

A Comprehensive Analysis of Data Imbalance in Deep Learning-Based Cognitive Diagnosis

, *, ***, ***, ***

Abstract—Cognitive diagnosis aims to quantify students’ learning status and mastery level of related knowledge concepts based on their responses to given exercises. This is a fundamental yet critical research task in the field of intelligent education, which helps to reveal students’ proficiency on multiple knowledge concepts they have learned, thereby providing personalized learning services for each student. In recent years, researchers have succeed in improving model’s diagnosis accuracy by designing diagnostic functions based on deep neural networks or integrating richer contextual features to enhance the representation learning of students and exercises. However, datasets used for deep learning-based cognitive diagnosis model training often present an imbalanced distribution, being a large number of students only answered a few exercises, and a large number of exercises were answered by only a few students, which may have a certain impact on the performance of the model. To verify this problem, we conducted considerable experiments on four well known models and two widely used datasets of deep learning-based cognitive diagnosis in this paper. Firstly, we analyzed the correlation between the model’s predictive accuracy for individual student’s response performance and the number of exercises answered by this student. Secondly, we studied the correlation between the model’s predictive accuracy for each exercise and the number of the exercise being answered by students in the dataset. Finally, we analyzed whether the model’s predictive accuracy for individual student would be over-fitting¹ during multiple epochs and whether the maximum predictive accuracy achieved is affected by the number of exercises answered by this student. The experimental results indicate that there are no evident statistics supporting the strong correlation between the model’s predictive accuracy for individual student and the number of exercises answered by this student. The same case happens with the correlation between the model’s prediction accuracy on each exercise and the number of the exercise being answered by students in the dataset. Notably, we observe that models are more likely to be over-fitting for students who have answered a larger number of exercises.

Index Terms—cognitive diagnosis; intelligent education; imbalanced distribution

I. INTRODUCTION

Cognitive diagnosis is a fundamental yet critical task in the field of intelligent education, which helps to reveal students’ proficiency on different knowledge concepts they have learned and provides personalized learning services for each student.

¹In this work, over-fitting refers to the phenomenon that the model’s predictive accuracy for individual student achieves the maximum accuracy in a certain epoch during training, but then decreases as the model continues to train.

Generally, students first answer a set of prepared exercises and leave a record of their responses. Subsequently, the goal of cognitive diagnosis is to infer students’ mastery level of specific knowledge concepts based on their answering logs and the correlation between exercises and knowledge concepts. This is a very important task in the daily teaching. With the help of diagnostic results, we can guide students in a targeted manner and help them to improve their comprehension of knowledge concepts.

Early cognitive diagnosis models primarily originated from the field of educational psychology, with many of these models relying heavily on statistical methods, such as Item Response Theory (IRT) [1] and the DINA (deterministic inputnoisy ‘and’ gate) model [2]. However, these models are limited in performance and struggle to address the challenges of processing large-scale data due to their reliance on manually designed functions. With the advancement of neural network technology, many fields have achieved state-of-the-art (SOTA) performance by applying it, and the field of cognitive diagnosis is no exception. Deep learning-based cognitive diagnosis models focus on designing diagnostic functions by neural networks to better model the complex interaction between students and exercises, or attempt to integrate richer contextual features and relationships among multiple knowledge concepts to enhance the representation learning of students and exercises.

Although current deep learning-based cognitive diagnosis models have been proven effective, they require large-scale data of student-exercise interactions for training. However, in the real world, datasets often exhibit imbalanced distribution. For most real-world datasets in cognitive diagnosis, many students answered only a few exercises, and there is a significant variation in the frequency of different exercises were answered. Fig. 1 clearly illustrates the distribution of the number of students considering their interaction frequencies, as well as the distribution of the number of exercises considering their frequencies of being answered in ASSIST0910. It is obvious that numerous students have few interactions, while many exercises are answered only a few times. Generally speaking, the model is well trained on the high-resource group with abundant data, but is poorly trained on the low-resource group with insufficient data, resulting in models exhibit commendable performance in high-resource groups but a significant decrease in predictive accuracy within low-resource groups.

