applied sciences

MDPI

Article

Deep Knowledge Tracing Based on Spatial and Temporal Representation Learning for Learning Performance Prediction

Liting Lyu [1], Zhifeng Wang [2,*], Haihong Yun [1], Zexue Yang [1] and Ya Li [1]

[1] School of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin 150050, China; lyuliting2022@163.com (L.L.); yunhh126@126.com (H.Y.); yzx1978@126.com (Z.Y.); ly.218@163.com (Y.L.)
[2] School of Educational Information Technology, Central China Normal University, Wuhan 430079, China
* Correspondence: zfwang@ccnu.edu.cn

Abstract: Knowledge tracing (KT) serves as a primary part of intelligent education systems. Most current KTs either rely on expert judgments or only exploit a single network structure, which affects the full expression of learning features. To adequately mine features of students' learning process, Deep Knowledge Tracing Based on Spatial and Temporal Deep Representation Learning for Learning Performance Prediction (DKT-STDRL) is proposed in this paper. DKT-STDRL extracts spatial features from students' learning history sequence, and then further extracts temporal features to extract deeper hidden information. Specifically, firstly, the DKT-STDRL model uses CNN to extract the spatial feature information of students' exercise sequences. Then, the spatial features are connected with the original students' exercise features as joint learning features. Then, the joint features are input into the BiLSTM part. Finally, the BiLSTM part extracts the temporal features from the joint learning features to obtain the prediction information of whether the students answer correctly at the next time step. Experiments on the public education datasets ASSISTment2009, ASSISTment2015, Synthetic-5, ASSISTchall, and Statics2011 prove that DKT-STDRL can achieve better prediction effects than DKT and CKT.

Keywords: prediction; learning performance; e-learning; deep learning; knowledge tracing; knowledge representation; spatial feature; temporal feature; convolutional neural network; bidirectional long short-term memory

Citation: Lyu, L.; Wang, Z.; Yun, H.; Yang, Z.; Li, Y. Deep Knowledge Tracing Based on Spatial and Temporal Representation Learning for Learning Performance Prediction. *Appl. Sci.* **2022**, *12*, 7188. https://doi.org/10.3390/app12147188

Academic Editors: William Yu Chung Wang, Yung-Chun Chang, Jheng-Long Wu and Hong-Jie Dai

Received: 1 June 2022
Accepted: 14 July 2022
Published: 17 July 2022

1. Introduction

At present, learning management systems (LMSs), are widely welcomed [1]. LMSs are software systems for distance education based on the internet. LMSs have the functions of managing students and learning resources, providing students with the services of online learning, exercises, tests, communication, registration, scheduling, logging, and so on [2]. The typical LMSs can be divided into categories of open LMSs (e.g., EdX, Moodle [3]), customized LMSs (e.g., ASSISTments [4]), and learning management ecosystems [2,5]. LMSs have rich resources, flexibility, and convenience, which brings new development opportunities for intelligent education [6]. As a crucial research branch of intelligent education, knowledge tracing (KT), which promotes solving the problem of personalized tutoring for learners, has attracted more and more attention [7]. KT can model and analyze the data of interactions from learners practicing online to obtain the potential law of the change of different students' knowledge status, and then predict students' future learning performance [8]. Specifically, knowledge tracing can model the students' practice process by logistic function, machine learning (such as hidden Markov models) or deep learning (such as recurrent neural networks, graph neural networks) algorithm models based on the students' practice records collected by LMSs such as ASSISTments and Coursera. Through the modeling and analysis of the interaction process between students and exercises, KT models can learn parameters about the hidden state of students' mastery of knowledge

points (the probability of successfully mastering the corresponding knowledge points of exercises) from the explicit answer history sequence and then can output the prediction of whether the student can answer the next question correctly or not. This is because it is generally believed in the field of education that exercises can test students' grasp of the knowledge points contained in the exercises, and students' grasp of the knowledge points can also be reflected through exercises. KT can analyze and predict the learning status of different students, which is helpful for LMSs to provide personalized learning programs and help students improve learning efficiency [9]. When most students can achieve better learning results, the educational level and social economy are expected to usher in further prosperity and development. As a result, more research has been conducted on KT.

Today, there are many KT models. First of all, Bayesian knowledge tracing (BKT) was proposed in 1994 [8], which took whether students have mastered a certain knowledge point or not as a hidden variable and used a hidden Markov model (HMM) to model its changes in the learning process. BKT has laid an important foundation for the development of KT. For a long period of time, the research on KT has mainly focused on the improvement and application of BKT, such as Knowledge Tracing: Item Difficulty Effect Model (KT-IDEM) [10] and Personalized Clustered BKT (PC-BKT) [11]. However, the traditional KT models represented by BKT were limited by manual tags and were limited by modeling and analyzing single concepts. Then, deep knowledge tracing (DKT) [12] was put forward in 2015, which applied recurrent neural networks (RNNs) to modeling the learning process of students and predicting their future performance. DKT solved the problem of the discrete modeling of knowledge points without manually annotating the relationship between exercises and knowledge points. Since then, more and more deep learning (DL) techniques have been introduced into KT modeling tasks. For example, dynamic key-value memory networks (DKVMNs) [13] brought in and improved memory-augmented neural networks (MANNs) to automatically discover knowledge points in exercises and predict students' knowledge state; convolutional neural networks (CNNs) were introduced into CKT [14] to analyze students' personalized phased learning characteristics and predict learning performance. GKT [15] built a graph neural network (GNN) by using the correlation between knowledge points to predict learning performance. The KT models based on deep learning had significant advantages in automatically extracting features, which was suitable for mining potential rules from massive data without manual operation. Compared with traditional KTs, KTs on the basis of DL catered to the development of big data in education. However, the existing KT models had the problem that they expressed features inadequately. It can be found that most of the existing KTs on the basis of DL used relatively single network structures. For example, DKT just used RNN or LSTM to extract the temporal features of students' learning sequences, CKT just used CNN to extract the spatial features of students' learning sequences, and GKT just used GNN to learn the correlations between knowledge points. The feature extraction abilities of different DL network structures have different advantages and disadvantages. So, KTs based on single networks had limitations in sufficiently extracting features, resulting in insufficiently mining and utilizing information based on original data and a difficulty in obtaining a more accurate prediction.

