dataset, which makes more contradictions in logs. Longer logs with more contradictions would decrease DOA.

6.5 Analyses On Content-based Extensions

Exercise text contents provide useful supplementary information for cognitive diagnosis. In Table 5 we present an example from Math dataset that demonstrates how CNCD-Q leverages text content to refine the Q-matrix and therefore improve the diagnostic performance. From the table we could observe that there is only one KC, i.e., "Number and formula" labeled by Q-matrix, which is inaccurate to describe the knowledge concepts tested by this exercise. Such inaccuracy could results from multiple reasons, such as experts' focus on main knowledge concepts of an exercise, or lack of systematically organizing the knowledge concepts. From the text content, CNCD-Q predicts KCs that are relevant to the exercise, such as "Positional relationship between a line a plane in space". The relevancies of the KCs labeled by Q-matrix and predicted bv CNCD-Q are discriminated (0.87 and 0.45±0.01 respectively). As we fix the total number of predicted KCs (i.e., 20), and limited by the performance of knowledge prediction component in CNCD-Q, some predicted KCs might not be relevant to the exercise (KCs without underline). Currently, CNCD-Q could not differentiate the relevancies of the predicted KCs, and this requires future improvement.

TABLE 5
An example of Q-matrix refinement.

Exercise text content	Let m, n be two different lines, α, β, γ be three different planes. Of the following propositions, select the correct ones: ____ (1) If $m \parallel n$ and $n parallel \alpha$, then $m \parallel \alpha$ or $m \subset \alpha$. (2) If $m \parallel \alpha, n \parallel \alpha, m \subset \beta, n \subset \beta$, then $\alpha \parallel \beta$. (3) If $\alpha \perp \gamma, \beta \perp \gamma$, then $\alpha \parallel \beta$. (4) If $\alpha \parallel \beta, \beta \parallel \gamma, m \perp \alpha$, then $m \perp \gamma$.
KCs labeled by Q-matrix (relevancy: 0.87)	"Number and formula"
KCs predicted by CNCD-Q (relevancy: 0.45±0.01)	"Positional relationship between a line and a plane in space", "Positional relationship between planes", "Positional relationship between lines in space", "Basic properties and applications of plane", "Judgment of perpendicularity between planes", "Judgment of perpendicularity between a line and a plane", "Judgment of parallelism between a line and a plane", "Properties of perpendicularity between planes", "Properties of perpendicularity between a line and a plane", "Properties of parallel lines and planes", "Angle between a line and a plane", "Angle formed by skew lines", "Skew lines", "Set", "Simple polyhedron", "Side area, surface area and volume", "Distance among points, lines and faces", "Full quantifier and existential quantifier", "Judgment of necessary, sufficient and sufficient conditions"

An additional effect that content-based extension brings is better tolerance of the knowledge coverage problem. From the results in Table 4, we could observe that although CNCD-Q and CNCD-F are not designed specifically for weak-knowledge-coverage scenario, they still perform well, and sometimes even better than KaNCD, in this scenario. The reason could be that in CNCD-Q, the predicted knowledge concepts significantly increases the knowledge coverage of students' response logs. As for CNCD-F, the extended dimension in $\boldsymbol{h}^s$ that corresponding to text factor $\boldsymbol{h}^{text}$ serves as an important indicator of the students' abilities, and affect the students' performance on all exercises regardless the knowledge coverage problem. In summary, when exercise text contents are available, taking advantage of the content information is a better solution to overcome the knowledge coverage problem.

6.6 Embedding Visualization in KaNCD

Using the trained KaNCD model on ASSIST, we visualize the embedding vectors of knowledge concepts ($\boldsymbol{l}^k$) by projecting them to 2-D points with t-SNE [56]. Figure 11 shows the visualization result of knowledge concept embeddings. We group the knowledge concepts into 7 clusters according to their positions and differentiate them with different colors. The clusters reveals some reasonable results. For example, the knowledge concepts in the first cluster are basically about basic algebra. Some relevant knowledge concepts are close (e.g. *88 Area Rectangle* and *90 Area Triangle*). These knowledge associations are implicitly captured by the embeddings learn from response logs, which helps improve the diagnostic results on weak-KCs.