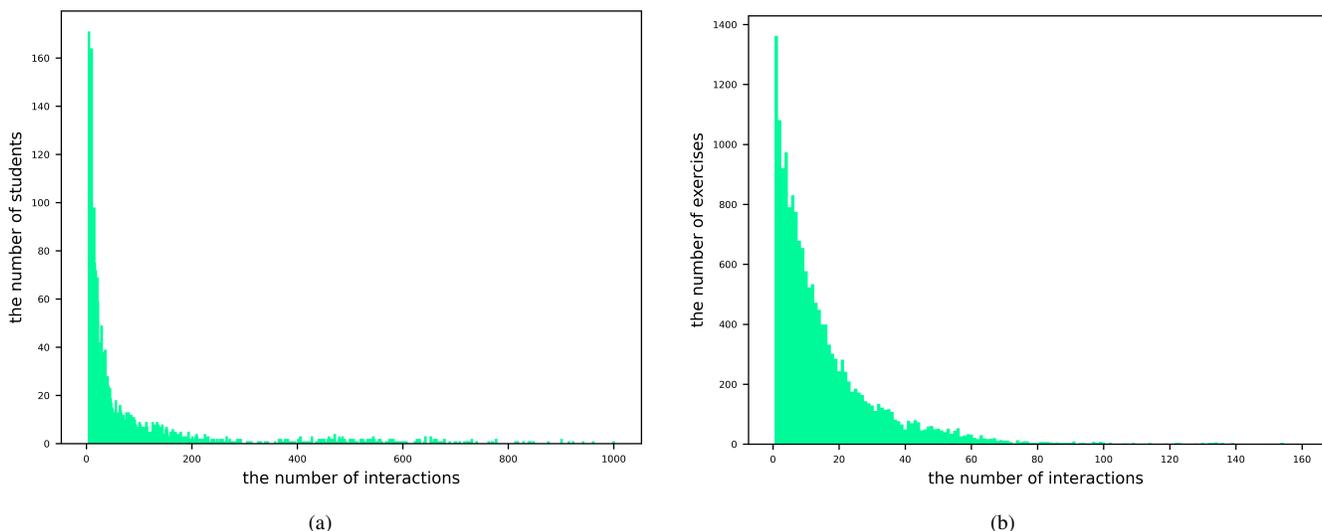


Fig. 1: (a) Statistical distribution of student quantity based on the interaction frequency in ASSIST0910; (b) statistical distribution of exercise quantity based on the number of times they were answered in ASSIST0910.

Thus, in the realm of deep learning-based cognitive diagnosis, might we encounter a similar phenomenon?

This work aims to deeply analyze the impact of data imbalance on model performance in the field of deep learning-based cognitive diagnosis. To the best of our knowledge, it is the first work studying this issue. We conducted considerable experiments with four models and two datasets that are famous and widely used in this field. Main contributions of this paper can be summarized as follows:

- Firstly, we explored the problem whether the model’s predictive accuracy for individual student correlates to the number of exercises answered by the student.
- Secondly, we studied the problem whether the model’s predictive accuracy for each exercise correlates to the number of the exercise being answered by students.
- Finally, we analyzed whether the model’s predictive accuracy for individual student would be over-fitting during multiple epochs and whether the maximum predictive accuracy achieved is affected by the number of exercises answered by this student.

Section II reviews the related works. Section III introduces the research questions and the corresponding experiments. Conclusions is in Section IV.

II. RELATED WORKS

Early cognitive diagnosis models (CDMs) mainly come from the field of educational psychology, and they are mainly based on statistical methods. Among them, IRT [1] and DINA [2] are the two most classical ones. Many subsequent CDMs are improved from these two models. IRT used one-dimension and continuous latent traits (i.e., students’ proficiency on knowledge concepts, the difficulty of exercises, and the discrimination of exercises) to represent students and exercises. It predicted the probability of students can answer exercise

correctly through a logistic function. Nevertheless, this model can only represent students’ abilities on a single dimension. In contrast, Multidimensional Item Response Theory (MIRT) model [3] extended students’ latent traits and exercise parameters into a multidimensional space to comprehensively represent students’ abilities across various dimensions. While the aforementioned two models can only provide an assessment value of students’ ability, DINA model aligned exercises with specific knowledge concepts by introducing the Q-matrix so as to accurately express students’ specific knowledge mastery states. It used a binary vector to represent students’ mastery of each knowledge concept, operating on a non-compensatory assumption that students can only answer correctly when they have mastered all knowledge concepts related to the exercise. Additionally, the model incorporated the guessing and slipping parameters to accommodate noisy data.

CDMs based on statistical methods rely on manually designed functions with limited performance. It is hard to handle large-scale data using these methods. With the advancement of neural network technology, many fields have achieved SOTA performance by applying it, and the field of cognitive diagnosis is no exception. Deep Item Response Theory (DIRT) model [4] was based on IRT and used deep learning technology to enhance the representation of students and exercises by combining exercise-texts and the relationship between exercises and knowledge concepts. To address the problem that artificially designed functions cannot fit the complex relationship between students and exercises well, Neural Cognitive Diagnosis (NeuralCD) framework [5] utilized multiple neural network layers to model the interaction between students and exercises and employed the monotonicity assumption to ensure the interpretability of the model. NeuralCD primarily considered students’ mastery of knowledge concepts, the difficulty of knowledge concepts and the discrimination of exercises