To solve the problem that the KT model does not adequately express deep features, we propose Deep Knowledge Tracing Based on Spatial and Temporal Deep Representation Learning for Learning Performance Prediction (DKT-STDRL) , which can fully explore and comprehensively utilize the spatial and temporal characteristics of students' exercise sequences, so as to obtain more accurate prediction results.. Inspired by DKT [12] and CKT [14], this model combines the students' exercise history and the spatial features as the joint learning features to further extract the temporal features of students' learning process. Specifically, for a given sequence of students' answer performance, we first extract the spatial features of students' exercise sequences by using multilayer convolution neural networks. The spatial features can represent the learners' personalized learning efficiency, which can be reflected from the performance of continuous problem solving [14].

Then, we extract the temporal characteristics of learners' learning performance sequences through bidirectional long short-term memory (BiLSTM) [16] based on the combination of spatial learning features and the original response performance. The temporal features can represent the changing process of students' knowledge state. Because of the sufficient utilization of bidirectional sequence signals by BiLSTM [17], the DKT-STDRL considers both the past and future learning performance of students so that a better judgment of the current time can be obtained.

We sum up our main contributions as follows:

1. A new KT model is proposed by us, which is called Deep Knowledge Tracing Based on Spatial and Temporal Deep Representation Learning for Learning Performance Prediction (DKT-STDRL). This model tries to combine the advantages of CKT [14] and DKT [12] in extracting the spatial features and temporal features of students' learning process, so as to mine students' learning characteristics in two ways;
2. The DKT-STDRL model combines the spatial features of students learning sequence over the past period of time extracted by the multilayer convolutional neural network with the data of students' historical answers to obtain the joint learning features. Then, we employ BiLSTM to further learn the time-series information of the joint learning features. The overlapped neural networks make the model take into account the spatial learning features when extracting the temporal features of students learning, so it is conducive to mining and obtaining deeper learning information of students;
3. We use BiLSTM to perform a two-way time series analysis of the joint learning features, which enables the model to analyze the learning features of students at each time step from both a past and future perspective, so as to obtain more accurate predictions;
4. We conducted sufficient experiments on five commonly used public educational datasets to demonstrate that DKT-STDRL obviously outperforms the CKT [14] and DKT [12] models on ACC, AUC, r^2, and RMSE;
5. We completed enough experiments to compare the prediction performance of DKT-STDRL with DKT [12], CKT [14], and four variants of DKT-STDRL (DKT-TDRL, DKT-SDRL1, DKT-STDRRP, and DKT-STDRRJ) when they are set up with the same hyperparameters on the same datasets, to observe the impacts of the different aspects of spatial features, temporal features, prior features, and joint features on the prediction metrics.

The rest of the paper is organized as follows: first, Section 2 provides a brief overview of the research work on KT and the research motivations. Then, in Section 3, we formally define the problem to be solved by the KT task, the implementation details of the components of the DKT-STDRL model structure, and the loss function. Section 4 compares the experimental results of DKT-STDRL with CKT, DKT, and the variants of the DKT-STDRL model on the same five open datasets. At last, the conclusion of this paper and prospective research work are pointed out in Section 5.

2. Related Work and Motivation

The work of KT is partitioned into two phases in general: the development of KT models on the basis of traditional research methods and the development of KT models on the basis of deep learning.

The first stage of KT development is from 1994 to around 2015. During this period, the research on KT was mainly based on traditional research methods, such as probability maps and psychological theories. Among them, probability map-based models represented by Bayesian knowledge tracing (BKT) [8] were the main part. BKT took the knowledge state of whether students mastered a knowledge point or not as a hidden variable, the result of whether they answered correctly the corresponding questions of the knowledge point or not as an observation variable, and used the hidden Markov model to represent the changes of the skills status of learners, and then predicted the probability that learners acquired the skill point. BKT solved the problem of KT for the first time. Subsequent research has also produced some variants of BKT that improved the performance of BKT in different

ways. For example, C. Carmona et al. proposed introducing a layer structure to represent the prerequisite relationships in actual educational scenarios, consequently acquiring a more accurate model used for predicting students' knowledge status [18]. KT-IDEM [10] was proposed to more accurately predict students' practice performance by adding the dependence of item difficulty to the probability map. BKT and its variants had simple structures and strong explanatory ability. During this period, BKT was of great significance in constructing intelligent tutorial systems and learning management systems, such as the ACT Programming Tutor [8] and edX [3]. For example, BKT improved the prediction of the test performance on the ACT Programming Tutor of students through modeling the change of the probability that the student grasped each rule [8]. EdX used variants of BKT to acquire students' practice performance on several questions in the past period, modeling students' mastery of knowledge components and, thus, predicting students' test performance in a later period. This helps to analyze the learning effect of students on the edX platform [3]. However, BKT and its variants ignored the long-term temporal dependence of students learning process because they were based on a Markov chain structure. Furthermore, BKT was not suitable for future scenarios with large amounts of data, because it relied on labeling the data manually. In addition, some scholars have applied psychological theory to KT, such as item response theory (IRT) [19], DINA [20], and so on. Although models based on psychometric measurement theory could be well interpreted, the simple parameter settings limited the models' ability to encode complex features. Therefore, though a number of effective models emerged in the first stage of KT, the model structures were too simple to express more complex features. Moreover, there was a bottleneck in the development of data processing due to the need for manual operation.

The second stage of KT is from 2015 to the present. In 2015, deep knowledge tracing (DKT) [12] was proposed, marking the beginning of an era in which DL technologies drove the evolution of KT. DKT used long short-term memory (LSTM) to learn the procedure that the learners' skill status changed and made a prediction about learners' future exercise performance. Compared with BKT, the LSTM model used by DKT could make use of long-term time series dependence, which conformed to the long-term dependence of the state of knowledge at each moment in the actual learning process and helped to obtain more accurate prediction results. In addition, DKT could automatically extract features without labeling them, which saved labor costs and avoided human errors. However, DKT only considered whether students answered correctly or not, ignoring other characteristics of the learning process. Liang Zhang et al. supplemented other relevant features, (such as time, reminder utilization, and attempt counts) collected by education information platforms to the DKT model and acquired more accurate prediction outputs [21]. Although the model could predict better by adding the exercise performance information in other aspects, these features were difficult to fully express the student's personalized learning features. Shuanghong Shen et al. proposed convolutional knowledge tracing (CKT) [14], which used convolutional neural networks (CNNs) to extract students' personalized prior knowledge and learning rates from their answer histories. Although extracting and utilizing personalized learning features could improve prediction accuracy, characterizing temporal changes was not sufficient because only a gate linear unit (GLU) [22] was used to control the forward temporal dependence of personalized prior knowledge. GritNet [17] bidirectionally analyzed students' exercise records by BiLSTM to predict students' future performance, which fully extracted the temporal features of the learning process, namely, the knowledge state for each knowledge point during each practice interaction. In the whole learning process, temporal factors such as remebering and forgetting could affect the students' mastery of knowledge at set intervals. Neural networks could learn representation features of the potential knowledge mastery state under the influence of these temporal factors. However, it did not take into account other learning features, such as personalized prior knowledge features or individualized learning rate features. Moreover, the experimental dataset, Udacity data, was not commonly used. Moreover, GritNet was only proved to be superior to the standard logistic-regression-based method.