Discussion. In KaNCD, we also represent each student and exercise with vector embeddings, which should also

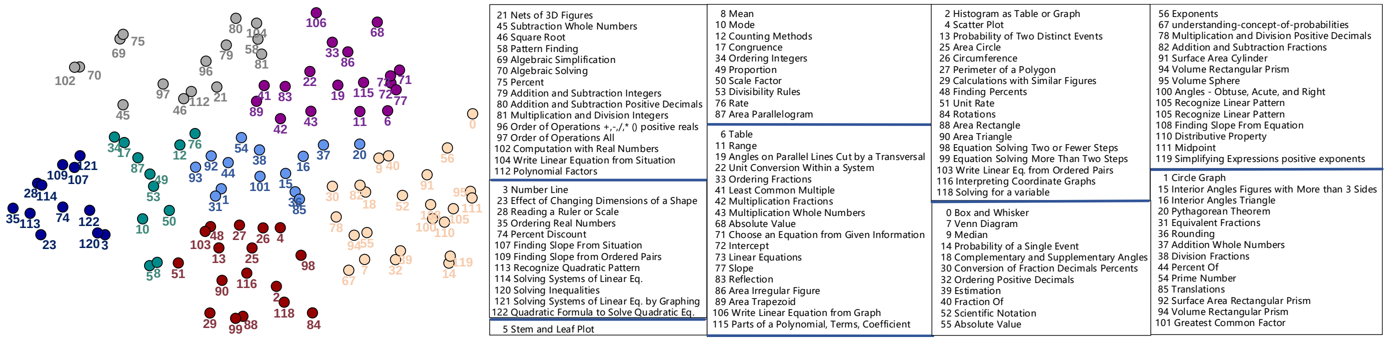
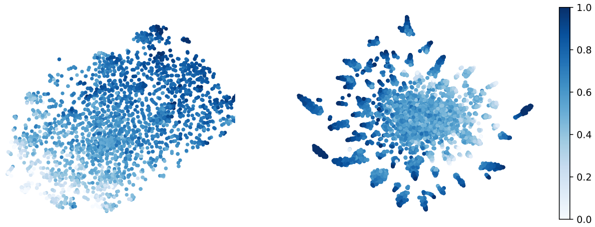


Fig. 11. t-SNE visualization of knowledge concept embeddings.



(a) Student embedding.

(b) Exercise embedding.

Fig. 12. t-SNE visualizations of student and exercise embeddings.

capture important information. Some information could probably be reflected by the distribution or the distances among embeddings. Thus, we try to visualize the student embeddings ($\boldsymbol{l}^s$) and exercise embeddings ($\boldsymbol{l}^e$) using t-SNE method. To ensure enough training data, we only visualize the embeddings of students with more than 15 responses and exercises with more than 10 responses. Finally we get Figure 12(a) and Figure 12(b), where we could observe that the points follow certain patterns instead of scatter randomly. Although it is difficult to figure out the reasons of why they appear such patterns, we do find an interesting phenomenon. If we color the student/exercise points with their correct rates (i.e., the correct rate of the student's all responses and the correct rate of all the responses of the exercise), we could find that the correct rates gradually increase from bottom-left to top-right in Figure 12(a) and from inside to outside in Figure 12(b). This indicate that the low-dimensioned student/exercise embeddings have captured the information about correct rates. There might be some other information (e.g. relations with knowledge concepts), and we leave this for future research.