during the diagnostic process, but neglected the importance of knowledge concepts. Thus it was insufficient to capture the complex interactions between students and exercises. Importance of Knowledge Point-Based Neural Cognitive Diagnosis (IK-NeuralCD) model [6] introduced the importance factor of knowledge concept and utilized the frequency of occurrence of knowledge concepts to express their significance, thereby improving the modeling of complex relationships between students and exercises. The two aforementioned models only constrained the parameters of the fully connected layer in the interaction stage to be positive to satisfy the monotonicity assumption, but did not consider this constraint during the model optimization process. Item Response Ranking (IRR) framework [7] introduced pairwise learning into cognitive diagnosis in order to satisfy the monotonicity assumption during the process of model optimization. Multitask Based Group-Level Cognitive Diagnosis (MGCD) framework [8] simultaneously modeled individual student performance and group performance. It transformed the information of students' response records into group representations through shared student representations, and used the attention mechanism to model the relationship between them. Previous models did not take the aggregation of knowledge concepts into account during the diagnostic process. CDGK [9] model applied neural networks to capture the nonlinear interaction among exercise features, students' scores and students' proficiency of knowledge concepts. Furthermore, it performed the aggregation of knowledge concepts by transforming them into a graph structure and focusing only on the leaf nodes of the knowledge concept hierarchy. However, it only aggregated knowledge concepts and did not consider the impact of different knowledge concepts on students' scores in answering exercises. Cognitive Diagnostic Model Focusing on Knowledge Concept (CDMFKC) model [10] focused on designing the difficulty and discrimination of each knowledge concept. It utilized multiple neural network layers to model the complex interactions between students and exercise attributes to obtain accurate and interpretable diagnostic results. Similar to previous models, it also overlooked the dependencies among knowledge concepts. In fact, students master sub-level knowledge concepts first, then parent-level knowledge concepts. Bayesian Network-Based Hierarchical Cognitive Diagnosis Framework (HierCDF) [11] utilized a Bayesian network to reasonably and efficiently model students' cognitive states in the attribute hierarchy. Subsequently, a CDM adapter was designed to bridge the gap between students' cognitive states and input features of the diagnosis model. While HierCDF has certain advantages in the attribute-hierarchy modeling and the interpretability of diagnosis results, the model faces challenges due to its high complexity, low computational efficiency and difficulties in parameter learning. Knowledge-Sensed Cognitive Diagnosis (KSCD) framework [12] first mapped students, exercises, and knowledge concepts into embedding representation matrices where intrinsic relationships among knowledge concepts are reflected. Then, the student knowledge mastery level was obtained by taking the product

of the knowledge-sensed student knowledge mastery vector, exercise vector and knowledge vector. This design made the students' mastery of non-interactive knowledge concepts be interpretably inferred. Quantitative Relationship Cognitive Diagnosis (QRCDM) model [13] assumed that exercises not only relate to knowledge concepts marked by experts, but also implicitly relate to some unmarked knowledge concepts. It calculated the dual contribution matrix of exercises and knowledge concepts via neural networks. Following this, the students' mastery of knowledge concepts and their scores were predicted based on the contribution matrix and the guessing and slipping parameters. It assumed that knowledge concepts are independent with each other and did not consider the dependencies among them. However, this assumption is not consistent with the actual situation. Interpretable Cognitive Diagnosis (ICD) model [14] added a neural network layer to fit the mutual impact among knowledge concepts.

The interaction relationship among students, exercises and knowledge concepts naturally presents as a graph structure. Graph neural network-based CDMs can effectively improve the representation quality of these three items by aggregating neighbor features, therefore predicting students' performance more accurately through interaction functions. Relation Map Driven Cognitive Diagnosis (RCD) model [15] constructed the interactions among the three in a hierarchical graph. The graph contained a student-exercise interaction graph, a concept-exercise association graph and a concept dependency graph extracted from the priori relationships among knowledge concepts. It used a multi-level attention neural network to achieve the node aggregation of hierarchical graphs. However, RCD model only simplified the student-exercise interaction into a binary interaction (i.e., interaction and non-interaction) and ignored the rich information contained in the different behavior patterns (correct and incorrect interactions) of students in answering exercises. Graph-based Cognitive Diagnosis (GCDM) model [16] divided interactions between students and exercises into two categories (correct answers and incorrect answers), and designed two graph-based layers (the information transfer layer and the knowledge aggregation layer). The former was used to propagate students' cognitive states through different types of graph edges, while the latter selectively collected information from adjacent graph nodes. Multi-Relational Cognitive Diagnosis (MRCD) framework [17] constructed two student-exercise interaction graphs based on the correctness of students' answers. Then, graph convolutional network was utilized to learn exercise-level representations of students and exercises based on different interaction graphs with graph contrastive learning employed to enhance the learning process. Considering that the number of knowledge concepts is relatively small, MRCD directly adopted the attention mechanism to generate concept-level representations of students and exercises. Afterwards, exercise-level and concept-level representations were fused and conveyed to a diagnostic function to predict student performance. Previous models ignored the impact of data imbalance in this field on model training, and discarded student samples with fewer interactions during the

model training process. Self-Supervised Cognitive Diagnosis (SCD) framework [18] leveraged self-supervised learning to assist graph-based cognitive diagnosis, removing edges based on specific rules to generate diverse sparse views. The model paid more attention to those long-tail students by maximizing cross-view consistency of node representations. Aforementioned research on cognitive diagnosis mainly focused on improving the accuracy of diagnostic results, often ignoring the important and practical task, being domain-level zero-shot cognitive diagnosis (DZCD). Transferable Knowledge Concept Graph Embedding Framework for Cognitive Diagnosis (TechCD) [19] constructed the relationship among students, exercises and knowledge concepts in a Knowledge Concept Graph (KCG), and used a graph convolutional network (GCN) to perform representation learning on the KCG. In order to capture propagation properties of embeddings, transferable student cognitive states and exercise features were built by discarding low-level embeddings of GCN and only aggregating high-level embeddings.