The dynamic key-value memory network (DKVMN) [13] employed the idea of memory-augmented neural networks (MANNs), which recorded and updated the knowledge points and students' mastery of knowledge with two matrices, respectively. DKVMN had an advantage in explanation. GKT [15] built a graph neural network (GNN) [23] on the basis of the association between knowledge points, which could help improve prediction and interpretation performance [15]. Sequential key-value memory networks (SKVMN) [24] employed the triangular membership function to diagnose students' mastery of knowledge concepts and used the improved LSTM to capture the features of the learning process, which showed good performance in prediction accuracy and explanation. Qi Liu et al. attempted to combine the students' exercise performance with the features extracted from the test content, and proposed an exercise-enhanced recurrent neural network (EERNN) and EKT [25], which exploited BiLSTM, a Markov property, or an attention mechanism. The experiment proved that supplementing the feature information of the problem content could help advance the prediction effect. The self-attention mechanism was utilized in self-attentive knowledge tracing (SAKT) [26] to model learners' performance on related questions, and then to predict learners' future responses. It solved the sparse dataset problem and achieved good prediction results. Although previous research has made up for the deficiencies in different aspects of KT models and promoted the development of KT based on deep learning, most models had relatively simple feature extraction methods, leading to the inadequate use of information on students' practice records. A comparison of previous main research in KT is shown in Table 1.

As far as we knew, there were spatial features and temporal features in students learning sequences, which had location correlations and time correlations. The spatial features can be extracted by CNN. Typical CNN [27] structures were mainly composed of convolution layers and pooling layers alternately, among which the core was the convolution layer. The convolution layer contained multiple convolution kernels. The convolution kernel contained multiple parameters. By translating the convolution kernel to scan different positions of samples, parameters in the convolution kernel could learn some local features in the sample space, which was conducive to the judgment of classification. The pooling layer played the role of sub-sampling to reduce model parameters and reduce model complexity. CNN was often used for two-dimensional image processing, such as handwriting recognition [27], due to its good spatial feature extraction ability. One-dimensional sequence samples could be regarded as special two-dimensional structures, so CNN was also suitable for processing time-series data, such as natural language processing [28], and only one-dimensional convolution kernels were required. In KT, we used one-dimensional convolution to extract the spatial features of students' practice sequences. In addition, temporal features can be extracted by BiLSTM [16]. BiLSTM is a method based on LSTM [29] for bidirectional parameter learning, which consists of circularly connected memory modules. Each module contains three gate units, which are used to control the input, output, and forgetting operations of transmitted information. Through this gating mechanism, the long-term dependence of the information transfer process was solved, which was conducive to the time series analysis of samples. LSTM and BiLSTM were often used in sequence learning tasks [16], such as word sequence processing. In KT, we used BiLSTM to extract the temporal features of student exercise sequences. Therefore, we tried to integrate CNN and BiLSTM to extract the spatial features and temporal features of interaction sequences, which helped to use feature information fully [30]. In addition, the BiLSTM further improved the adequate utilization of the information by extracting features in both directions.

Table 1. Comparison of previous main research in KT. The "Category" column represents the two main types of knowledge tracing methods, the "KTs" column stands for the different knowledge tracing models, the "Year" column is the year that the knowledge tracing model was firstly published, the "Technology" column means the main machine learning or deep learning techniques used, the "Key Article" column shows the key articles involved in the method, the "Utility" column serves as the educational application scenarios, and the "Characteristics" column summarizes the main features of the knowledge tracing method.

Category	KTs	Year	Technology	AUC	Key Article	Utility	Characteristics
Probability Graph based KT	BKT	1994	HMM	0.670	Corbett et al. [8]	Programming performance prediction	Single knowledge
	KT-IDEM	2011	HMM+IRT	0.690	Pardos et al. [10]	Performance prediction on math exercises	Ignore long term dependences
Deep Learning based KT	DKT	2015	RNN/LSTM	0.805	Piech et al. [12]	Performance prediction on math courses	Multiple knowledge, single feature
	DKT+	2017	RNN/LSTM	0.863	Zhang et al. [21]	Performance prediction on statistics courses	Multiple knowledge, multiple features
	DKVMN	2017	MVNN	0.816	Zhang et al. [13]	Performance prediction on math exercises	Adaptively update knowledge mastery
	SKVMN	2019	LSTM+MVNN	0.836	Abdelrahman et al. [24]	Performance prediction on scientific courses	Multiple knowledge, long-term dependencies
	GKT	2019	GNN	0.723	Nakagawa et al. [15]	Performance prediction on math courses	Model relationship between exercises and knowledge
	EKT	2019	LSTM	0.850	Liu et al. [25]	Performance prediction on math courses	Apply semantic information of exercises
	SAKT	2019	FFN+MSA	0.848	Pandey et al. [26]	Performance prediction on scientific courses	Model relationship among knowledge points
	CKT	2020	CNN	0.825	Shen et al. [14]	Performance prediction on math exercises	Extract spatial features

3. Proposed Method

3.1. Problem Definition

The function of KT is described as follows: for a given dataset of students' exercise performance history, the interaction sequence when a student answers l times can be expressed as $\boldsymbol{U}_l = (\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_t, \ldots \boldsymbol{u}_l)$, where l is the length of the sequence. $\boldsymbol{u}_t = (s_t, r_t)$ represents the exercise record of the student at time step t, s_t is the number of the question at time step t, and r_t represents the corresponding answer result. r_t has only two values: 0 or 1. When $r_t = 1$, the student answered the question correctly at the corresponding time step. When $r_t = 0$, the student answered the question incorrectly at the corresponding time step. KT uses $\boldsymbol{U}_l$ to learn the students' skill status, for predicting whether the students can successfully solve the problem s_{t+1} at each next time step $t + 1$, namely, r_{t+1}.