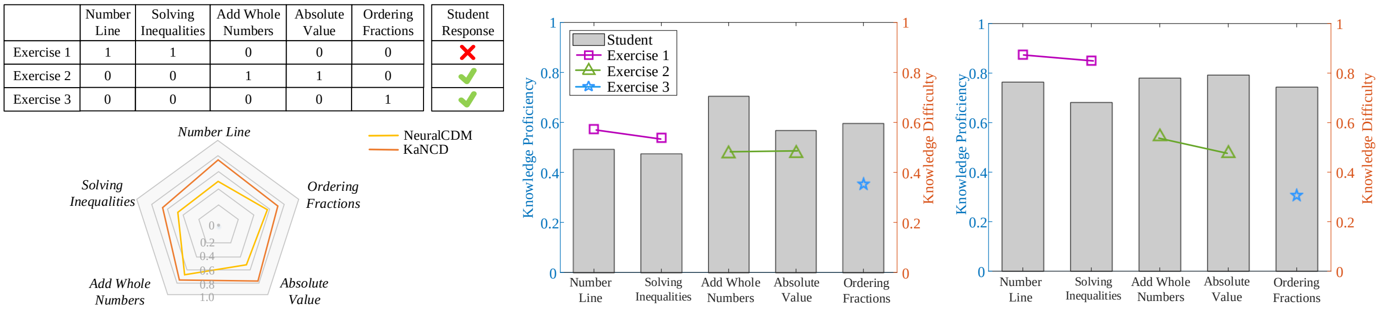
6.7 Case Study

In Figure 13 we present an example of a student's diagnostic results by NeuralCDM and KaNCD on the public dataset ASSIST. The table in Figure 13(a) shows the Q-matrix of three of the exercises answered by the student and the corresponding response results. The radar chart behind Q-matrix table presents the diagnosed results of the student on the corresponding knowledge concepts. As shown in the radar chart, NeuralCD models could provide explainable diagnosis reports that indicate students' proficiencies on different knowledge concepts. Then, we compare the knowledge difficulties and the student's proficiencies. As shown in Figure 13(a) and Figure 13(b), where the bars represent the student's proficiencies on each relevant knowledge concepts and the lines with different colors and markers represent the knowledge difficulties of relevant knowledge concepts. We can observe from the both of the subfigures that when the student answered correctly, the diagnosed knowledge proficiencies tend to be higher than knowledge difficulties. For example, in Figure 13(a), Exercise 3 requires the mastery of *'Ordering Fraction'* and corresponding difficulty is 0.35. The student answered it correctly, so the diagnosed proficiency is 0.6. Both knowledge difficulty ($\boldsymbol{h}^{diff}$) and knowledge proficiency ($\boldsymbol{h}^s$) in our models are explainable as expected. Furthermore, we plot all the proficiencies diagnosed by NeuralCDM and KaNCD in Figure 6.6. The knowledge concepts are resorted so that the first 45 concepts appeared in the student's training data and the last 78 knowledge concepts are weak-KCs. We can observe that although there are variations of NeuralCDM on the first 45 concepts, the proficiencies on weak-KCs are close to 0.5, which remain their initialized values and never trained. On the other hand, the proficiencies provided by KaNCD have reasonable variations on all the knowledge concepts and their average (0.65) is closer to the student's overall correct rate (0.75), which again confirms the observation in 6.6 that the low-dimensional student embeddings have captured the information about students' correct rates.

With a deeper observation of the Figure 6.6, we could find that the models provide different diagnostic results. This arouse a question: which result should we trust, or how we use the diagnosed results appropriately? As our NeuralCD framework and implemented models are data-driven, proficiencies diagnosed by different trained models (e.g., with different data or hyper-parameters), are not strictly guaranteed to be comparable. The explanation and usage of diagnosed proficiencies should be together with the estimated exercise attributes (e.g., difficulty), as they are in the same parameter scale. We leave the comparison of proficiencies from different trained models and the validation of their credibility for future exploration.

7 Conclusion and Discussion

In this paper, we proposed a neural cognitive diagnostic framework, NeuralCD framework, for students' cognitive diagnosis. Specifically, we first discussed fundamental student and exercise factors in the framework, and placed a monotonicity assumption on the neural-network-based framework to ensure both accuracy and interpretability of



	Number Line	Solving Inequalities	Add Whole Numbers	Absolute Value	Ordering Fractions	Student Response
Exercise 1	1	1	0	0	0	✗
Exercise 2	0	0	1	1	0	✓
Exercise 3	0	0	0	0	1	✓

(a) Q-matrix and diagnosed results by NeuralCDM and KaNCD.

(b) Difficulties and diagnosed proficiencies by NeuralCDM.

(c) Difficulties and diagnosed proficiencies by KaNCD.

Fig. 13. Diagnosed results of two students and their relation with knowledge difficulties on ASSIST.