Although existing researches have achieved great successes, few works have paid attention to the issue of imbalanced datasets in this field. For many deep learning-based tasks, such as image classification or character recognition, usually the model is well trained on categories with abundant samples, but is poorly trained on few shot categories. Is this also the case for deep learning-based cognitive diagnosis? A deep exploration and analysis is required.

III. EXPERIMENTS

In this section, we will describe the benchmark datasets and experimental setups in detail. Considerable experiments are conducted to answer the following research questions:

- RQ 1: Does the model’s prediction accuracy for individual student correlate to the number of his/her interactions?
- RQ 2: Does the model’s prediction accuracy on a certain exercise correlate to the frequency of the exercise being answered in the dataset?
- RQ 3: Whether the model’s predictive accuracy for individual student’s response performance would be over-fitting during training epochs? And does the over-fitting correlate to the number of his/her interactions?
- RQ 4: Whether the maximum prediction accuracy for individual student achieved during different training epochs is correlated to the number of his/her interactions?

A. Datasets

We conducted experiments on two real-world datasets: ASSIST09010² and JunYi³. ASSIST09010 is a public dataset collected by the online tutoring system ASSISTments, containing students’ response records during the academic year from 2009 to 2010. Further, the relationship between exercises and knowledge concepts is available in this dataset. JunYi originates from the online learning platform Junyi Academy,

comprising answering records from October 2012 to January 2015. Each exercise contains only one knowledge concept, and each knowledge concept is covered by only one exercise. To ensure adequate training samples, most studies filtered out student samples with a small number of interactions. In this work, as we study the problem of imbalanced distribution, only student samples with less than 5 interactions were filtered. This criteria was set because we split the dataset into training and test with an 8:2 ratio. Table I presents the statistics of two selected public datasets, including the number of students, the number of exercises, the number of knowledge concepts, the total number of interactions, the average number of interactions per student has and the average number of knowledge concepts per exercise contains.

TABLE I: The statistics of ASSIST0910 and Junyi.

Dataset	ASSIST0910	Junyi
#Students	3628	10000
#Exercises	16866	706
#Knowledge concepts	110	706
#Response logs	269269	224380
#Avg logs per student	74.22	22.438
#Avg concepts per exercise	1.18	1.0

B. Experimental Setups

- Experimental settings: We selected four prominent cognitive diagnosis models for experimentation: NCD [5], RCD [15], SCD⁴ [18], and ICD [14]. Training epochs of all models were set to 10 and other hyper-parameters were set according to their original papers. We implemented all the models using PyTorch and conducted all experiments on a Windows 10 server equipped with a 3.00 GHz Intel(R) Core (TM) i9-13900K CPU and a RTX 4090 GPU.
- The Pearson correlation coefficient: The Pearson correlation coefficient, also known as the Pearson product-moment correlation coefficient, is a statistical metric that measures the linear correlation between two variables. The closer the absolute value of the coefficient is to 1, the stronger the linear correlation between two variables, and vice versa. The Pearson correlation coefficient is calculated based on the covariance and standard deviation, and can be expressed using the formula:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (1)$$

where X_i and Y_i are the individual data points, \bar{X} and \bar{Y} are the respective means. Through this formula, we can quantify and comprehend the linear relationship between variables which is crucial for data analysis and scientific research. The Pearson correlation coefficient is widely utilized in various fields such as psychology, social sciences, bio-statistics and economics, aiding researchers in

²<https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data/skill-builder-data-2009-2010>

³<https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198>

⁴Here, we removed the data augmentation module from the original SCD model to avoid its possible effect on the prediction results.