3.2. Model Architecture

Deep Knowledge Tracing Based on Spatial and Temporal Deep Representation Learning for Learning Performance Prediction (DKT-STDRL) consists of 3 parts: the part of extracting the spatial features of students' learning sequences by CNN, the part of intermediate data processing, and the part of extracting students' temporal features in the learning process by BiLSTM. In the part of spatial feature extraction, a multi-layer convolutional structure [27] is used to obtain students' personalized learning efficiency information as supplementary information. The goal of the intermediate data processing part is to combine the students' practice history and personalized learning efficiency information for the subsequent sequence feature extraction part. Therefore, the intermediate data processing part merges the one-hot coded spatial features and the exercise history features into the joint features of student learning, and then passes them to the next part. The next part, namely, temporal feature extraction, adopts the BiLSTM [16] structure to further extract bidirectional time-sequence features from the student learning joint features, so as to obtain the information about the change of students' knowledge state in the learning process and then predict students' performance in the next time slice. The architecture diagram of DKT-STDRL is shown in Figure 1.

As shown in Figure 1, it can be noted that the DKT-STDRL improves the model structure of DKT [12] and CKT [14].

DKT directly used LSTM [29] to extract the temporal features (hidden knowledge state of students) of the sequences from the students' practice history, and then output the prediction results of the next time slice [12]. Compared with DKT, DKT-STDRL is improved in two aspects. First, the input information of the sequential learning structure is more abundant. The information data extracted from sequence features is no longer only the original student practice records, but also includes the spatial features extracted from the student practice sequences using CNN [27]. In this way, the spatial features can be regarded as abstract characteristics of students' personalized learning efficiency. Spatial features are added as the input of the temporal feature extraction structure for analyzing and predicting students' practice sequences from the spatial perspective before analyzing students' practice sequences from the temporal perspective. Thus, the learning analysis process of the model is more comprehensive, because the change of knowledge states of students in the learning process and the influencing factors of students' personalized learning efficiency are analyzed at the same time, and then a more accurate prediction can be obtained. Secondly, the learning ability of the structure for temporal features learning is improved. The temporal features learning structure is changed from LSTM [29] to BiLSTM [16], which enables the model to learn not only the forward sequence features, but also the reverse sequence features. The two-way learning mode enhances the ability of the model to adjust internal parameters so that it is easy to obtain more accurate prediction results. The extraction method of bidirectional temporal features is in line with the practical significance, which enables the KT model to consider students' future performance as well

as their past performance, which enables more accurate judgments to be obtained when analyzing students' knowledge mastery at each time step.

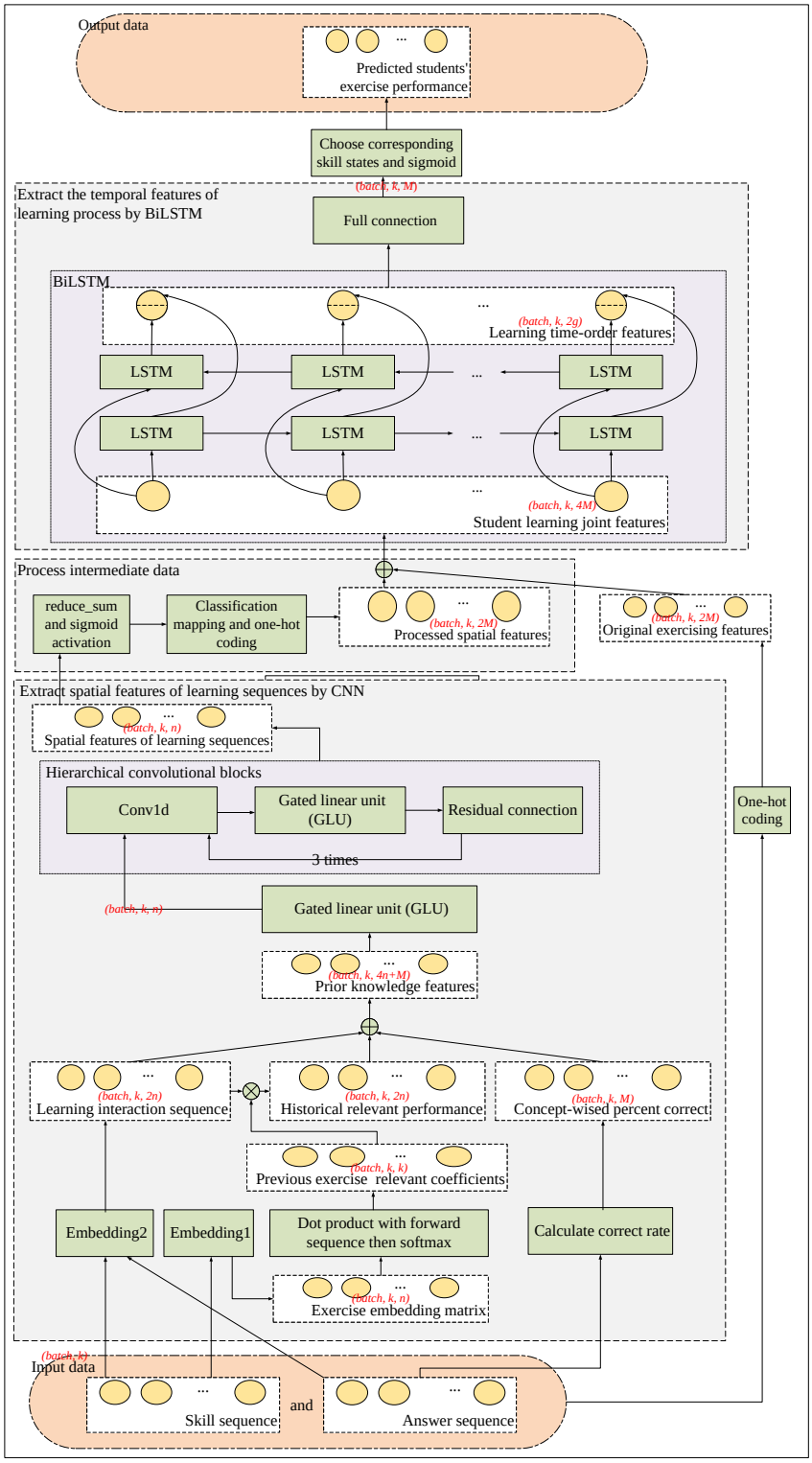

Figure 1. The architecture of the DKT-STDRL model. The red text in the figure shows the shape of the data, where *batch* equals the batch size. k represents the max sequence length. n is the number of dimensions of the embedding matrix. M is the count of different skills. g is the number of units of each LSTM module.