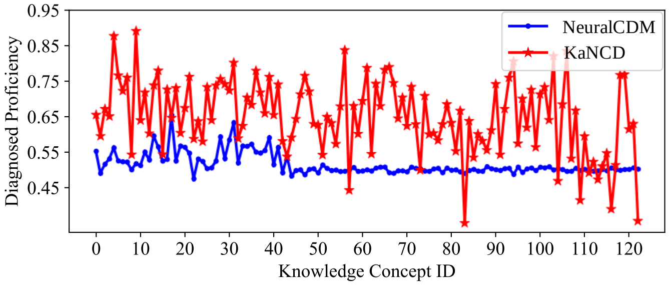


Fig. 14. Diagnosed results on knowledge concepts.

diagnostic results. We then implemented specific models under NeuralCD (i.e. NeuralCDM and three extended models), and with extensive experiments on real-world datasets, we proved the feasibility of the framework. We have preliminarily shown that neural network is competent for cognitive diagnosis and has better potential than traditional models. The NeuralCD framework we proposed is not only accurate and interpretable for cognitive diagnosis, but also has good generality and extensibility. Specifically, we demonstrated that some traditional models (e.g. MIRT) can be regarded as special cases of NeuralCD, while NeuralCD is more flexible to be extended so as to better simulate the cognitive process. For example, more types of response data (e.g. exercise text content) could be aggregated, and the structure of NeuralCD could be adjusted so as to deal with different situations (e.g. knowledge coverage problem).

Serving as the basis of intelligent education, cognitive diagnosis provides supports to numerous adaptive learning applications, such as learning feedback, computerized adaptive testing [8], [57], resource recommendation [58] and learning path planing [59], [60]. In addition to intelligent education, cognitive diagnosis is also widely applicable in areas where users' latent traits such as ability or psychological states require assessments. Clinical assessment is a typical application of cognitive diagnosis [2], including measuring psychological disorder [34], patient-report outcomes [3], etc. In game field, a common demand is to predict players' matchups and preferences [1], which requires assessing players' abilities as well as their cooperation and competition [61]. In career field, An et al. [62] tried to assess the proficiencies of trial lawyers. In summary, cognitive diagnosis is a fundamental task in many areas, and an important basis for extensive applications. Considering the high flexibility and potential of neural network, we hope this work could lead to further studies for cognitive diagnosis in different areas.

Acknowledgments

This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 61922073, U20A20229, 62106244) and the Fundamental Research Funds for the Central Universities (Grants No. WK2150110021), and the Iflytek joint research program.

References

[1] S. Chen and T. Joachims, "Predicting matchups and preferences in context," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 775–784.

[2] M. L. Thomas, "The value of item response theory in clinical assessment: a review," *Assessment*, vol. 18, no. 3, pp. 291–307, 2011.

[3] J. C. Cappelleri, J. J. Lundy, and R. D. Hays, "Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures," *Clinical therapeutics*, vol. 36, no. 5, pp. 648–662, 2014.

[4] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with massive online courses," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 687–698.

[5] H. Burns, C. A. Luckhardt, J. W. Parlett, and C. L. Redfield, *Intelligent tutoring systems: Evolutions in design*. Psychology Press, 2014.

[6] R. Wu, Q. Liu, Y. Liu, E. Chen, Y. Su, Z. Chen, and G. Hu, "Cognitive modelling for predicting examinee performance," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[7] G. D. Kuh, J. Kinzie, J. A. Buckley, B. K. Bridges, and J. C. Hayek, *Piecing together the student success puzzle: research, propositions, and recommendations: ASHE Higher Education Report*. John Wiley & Sons, 2011, vol. 116.

[8] H. Bi, H. Ma, Z. Huang, Y. Yin, Q. Liu, E. Chen, Y. Su, and S. Wang, "Quality meets diversity: A model-agnostic framework for computerized adaptive testing," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 42–51.

[9] J. De La Torre, "Dina model and parameter estimation: A didactic," *Journal of educational and behavioral statistics*, vol. 34, no. 1, pp. 115–130, 2009.

[10] S. E. Embretson and S. P. Reise, *Item response theory*. Psychology Press, 2013.