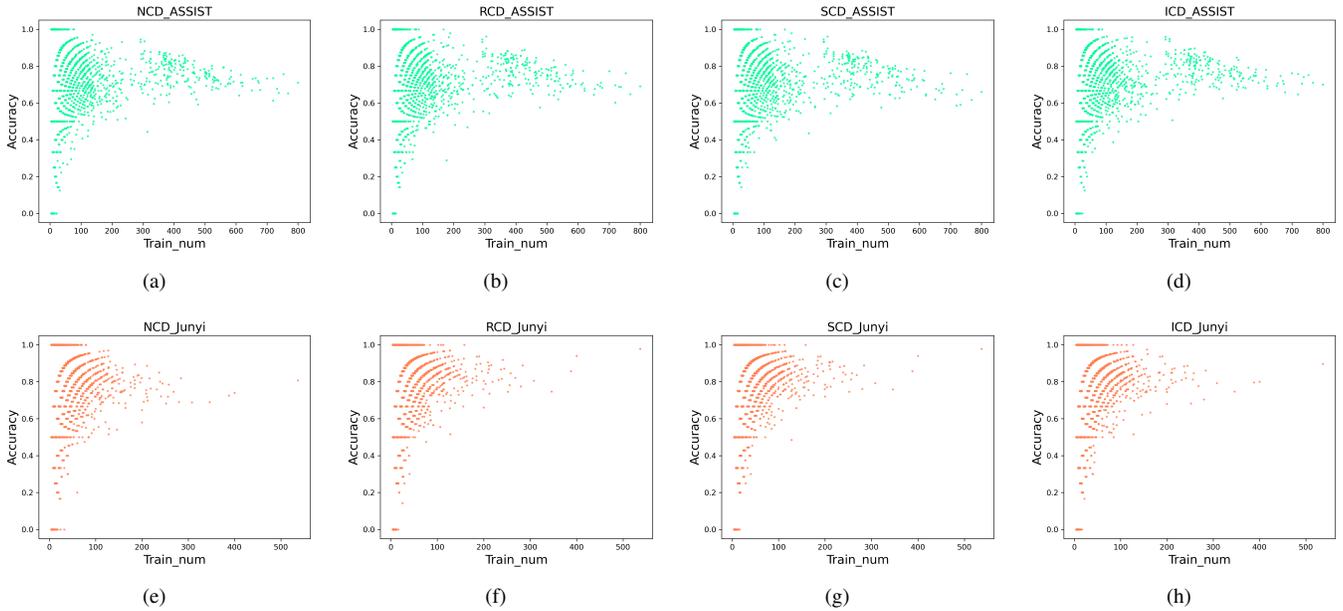


Fig. 2: The scatter plot depicting the distribution of the model’s prediction accuracy for individual student and the number of exercises answered by the student. (a-d) Statistical results of four models on ASSIST and (e-h) statistical results of four models on Junyi dataset.

uncovering potential correlations between variables and laying a foundation for further causal relationship investigations. Therefore, we chose the pearson correlation coefficient as the metric in this work.

TABLE II: The Pearson correlation coefficient between the model’s prediction accuracy for individual student and the number of exercises answered by the student.

Dataset \ Model	NCD	RCD	SCD	ICD
ASSIST0910	0.018793	0.021944	0.012207	0.029801
Junyi	0.030464	0.049657	0.047572	0.046319

TABLE III: The Pearson correlation coefficient between the model’s prediction accuracy for individual student with over 300 training records and the number of exercises answered by the student in ASSIST.

Dataset \ Model	NCD	RCD	SCD	ICD
ASSIST0910	-0.499492	-0.509427	-0.429614	-0.488778

C. RQ 1

To answer research question 1, we selected the prediction results from the epoch with the highest total prediction accuracy (#correct predictions/#all predictions) during the training process for each of four models. We counted the model’s prediction accuracy for individual student and the number of exercises answered by each student. Fig. 2 illustrates the scatter plot depicting their relationships, while the Pearson correlation coefficient between them is computed

and presented in Table II. The results indicate that there is no significant correlation between the model’s prediction accuracy for individual student and the number of exercises answered by the student. The scatter plot generated based on models’ performance on ASSIST0910 conspicuously illustrates that there exists a certain degree of negative correlation between two variables as the number of interactions reaches a certain threshold. To further validate this observation, we conducted a statistical analysis of students with over 300 records in the training set and computed the correlation coefficient between two variables as shown in Table III.

TABLE IV: The Pearson correlation coefficient between the model’s accuracy on a certain exercise and the frequency of the exercise being answered in the training set.

Dataset \ Model	NCD	RCD	SCD	ICD
ASSIST0910	-0.0093	-0.007092	-0.016533	-0.016679
Junyi	0.137817	0.082181	0.081158	0.126585

D. RQ 2

As same as RQ 1, we use the prediction results of the epoch with highest total prediction accuracy (#correct predictions/#all predictions) during the training process. Fig. 3 illustrates the scatter plot depicting the distribution of the model’s prediction accuracy on a certain exercise and the number of the exercise being answered in the training set. We do not observe any evident linear relationship between these two variables. Further computation of their Pearson correlation coefficient is presented in Table IV. The results indicate that there is no significant correlation between the

TABLE V: The proportion of students in each group whose predictive accuracy has achieved the maximum accuracy for each epoch, (a-d) the statistical results of the predictive performance of the four models on the ASSIST dataset and (e-h) the statistical results of the predictive performance of the four models on the Junyi dataset