CKT [14] extracted the characteristics of personalized prior knowledge from students' practice records, and then extracted the characteristics of the learning rate based on the characteristics of the personalized prior knowledge using CNN [27], then outputting the prediction. Although the spatial features of student practice sequences are extracted, CKT has defects in temporal feature extraction. Compared with CKT [14], DKT-STDRL enhanced the time series feature extraction capability of CKT. After simple processing, the spatial features extracted from sequences are input into the BiLSTM structure to further extract the bidirectional temporal features. In this way, the prediction output of the model is based on the comprehensive analysis of spatial features and temporal features, which is to say that our model can simultaneously analyze the personalized learning rate of students and the change of knowledge state of students in the learning process, and it is thereby easy to obtain more accurate prediction results.

3.2.1. Using CNN to Extract Spatial Features of Students' Learning Sequences

Because CKT has significant advantages in modeling students' personalized learning interaction process [14], DKT-STDRL borrows the CKT model structure. Different from CKT, this component of the DKT-STDRL model needs only to extract the spatial features of students' learning sequences without predicting the performance of the next time step. The steps are as follows:

1. Convert the input to an embedded matrix.
   In order to make the model better extract spatial features of students' learning process, CKT uses the embedding matrix to represent the given students' exercise history $\boldsymbol{U}_k$. k is the max sequence length. Specifically, s_t is randomly initialized as an n-dimensional vector $\boldsymbol{s}_t$ by embedding; n is far less than the total counts of knowledge points in the whole dataset M. Then, according to whether the answer result is correct or not, the same n-dimensional zero vector is spliced on the right or left side of $\boldsymbol{s}_t$, so that $\boldsymbol{U}_k$ is transformed into a $k \times 2n$ dimensional matrix through embedding [14];
2. Calculate the prior knowledge of different students.
   Because the learners' prior knowledge varies with each individual and can be reflected by students' past exercise performance and correct ratio on relevant skills [31], CKT obtains the prior knowledge of different students by calculating values for historical relevant performance (HRP) and concept-wise percent correct (CPC) [14];
3. Calculate the learning rate of different students.
   Because the learning efficiency of different students is different, and the continuous exercise performance of students over a period can reflect the learning efficiency of different students in different learning stages, CKT extracts the learning rate of different students through multi-layer CNN.

Through the above process, the spatial learning features of students' learning sequences can be obtained. The process is visualized by the formula as follows:

$$\boldsymbol{e}_t = \begin{cases} [\boldsymbol{s}_t \oplus \boldsymbol{0}], & \text{if } r_t = 1 \\ [\boldsymbol{0} \oplus \boldsymbol{s}_t], & \text{if } r_t = 0 \end{cases} \tag{1}$$

$$\boldsymbol{F}_{LIS} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k)^\top \tag{2}$$

$$\begin{cases} \boldsymbol{relation}t(j) = \text{ Masking } (\boldsymbol{s}_j \cdot \boldsymbol{s}_t), j \in (t,k) \\ \boldsymbol{weight}_t(j) = \text{ Softmax } (\boldsymbol{relation}_t(j)), j \in (1,k) \end{cases} \tag{3}$$

$$\boldsymbol{F}_{HRP_t}(t) = \sum_{j=1}^{t-1} \boldsymbol{weight}_t(j)\boldsymbol{e}_j \tag{4}$$

$$\boldsymbol{F}_{CPC_t}(m) = \frac{\sum_{j=0}^{t-1} r_j^m == 1}{\text{total}(s^m)} \tag{5}$$

$$\begin{cases} \boldsymbol{F}_{JPF} = \boldsymbol{F}_{LIS} \oplus \boldsymbol{F}_{HRP} \oplus \boldsymbol{F}_{CPC} \\ \boldsymbol{F}_{ILA} = \left(\boldsymbol{F}_{JPF} * \boldsymbol{W}_3 + \boldsymbol{b}_3\right) \otimes \sigma\left(\boldsymbol{F}_{JPF} * \boldsymbol{W}_4 + \boldsymbol{b}_4\right) \end{cases} \tag{6}$$

$$\boldsymbol{F}_{KS} = \{GLU(\text{Conv}\,1d(\boldsymbol{F}_{ILA}))\}_3 \tag{7}$$

where $\boldsymbol{F}_{LIS}$ is the embedded matrix expression of the learning interaction sequence $\boldsymbol{U}_k$. *Masking* is an operation used to exclude subsequences. ***weight*** is used to estimate the resemblance between the present question and the previous question, and to then comprehensively analyze the impact of previous exercise performance on the current exercise. $\boldsymbol{F}_{HRP}$ is a historically related performance feature. m is the number of the knowledge point. $\boldsymbol{F}_{CPC_t}(m)$ indicates the correct rate when the student answers the question of a knowledge point at a certain time. $\boldsymbol{F}_{JPF}$ is the prior knowledge feature obtained by connecting $\boldsymbol{F}_{LIS}$, $\boldsymbol{F}_{HRP}$, and $\boldsymbol{F}_{CPC}$. The spatial features of people's studying processes $\boldsymbol{F}_{KS}$ can be obtained from $\boldsymbol{F}_{JPF}$ through 3 layers of nonlinear transformations of a gated linear unit (GLU) and one-dimensional convolution operation [14].

3.2.2. Intermediate Data Processing

The main goal of intermediate data processing is to take the spatial features extracted from the previous part as supplementary information regarding students' original exercise features, and then input them into the next part to extract the temporal features of students' learning process. Specifically, firstly, the spatial learning features $\boldsymbol{F}_{KS}$ are activated by the sigmoid function. Then, the processed spatial features are suitable for the next part after classification and one-hot coding. Finally, the spatial features are combined with the student learning record sequences processed by one-hot coding to form joint features as the input of BiLSTM. The process is shown by the formula as follows:

$$\boldsymbol{F}'_{KS} = \sigma(\boldsymbol{F}_{KS}) \tag{8}$$

$$\boldsymbol{F}''_{KS}(t) = \begin{cases} 1, & \text{if } \boldsymbol{F}'_{KS} > 0.5 \\ 0, & \text{if } \boldsymbol{F}'_{KS} \leq 0.5 \end{cases} \tag{9}$$