[11] M. D. Reckase, "Multidimensional item response theory models," in *Multidimensional Item Response Theory*. Springer, 2009, pp. 79–112.

[12] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, no. 8, pp. 30–37, 2009.

[13] Q. Liu, "Towards a new generation of cognitive diagnosis." in *IJCAI*, 2021, pp. 4961–4964.

[14] M. Zhang, W. Wang, X. Liu, J. Gao, and Y. He, "Navigating with graph representations for fast and scalable decoding of neural language models," in *Advances in Neural Information Processing Systems*, 2018, pp. 6308–6319.

[15] K. Song, M. Ji, S. Park, and I.-C. Moon, "Hierarchical context enabled recurrent neural network for recommendation," *arXiv preprint arXiv:1904.12674*, 2019.

[16] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Chen, Y. Yin, Z. Huang, and S. Wang, "Neural cognitive diagnosis for intelligent education systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6153–6161.

[17] M. von Davier, "The dina model as a constrained general diagnostic model: Two variants of a model equivalency," *British Journal of Mathematical and Statistical Psychology*, vol. 67, no. 1, pp. 49–71, 2014.

[18] G. H. Fischer, "Derivations of the rasch model," in *Rasch models*. Springer, 1995, pp. 15–38.

[19] F. M. Lord, *Applications of item response theory to practical testing problems*. Routledge, 2012.

[20] R. J. Adams, M. Wilson, and W.-c. Wang, "The multidimensional random coefficients multinomial logit model," *Applied psychological measurement*, vol. 21, no. 1, pp. 1–23, 1997.

[21] A. Toscher and M. Jahrer, "Collaborative filtering applied to educational data mining," *KDD cup*, 2010.

[22] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Computer Science*, vol. 1, no. 2, pp. 2811–2819, 2010.

[23] N. Thai-Nghe and L. Schmidt-Thieme, "Multi-relational factorization models for student modeling in intelligent tutoring systems," in *Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on*. IEEE, 2015, pp. 61–66.

[24] P.-W. Lei and H. Li, "Performance of fit indices in choosing correct cognitive diagnostic models and q-matrices," *Applied Psychological Measurement*, vol. 40, no. 6, pp. 405–417, 2016.

[25] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[26] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[27] W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8957–8964.

[28] Z. Huang, Q. Liu, E. Chen, H. Zhao, M. Gao, S. Wei, Y. Su, and G. Hu, "Question difficulty prediction for reading problems in standard tests." in *AAAI*, 2017, pp. 1352–1359.

[29] M. Wu, M. Mosse, N. Goodman, and C. Piech, "Zero shot learning for code education: Rubric sampling with deep learning inference," 2019.

[30] Y. Yin, Z. Huang, E. Chen, Q. Liu, F. Zhang, X. Xie, and G. Hu, "Transcribing content from structural images with spotlight mechanism," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2643–2652.

[31] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Advances in Neural Information Processing Systems*, 2015, pp. 505–513.

[32] P. Chen, Y. Lu, V. W. Zheng, and Y. Pian, "Prerequisite-driven deep knowledge tracing," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 39–48.

[33] Q. Liu, Z. Huang, Y. Yin, E. Chen, H. Xiong, Y. Su, and G. Hu, "Ekt: Exercise-aware knowledge tracing for student performance prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 1, pp. 100–115, 2019.

[34] J. L. Templin and R. A. Henson, "Measurement of psychological disorders using cognitive diagnosis models." *Psychological methods*, vol. 11, no. 3, p. 287, 2006.

[35] E. Maris, "Estimating multiple classification latent class models," *Psychometrika*, vol. 64, no. 2, pp. 187–212, 1999.

[36] J. P. Leighton, M. J. Gierl, and S. M. Hunka, "The attribute hierarchy method for cognitive assessment: A variation on tatsuoka's rule-space approach," *Journal of educational measurement*, vol. 41, no. 3, pp. 205–237, 2004.

[37] J. De La Torre and J. A. Douglas, "Higher-order latent trait models for cognitive diagnosis," *Psychometrika*, vol. 69, no. 3, pp. 333–353, 2004.