(a) NCD_ASSIST			(b) RCD_ASSIST			(c) SCD_ASSIST			(d) ICD_ASSIST		
Epoch	≤ 100	>100									
1	0.0826	0.2167	1	0.1056	0.1685	1	0.0512	0.0537	1	0.0372	0.0574
2	0.1143	0.3259	2	0.1635	0.3389	2	0.1496	0.2315	2	0.0725	0.1315
3	0.1493	0.4000	3	0.2011	0.4556	3	0.2021	0.3463	3	0.1114	0.2037
4	0.1661	0.4611	4	0.2218	0.5074	4	0.2222	0.4222	4	0.1460	0.2685
5	0.1846	0.4963	5	0.2345	0.5352	5	0.2438	0.4944	5	0.1881	0.3370
6	0.2069	0.5500	6	0.2481	0.5630	6	0.2594	0.5426	6	0.2341	0.4444
7	0.2312	0.6130	7	0.2556	0.5852	7	0.2801	0.6037	7	0.2707	0.5259
8	0.2536	0.6667	8	0.2610	0.5926	8	0.3025	0.6518	8	0.3131	0.6167
9	0.2843	0.7370	9	0.2636	0.6037	9	0.3332	0.7296	9	0.3582	0.6981
10	1.0000	1.0000	10	1.0000	1.0000	10	1.0000	1.0000	10	1.0000	1.0000

(e) NCD_Junyi			(f) RCD_Junyi			(g) SCD_Junyi			(h) ICD_Junyi		
Epoch	≤ 50	>50									
1	0	0	1	0.0520	0.0625	1	0.0393	0.0609	1	0.0337	0.0453
2	0.0013	0.0109	2	0.0755	0.0797	2	0.0611	0.0844	2	0.0583	0.0703
3	0.0437	0.1484	3	0.0888	0.0984	3	0.0765	0.1031	3	0.0721	0.0891
4	0.0556	0.2125	4	0.0994	0.1156	4	0.0867	0.1344	4	0.0872	0.1516
5	0.0735	0.2578	5	0.1080	0.1328	5	0.0972	0.1594	5	0.0964	0.1766
6	0.0999	0.3109	6	0.1140	0.1563	6	0.1103	0.2031	6	0.1099	0.2188
7	0.1138	0.3734	7	0.1192	0.1766	7	0.1241	0.2359	7	0.1262	0.2656
8	0.1251	0.4406	8	0.1244	0.1906	8	0.1501	0.3328	8	0.1472	0.3188
9	0.1487	0.5281	9	0.1295	0.2109	9	0.1851	0.4344	9	0.1885	0.4484
10	1.0000	1.0000	10	1.0000	1.0000	10	1.0000	1.0000	10	1.0000	1.0000

model’s accuracy on a certain exercise and the frequency of the exercise being answered.

E. RQ 3

In the previous two sections, it was observed that models within the realm of cognitive diagnosis do not exhibit the similar behavior as typical deep learning models, namely predictive capabilities are weaker for categories with fewer samples and stronger for those with more samples. Surprisingly, there was no evident correlation observed between the model’s predictive ability and the quantity of training samples in the realm of cognitive diagnosis. Opting for the prediction results of the epoch with the highest overall prediction accuracy during the training process may also potentially influence the statistical outcomes. We compared the model’s prediction accuracy for each student’s response performance across ten epochs, and discovered that there is an over-fitting phenomenon in the model’s prediction accuracy for partial students. We divided students into two groups based on the number of exercises they answered (split point is 100 for ASSIST0910 and 50 for Junyi). We calculated the proportion of students in each group whose prediction accuracy has achieved the maximum accuracy (including those have achieved the maximum accuracy before the current epoch) for each epoch, as shown in Table V. Since the model was trained for only 10 epochs, the prediction accuracy for some students reached the maximum value in the last epoch. It remains uncertain whether there will be an over-fitting phenomenon on these students as the model continues to train. From the table, we can observe that

the model is more likely to be over-fitting for students with a larger number of interactions.

TABLE VI: The Pearson correlation coefficient between the maximum prediction accuracy for individual student achieved during the training and the number of exercises answered by the student, only considering students exhibiting over-fitting phenomenon.

Dataset \ Model	NCD	RCD	SCD	ICD
ASSIST0910	-0.098165	-0.149055	-0.144242	-0.086246
Junyi	-0.074608	-0.152365	-0.127435	-0.071997

F. RQ 4

In Section III-E, we observed an over-fitting phenomenon in the model’s prediction accuracy for individual student across multiple epochs. We hypothesized that there might be a correlation between the maximum prediction accuracy for individual student achieved during the training and the number of exercises answered by the student. However, the scatter plot in Fig. 4 reveals no obvious correlation between these two variables either. Further calculation of the Pearson correlation coefficient between them is shown in Table VI. The results indicate that there is no significant correlation between the maximum prediction accuracy for individual student achieved during the training and the number of exercises answered by the student.