$$\boldsymbol{AP}_1(t) = \boldsymbol{F}''_{KS}(t) \cdot M + s_t \tag{10}$$

$$\boldsymbol{F}_{ILF} = OneHot(\boldsymbol{AP}_1, 2M) \tag{11}$$

$$\boldsymbol{AP}_2(t) = (1 - r_t) \cdot M + s_t \tag{12}$$

$$\boldsymbol{F}_{LRS} = OneHot(\boldsymbol{AP}_2, 2M) \tag{13}$$

$$\boldsymbol{F}_{LJF}(t) = \boldsymbol{F}_{ILS}(t) \oplus \boldsymbol{F}_{LRS}(t) \tag{14}$$

where $\boldsymbol{AP}_1$ and $\boldsymbol{AP}_2$ are used to calculate the position of 1, which is used for the one-hot coding transformation of the spatial features and learning record sequence. $\boldsymbol{F}_{ILF}$ represents the spatial features encoded by one-hot, and $\boldsymbol{F}_{LRS}$ represents the learning record sequences encoded by one-hot. The two are connected and combined to form the joint feature of students' learning $\boldsymbol{F}_{LJF}$. $\boldsymbol{F}_{LJF}$ is a $k \times 4M$ matrix.

3.2.3. Using BiLSTM to Extract the Temporal Features of Students' Learning Process

Inspired by GritNet [17], this part of the model introduces BiLSTM to extract the temporal features of students' studying procedures and make a prediction about learners' recent learning effect. Because BiLSTM can pass information forward and backward at the same time, the DKT model with the BiLSTM structure can make use of students' past

exercise performance information and future exercise performance information at the same time. Compared with the original DKT consisting of the LSTM structure, which can only use past information, the DKT based on the BiLSTM structure can use more sufficient information to adjust the parameters of the model, so as to obtain more accurate judgment results. The processing process of intermediate data in this part is as follows:

$$\overrightarrow{h_t} = \text{LSTM}\left(F_{LJF}(t), \overrightarrow{h_{t-1}}\right) \tag{15}$$

$$\overleftarrow{h_t} = \text{LSTM}\left(F_{LJF}(t), \overleftarrow{h_{t+1}}\right) \tag{16}$$

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \tag{17}$$

$$h'_t = h_t * W_5 + b_5 \tag{18}$$

where $\overrightarrow{h_t}$ is the hidden state in chronological order when it is time step t. $\overleftarrow{h_t}$ is the hidden state in reverse time order when it is time step t. $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are combined to obtain h_t, and h_t is the temporal features of students' learning process extracted by BiLSTM. h'_t can be regarded as the hidden knowledge state after comprehensive analysis. Through h'_t and an exercise sequence, the exercise performance at the next time step p_{t+1} can be easily obtained by simply finding the knowledge state of the corresponding problem and being dealt with by the sigmoid function.

3.3. Optimization

In order to optimize the model, the loss function of cross-entropy and the Adam optimizer are adopted. The loss function is as follows:

$$\text{Loss} = -\sum_{i=1}^{l}(r_i \log p_i + (1 - r_i)\log(1 - p_i)) \tag{19}$$

4. Results and Discussion

4.1. Experimental Datasets

For the purpose of fair comparison with other KTs, experiments were conducted on the same public datasets to obtain the models' performance. These datasets were the classical datasets that are commonly exploited in KT research. Descriptions of the datasets are shown in Table 2.

Table 2. Introduction of the datasets.

Dataset	Students	Skills	Records
ASSISTment2009	4151	110	325,637
ASSISTment2015	19,840	100	683,801
Synthetic-5	4000	50	200,000
ASSISTchall	1709	102	942,816
Statics2011	333	1223	189,297

The ASSISTment2009 dataset comes from the log data of students performing math exercises in 2009, which were collected by the ASSISTments platform [32]. The original version of the dataset contained 123 skills. Skills corresponded to exercise tags, such as prime numbers or linear equations in math problems. For the convenience of representation, the exercise tags were mapped to some numbers in the dataset. In other words, a skill ID represented an exercise tag in these KT datasets. Later, it was found that there were duplicate records of the data, which affected the reliability of the prediction results of the DKT model [33]. Therefore, experiments in this paper adopted the version after removing

duplicate records. The total number of skills involved in this version of the dataset has been reduced to 110.

The ASSISTment2015 dataset comes from the data collected by the ASSISTments platform in 2015.

The Synthetic-5 dataset is from the paper which proposed DKT. The dataset is a set of unreal datasets constructed for experiments and does not correspond to the actual skills [12]. However, the dataset is of good quality and is still suitable for many experimental studies in KT founded on DL.

The ASSISTchall dataset was gathered when learners utilized the ASSISTments platform and was used for an educational data mining competition in 2017.

The Statics2011 dataset originates from online college-level statistics lessons [34].

When using the datasets mentioned above, the student whose records are less than or equal to two were regarded as invalid records and were removed from the sample sets, as the student records with too few practice records could not reflect the real learning situation of the student and were not applicable to the time-series model [21]. There were various situations leading to a student only practicing twice. The student may have already mastered a particular skill, and thus did not need practice. The student may also have given up the exercise because it was too difficult. So, the student records with two practice records did not seem reliable enough to reflect students' real learning situation. The training of neural networks relies on reliable samples on a large scale to obtain accurate results. So, we removed the students with less than two practice records, which could affect the whole model.

A portion of the dataset is shown in Figure 2.

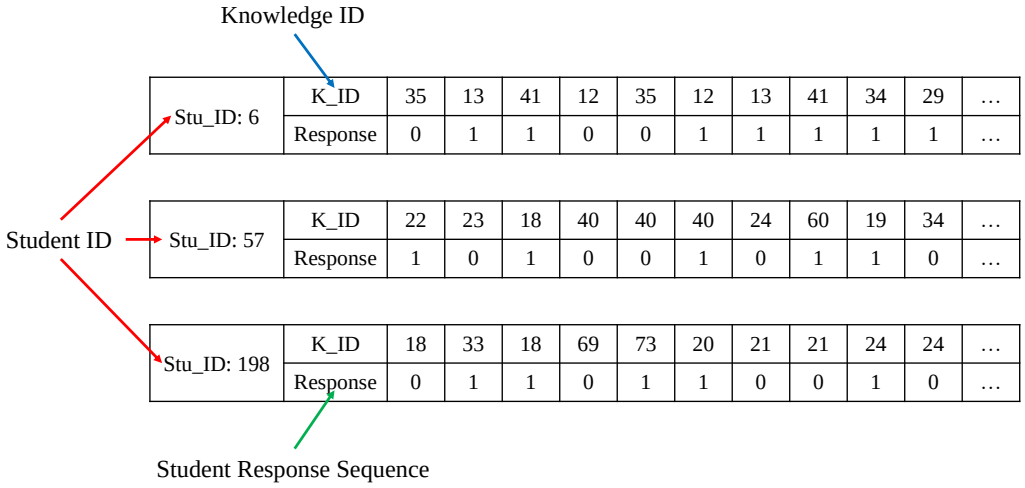

Figure 2. Description of a portion of the dataset.