[38] Q. Liu, R. Wu, E. Chen, G. Xu, Y. Su, Z. Chen, and G. Hu, "Fuzzy cognitive diagnosis for modelling examinee performance," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 4, p. 48, 2018.

[39] L. V. DiBello, L. A. Roussos, and W. Stout, "31a review of cognitively diagnostic assessment and a summary of psychometric models," *Handbook of statistics*, vol. 26, pp. 979–1030, 2006.

[40] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[41] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.

[42] Y. Su, Q. Liu, Q. Liu, Z. Huang, Y. Yin, E. Chen, C. Ding, S. Wei, and G. Hu, "Exercise-enhanced sequential modeling for student performance prediction," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[43] J. Liu, G. Xu, and Z. Ying, "Data-driven learning of q-matrix," *Applied psychological measurement*, vol. 36, no. 7, pp. 548–564, 2012.

[44] W. Huang, E. Chen, Q. Liu, Y. Chen, Z. Huang, Y. Liu, Z. Zhao, D. Zhang, and S. Wang, "Hierarchical multi-label text classification: An attention-based recurrent network approach," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1051–1060.

[45] R. Wu, G. Xu, E. Chen, Q. Liu, and W. Ng, "Knowledge or gaming? cognitive modelling based on multiple-attempt response," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 321–329.

[46] A. Rakhlin, "Convolutional neural networks for sentence classification," *GitHub*, 2016.

[47] J. Yu, G. Luo, T. Xiao, Q. Zhong, Y. Wang, W. Feng, J. Luo, C. Wang, L. Hou, J. Li *et al.*, "Mooccube: a large-scale data repository for nlp applications in moocs," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3135–3142.

[48] X. Huang, Q. Liu, C. Wang, H. Han, J. Ma, E. Chen, Y. Su, and S. Wang, "Constructing educational concept maps with multiple relationships from multi-source data," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 1108–1113.

[49] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," *User Modeling and User-Adapted Interaction*, vol. 19, no. 3, pp. 243–266, 2009.

[50] X. Xiong, S. Zhao, E. G. Van Inwegen, and J. E. Beck, "Going deeper with deep knowledge tracing." *International Educational Data Mining Society*, 2016.

[51] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[52] H. Pei, B. Yang, J. Liu, and L. Dong, "Group sparse bayesian learning for active surveillance on epidemic dynamics," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[53] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[54] A. Pirotte, J.-M. Renders, M. Saerens *et al.*, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Transactions on Knowledge & Data Engineering*, no. 3, pp. 355–369, 2007.

[55] Y. Chen, Q. Liu, Z. Huang, L. Wu, E. Chen, R. Wu, Y. Su, and G. Hu, "Tracking knowledge proficiency of students with educational priors," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 989–998.

[56] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[57] L. M. Rudner, "An examination of decision-theory adaptive testing procedures," in *annual meeting of the American Educational Research Association*, 2002.

[58] Z. Huang, Q. Liu, C. Zhai, Y. Yin, E. Chen, W. Gao, and G. Hu, "Exploring multi-objective exercise recommendations in online education systems," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 1261–1270.

[59] H. Zhu, F. Tian, K. Wu, N. Shah, Y. Chen, Y. Ni, X. Zhang, K.-M. Chao, and Q. Zheng, "A multi-constraint learning path recommendation algorithm based on knowledge map," *Knowledge-Based Systems*, vol. 143, pp. 102–114, 2018.

[60] Q. Liu, S. Tong, C. Liu, H. Zhao, E. Chen, H. Ma, and S. Wang, "Exploiting cognitive structure for adaptive learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 627–635.

[61] Y. Gu, Q. Liu, K. Zhang, Z. Huang, R. Wu, and J. Tao, "Neuralac: Learning cooperation and competition effects for match outcome prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4072–4080.

[62] Y. An, Q. Liu, H. Wu, K. Zhang, L. Yue, M. Cheng, H. Zhao, and E. Chen, "Lawyerpan: a proficiency assessment network for trial lawyers," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 5–13.