IV. CONCLUSIONS

In this work, we analyzed the impact of data imbalance on model training in the field of deep learning-based cog-

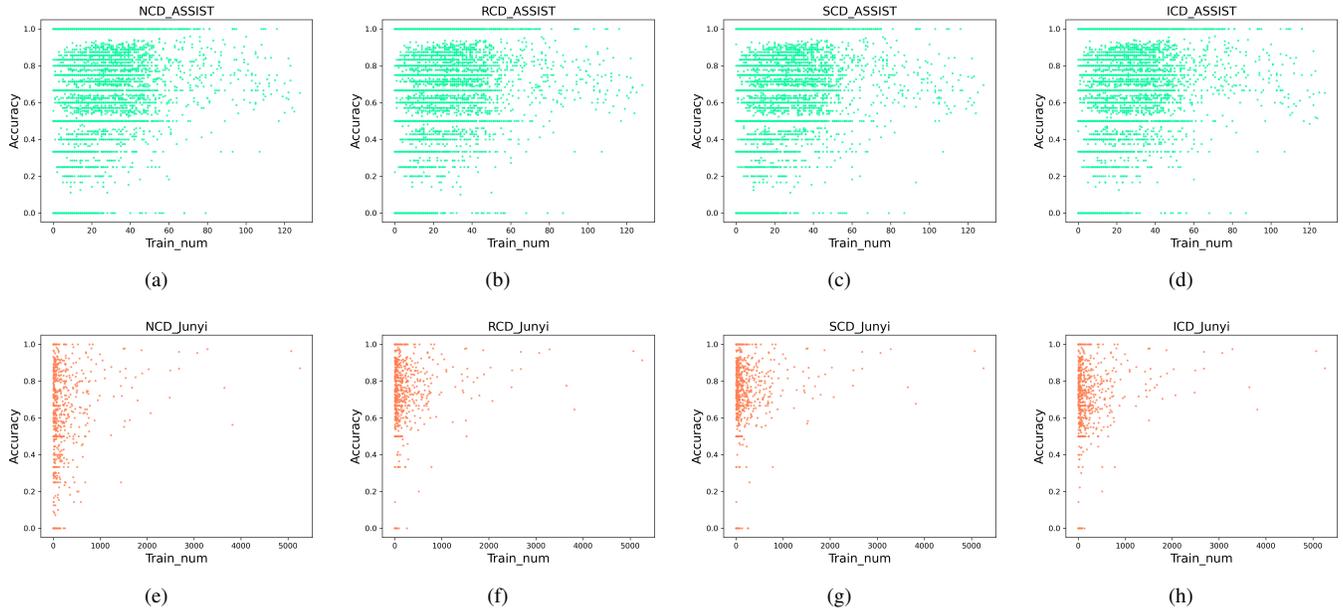


Fig. 3: The scatter plot depicting the distribution of the model’s prediction accuracy on a certain exercise and the number of the exercise being answered in the training set. (a-d) Statistical results of four models on ASSIST and (e-h) statistical results of four models on Junyi.

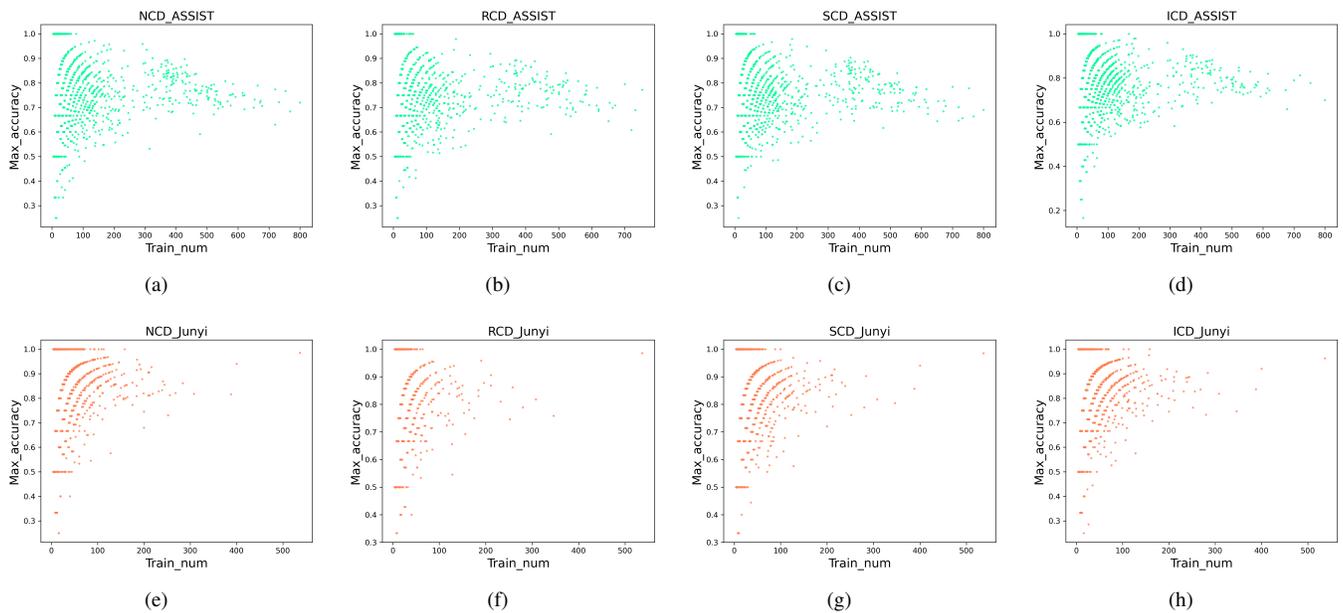


Fig. 4: The scatter plot depicting the distribution of the maximum prediction accuracy for individual student achieved during the training and the number of exercises answered by the student, only considering students exhibiting over-fitting phenomenon. (a-d) Statistical results of four models on ASSIST and (e-h) statistical results of four models on Junyi.

nitive diagnosis. The experimental results indicate that the model's prediction accuracy for individual student's response performance is not significantly correlated to the number of questions answered by the student. Similarly, the model's prediction accuracy on a certain exercise is not clearly correlated with the frequency of the exercise being answered. Notably, we observed that models are more likely to be over-fitting for students who have answered a larger number of exercises. Cognitive diagnosis attempts to model human beings' cognitive status which is very complicated in fact. We think the data imbalance problem in this field can not be defined only from the perspective of the quantity of training samples.

REFERENCES

- [1] Wendy M Yen and Anne R Fitzpatrick. Item response theory. *Educational measurement*, 4:111–153, 2006.
- [2] Jimmy De La Torre. Dina model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, 34(1):115–130, 2009.
- [3] Terry A Ackerman. Multidimensional item response theory models. *Wiley StatsRef: Statistics Reference Online*, 2014.
- [4] Song Cheng, Qi Liu, Enhong Chen, Zai Huang, Zhenya Huang, Yiyang Chen, Haiping Ma, and Guoping Hu. Dirt: Deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2397–2400, 2019.
- [5] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6153–6161, 2020.
- [6] Yan Cheng, Meng Li, Haomai Chen, Yingying Cai, Huan Sun, Gang Wu, Zhuang Cai, and Guanghe Zhang. Neural cognitive modeling based on the importance of knowledge point for student performance prediction. In *2021 16th International Conference on Computer Science & Education (ICCSE)*, pages 495–499. IEEE, 2021.
- [7] Shiwei Tong, Qi Liu, Runlong Yu, Wei Huang, Zhenya Huang, Zachary A Pardos, and Weijie Jiang. Item response ranking for cognitive diagnosis. In *IJCAI*, pages 1750–1756, 2021.
- [8] Jie Huang, Qi Liu, Fei Wang, Zhenya Huang, Songtao Fang, Runze Wu, Enhong Chen, Yu Su, and Shijin Wang. Group-level cognitive diagnosis: A multi-task learning perspective. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE, 2021.
- [9] Xiping Wang, Caidie Huang, Jinfang Cai, and Liangyu Chen. Using knowledge concept aggregation towards accurate cognitive diagnosis. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2010–2019, 2021.
- [10] Sheng Li, Quanlong Guan, Liangda Fang, Fang Xiao, Zhenyu He, Yizhou He, and Weiqi Luo. Cognitive diagnosis focusing on knowledge concepts. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3272–3281, 2022.
- [11] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. Hiercdf: A bayesian network-based hierarchical cognitive diagnosis framework. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 904–913, 2022.
- [12] Haiping Ma, Manwei Li, Le Wu, Haifeng Zhang, Yunbo Cao, Xingyi Zhang, and Xuemin Zhao. Knowledge-sensed cognitive diagnosis for intelligent education platforms. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 1451–1460, 2022.
- [13] Haowen Yang, Tianlong Qi, Jin Li, Longjiang Guo, Meirui Ren, Lichen Zhang, and Xiaoming Wang. A novel quantitative relationship neural network for explainable cognitive diagnosis model. *Knowledge-Based Systems*, 250:109156, 2022.
- [14] Tianlong Qi, Meirui Ren, Longjiang Guo, Xiaokun Li, Jin Li, and Lichen Zhang. Icd: A new interpretable cognitive diagnosis model for intelligent tutor systems. *Expert Systems with Applications*, 215:119309, 2023.
- [15] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 501–510, 2021.
- [16] Yu Su, Zeyu Cheng, Jinze Wu, Yanmin Dong, Zhenya Huang, Le Wu, Enhong Chen, Shijin Wang, and Fei Xie. Graph-based cognitive diagnosis for intelligent tutoring systems. *Knowledge-Based Systems*, 253:109547, 2022.
- [17] Kaifang Wu, Yonghui Yang, Kun Zhang, Le Wu, Jing Liu, and Xin Li. Multi-relational cognitive diagnosis for intelligent education. In *CAAI International Conference on Artificial Intelligence*, pages 425–437. Springer, 2022.
- [18] Shanshan Wang, Zhen Zeng, Xun Yang, and Xingyi Zhang. Self-supervised graph learning for long-tailed cognitive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 110–118, 2023.
- [19] Weibo Gao, Hao Wang, Qi Liu, Fei Wang, Xin Lin, Linan Yue, Zheng Zhang, Rui Lv, and Shijin Wang. Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 983–992, 2023.