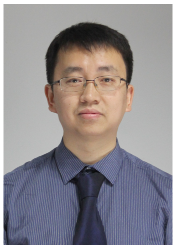
Fei Wang received the BE degree in computer science and technology from the University of Science and Technology of China, Hefei, China. He is currently working toward the Ph.D. degree majoring in Applied Computer Technology with the School of Computer Science and Technology, University of Science and Technology of China. His research interests include data mining and intelligent education systems.

Qi Liu eceived the Ph.D. degree from University of Science and Technology of China (USTC), Hefei, China, in 2013. He is currently a Professor in the School of Computer Science and Technology at USTC. His general area of research is data mining and knowledge discovery. He has published prolifically in refereed journals and conference proceedings (e.g., TKDE, TOIS, KDD). He is an Associate Editor of IEEE TBD and Neurocomputing. He was the recipient of KDD' 18 Best Student Paper Award and ICDM' 11 Best Research Paper Award. He is a member of the Alibaba DAMO Academy Young Fellow. He was also the recipient of China Outstanding Youth Science Foundation in 2019.eceived the Ph.D. degree from University of Science and Technology of China (USTC), Hefei, China, in 2013. He is currently a Professor in the School of Computer Science and Technology at USTC. His general area of research is data mining and knowledge discovery. He has published prolifically in refereed journals and conference proceedings (e.g., TKDE, TOIS, KDD). He is an Associate Editor of IEEE TBD and Neurocomputing. He was the recipient of KDD' 18 Best Student Paper Award and ICDM' 11 Best Research Paper Award. He is a member of the Alibaba DAMO Academy Young Fellow. He was also the recipient of China Outstanding Youth Science Foundation in 2019.r

Enhong Chen (SM'07) is a professor and vice dean of the School of Computer Science at University of Science and Technology of China (USTC). He received the Ph.D. degree from USTC. His general area of research includes data mining and machine learning, social network analysis and recommender systems. He has published more than 200 papers in refereed conferences and journals, including IEEE TKDE, IEEE TMC, KDD, ICDM, NIPS, and CIKM. He was on program committees of numerous conferences including KDD, ICDM, SDM. He received the Best Application Paper Award on KDD-2008, the Best Student Paper Award on KDD-2018 (Research), the Best Research Paper Award on ICDM-2011 and Best of SDM-2015. His research is supported by the National Science Foundation for Distinguished Young Scholars of China. He is a senior member of the IEEE.

Zhenya Huang recieved the BE degree from Shandong University, Ji'nan, China, in 2014 and the Ph.D. degree from University of Science and Technology of China, Hefei, China, in 2020. He is currently working as an associate researcher of the School of Computer Science and Technology, University of Science and Technology of China. His main research interests include data mining, knowledge discovery, recommender systems and intelligent education systems. He has published more than 40 papers in refereed journals and conference proceedings, including TKDE, TOIS, KDD, AAAI, CIKM.

Yu Yin received the BE degree in computer science from University of Science and Technology of China, in 2017. He is currently working toward the Ph.D. degree in the School of Computer Science and Technology at University of Science and Technology of China. His main research interests include data mining, intelligent education systems and image recognition. He won the first prize in the Second Student RDMA Programming Competition, 2014. He has published papers in refereed conference proceedings, such as AAAI and KDD.



Shijin Wang received the BE degree in electronic science and Technology from University of Science and Technology of China, Hefei, China, in 2003 and Ph.D. degree in pattern recognition & intelligent system from Institute of Automation, Chinese Academy of Science, Beijing, China. He is currently the vice president of IFLYTEK Co., Ltd. and the president of IFLYTEK AI Research (Central China). He has published more than 60 papers in refereed conferences such as ICASSP, ACL, KDD, SIGIR and AAAI. His research interests include speech processing, natural language processing and intelligent education. He led the team that won more than ten championships in international technical evaluation such as Blizzard Challenge and CHiME.

Yu Su received the Ph.D. degree from Anhui University. He is currently an associate professor in Hefei Normal University. His main area of research includes data mining, machine learning, recommender systems and intelligent education systems. He has published several papers in referred conference proceedings and journals, such as IJCAI'2015, AAAI'2017, AAAI'2018, KDD'2018, CIKM'2017, DASFAA'2016, ACM TIST, IEEE TKDE, KBS.