4.2. Experimental Environment

The experiments for analyzing the models cannot be completed without the support of an effective software and hardware environment. A description of the main hardware and software environment configurations is shown in Table 3.

Table 3. Description of the experimental environment.

Configuration Environment	Configuration Parameters
Operating System	Windows 10 64-bit
GPU	gtx 1080ti
CPU	E5 Series (4 cores)
Memory	16 GB
Programming language	Python 3.6
Deep learning framework	Tensorflow 1.5
Python library	Numpy, Scikit-learn, Pandas, Matplotlib

4.3. Results and Discussion

During our experiments, we partitioned each dataset into three parts: a training set, a validation set, and a test set. They, respectively, accounted for about 55%, 15%, and 30% of the total students. The training set was used for training the model. The validation set was used for adjusting hyperparameters. The test set was used for evaluating the model. To obtain more reliable results, the experiments of each model on each dataset were repeated three times, and the average values were taken as the evaluation result in terms of RMSE, AUC, ACC, and r^2. Considering the balance of the calculation resource and the prediction accuracy, the hyperparameters were set as in Table 4. In fact, the DKT-STDRL model can obtain more accurate predictions by changing the hyperparameters, regardless of calculation resources.

Table 4. Hyperparameters of the DKT-STDRL.

Hyperparameters	Value
Learning rate	0.001
Rate of decay for the learning rate	0.3
Steps before the decay	8
Batch size	50
Epochs	10
Shape of filters of the conv1d	(16, 50, 50)
Layers of the hierarchical convolution	3
Keep probability for dropout of the convolution	0.2
Number of units of each LSTM cell	30
Output keep probability of LSTM cell	0.3

Baselines comparison. For the sake of better verifying the validity of the new model, DKT-STDRL was compared with DKT based on LSTM and CKT on the same dataset. As the starting model of deep knowledge tracing, the DKT model based on LSTM had an important reference value. In addition, because DKT-STDRL is an improvement based on the CKT model, the comparison with the CKT model was essential. We refer to and reproduce the code of DKT and CKT as baselines, and obtain the evaluation results approach to the reported results.

Table 5 compares the prediction effects of DKT-STDRL with DKT and CKT. It was found in the experiments on the five datasets that our DKT-STDRL outperforms the DKT and the CKT in terms of RMSE, AUC, ACC, and r^2. Specifically, for ASSISTments2009, DKT-STDRL achieved an RMSE of 0.2826. It decreased by 11.8% and 10.86% compared, respectively, with DKT and CKT. DKT-STDRL achieved an AUC of 0.9591. It increased by 15.42% and 13.54% compared, respectively, with DKT and CKT. DKT-STDRL achieved an ACC of 0.8904. It increased by 12.41% and 11.56% compared, respectively, with DKT and CKT. DKT-STDRL achieved r^2 of 0.644. It increased by 36.05% and 32.59% comparedm respectively, with DKT and CKT. For ASSISTments2015, Synthetic-5, ASSISTchal, and Statics2011, DKT-STDRL had similar performance. Experiments verify the prediction effect of the model from two aspects: classification and regression. AUC and ACC of DKT-STDRL

were gained. Therefore, the DKT-STDRL model can better classify the prediction results of whether students can answer correctly or not. Moreover, from the perspective of regression, the reduction of RMSE with DKT-STDRL reflects that the new model is more accurate in predicting the probability of students answering correctly or incorrectly. In addition, the r^2 of DKT-STDRL has been significantly improved, which shows that the prediction results of the new model are highly correlated with the actual exercise performance of students. The new model learns the essential law of the change of students' knowledge state from the sample data. Therefore, the DKT-STDRL model promotes present KT models predicting more accurately.

Table 5. Comparison of experimentation results among DKT-STDRL, DKT, and CKT models.

Datasets	Models	RMSE	AUC	ACC	r^2
ASSISTment2009	**Ours**	**0.2826**	**0.9591**	**0.8904**	**0.6440**
	DKT	0.4006	0.8049	0.7663	0.2835
	CKT	0.3912	0.8237	0.7748	0.3181
ASSISTment2015	**Ours**	**0.0766**	**0.9996**	**0.9948**	**0.9698**
	DKT	0.4131	0.7235	0.7504	0.1235
	CKT	0.4107	0.7322	0.7542	0.1338
Synthetic-5	**Ours**	**0.3167**	**0.9482**	**0.8790**	**0.5766**
	DKT	0.4109	0.8167	0.7475	0.2868
	CKT	0.4051	0.8279	0.7553	0.3075
ASSISTchall	**Ours**	**0.2716**	**0.9710**	**0.9078**	**0.6847**
	DKT	0.4538	0.7022	0.6791	0.1198
	CKT	0.4500	0.7127	0.6857	0.1342
Statics2011	**Ours**	**0.2772**	**0.9499**	**0.9010**	**0.5621**
	DKT	0.3697	0.8012	0.8054	0.2216
	CKT	0.3630	0.8232	0.8101	0.2496

Comparison of the variants of the DKT-STDRL. The experiments for comparing DKT-STDRL with its variants, by removing different parts, are helpful to understand the importance of various components in the DKT-STDRL model for advancing the prediction effect of the whole model. Therefore, in addition to the CKT and DKT, four variants were designed by us to observe the different impacts on the prediction effect from the aspects of spatial features, temporal features, prior features, and joint features. The four variants are described as follows:

- *DKT-TDRL.* To study the influence of spatial features on the prediction effect of the DKT-STDRL model, we removed the part of extracting spatial features with CNN from the DKT-STDRL model and obtaine dthe variant model DKT-TDRL. Specifically, the DKT-TDRL model first takes the input data represented by the embedding matrix and the students' personalized prior knowledge state as the prior feature. Then, through the intermediate data processing process, it is input into BiLSTM for bidirectional time feature extraction. Finally, the prediction result is output;
- *DKT-SDRL1.* For the purpose of studying the impact of time features of DKT-STDRL on the prediction effect, the part of extracting time features was removed from the DKT-STDRL model. Since DKT-STDRL adopts BiLSTM to the express bidirectional feature of the sequence, two schemes can be obtained to study the influence of unidirectional and bidirectional temporal feature extraction, respectively. The first scheme is to remove the bidirectional temporal feature extraction structure of the DKT-STDRL model. Following this idea, to obtain the prediction results of the next time step, the model is transformed into the CKT model. The second scheme is to change the BiLSTM structure of the DKT-STDRL model into the LSTM structure, which can show the influence of one-way temporal output on the prediction effect of the model. Then, we can compare the difference between two-way time characteristics and one-way time

characteristics in solving the prediction problem of students' learning performance. We abbreviate the second scheme as DKT-SDRL1. Specifically, DKT-SDRL1 first extracts and uses the prior learning features, and then a multi-layer convolution structure is used to extract students' spatial learning features. Then, after a simple intermediate process, it is input to the one-way temporal feature extraction layer based on LSTM. Finally, the prediction results are output;

- *DKT-STDRRP.* In order to study the influence of the prior learning features of the DKT-STDRL model on the prediction effect, we removed the part of extracting the prior learning features from the DKT-STDRL model. This variant model is called DKT-STDRRP. DKT-STDRRP transforms the input students' learning history data into an embedded matrix, and then extracts spatial features through CNN layers and extracts bidirectional temporal features through the BiLSTM layer, so as to predict students' learning performance. By comparing the prediction results before and after removing the prior learning features, it is helpful to intuitively understand the role of the prior features in the task of predicting learning performance;
- *DKT-STDRRJ.* For studying the impact of the joint features in the DKT-STDRL model on the prediction, the process of merging the one-hot coded exercise history data is removed from the DKT-STDRL model in the intermediate data processing. In other words, the variant scheme keeps the structure of the first half of the DKT-STDRL model unchanged, inputs the one-hot coded spatial features directly into the temporal feature extraction part, and then obtains the final prediction result. This scheme is called DKT-STDRRJ.

Table 6 shows the prediction performance of DKT-STDRL, CKT, DKT, and the variants of DKT-STDRL (DKT-TDRL, DKT-SDRL1, DKT-STDRRP, and DKT-STDRRJ) on the ASSISTment2009, ASSISTment2015, Synthetic-5, ASSISTchall, and Statics2011 datasets. It should be noted that different models are based on the same hyperparameters in the experiments on the same datasets. Because as parts of the DKT-STDRL, other models should keep the same settings to make only the model structure different in each experiment.

In order to more intuitively compare the prediction effect of these models, Figure 3 displays the prediction results of all models on the ASSISTment2009 dataset in the bar charts. Figure 4 displays the prediction results of all models on the ASSISTment2015 dataset in the bar charts. Figure 5 displays the prediction results of all models on the Synthetic-5 dataset in the bar charts. Figure 6 displays the prediction results of all models on the ASSISTchall dataset in the bar charts. Figure 7 displays the prediction results of all models on the Statics2011 dataset in the bar charts. From these figures, it can be seen, obviously, that the DKT-STDRL had a higher AUC, ACC, and r^2, and a lower RMSE, than CKT, DKT, DKT-SDRL1, and DKT-STDRRJ. Moreover, the AUC, ACC, and r^2 of the DKT-STDRL were a little higher than DKT-TDRL and DKT-STDRRP, and the RMSE of DKT-STDRL was a little lower than DKT-TDRL and DKT-STDRRP. The model with higher AUC, ACC, and r^2, and lower RMSE, is better. So, DKT-STDRL is significantly better than CKT, DKT, DKT-SDRL1, and DKT-STDRRJ. Moreover, DKT-STDRL is only a little better than DKT-TDRL and DKT-STDRRP. Firstly, DKT-TDRL, regardless of extracting spatial features, raised the RMSE no more than 2% and dropped the AUC, ACC, r^2 no more than 2%, 0.7%, and 5%, respectively. So, the spatial features have little benefit in improving prediction results. Secondly, different from DKT-STDRL, DKT-SDRL1 learns unidirectional temporal features instead of bidirectional features. As the results show, DKT-SDRL1 is significantly worse than DKT-STDRL. Thus, the extracted bidirectional temporal features are very important for DKT-STDRL. Thirdly, DKT-STDRL has many advantages compared with DKT-STDRRP. Changing from DKT-STDRRP to DKT-STDRL can decrease the RMSE by over 2% and increase AUC, ACC, and r^2 by over 2%, 1.8%, 6.8%, repsectively. So, prior learning features are also meaningful for obtaining a better model. Fourthly, DKT-STDRRJ has obvious weaknesses in all the metrics. It can be inferred that merging exercise performance history information with spatial features is necessary for the DKT-STDRL because the extracted spatial features have information errors compared with the original data and cannot be

directly input into BiLSTM. So, the spatial features, temporal features, prior features, and joint features of DKT-STDRL play various roles in improving prediction.

Table 6. Comparison of experimentation results among the DKT-STDRL, CKT, DKT, and variants of DKT-STDRL.

Datasets	Models	RMSE	AUC	ACC	r^2
ASSISTment2009	**DKT-STDRL**	**0.2826**	**0.9591**	**0.8904**	**0.6440**
	DKT-TDRL	0.2845	0.9574	0.8896	0.6394
	CKT	0.3953	0.8154	0.7687	0.3039
	DKT-SDRL1	0.4265	0.7567	0.7370	0.1894
	DKT	0.4293	0.7509	0.7338	0.1788
	DKT-STDRRP	0.2878	0.9557	0.8861	0.6310
	DKT-STDRRJ	0.4531	0.6736	0.6890	0.0853
ASSISTment2015	**DKT-STDRL**	**0.0766**	**0.9996**	**0.9948**	**0.9698**
	DKT-TDRL	0.0799	0.9995	0.9942	0.9672
	CKT	0.4099	0.7343	0.7546	0.1370
	DKT-SDRL1	0.4176	0.7070	0.7475	0.1043
	DKT	0.4183	0.7042	0.7468	0.1015
	DKT-STDRRP	0.0801	0.9995	0.9943	0.9671
	DKT-STDRRJ	0.4053	0.7581	0.7511	0.1563
Synthetic-5	**DKT-STDRL**	**0.3167**	**0.9482**	**0.8790**	**0.5766**
	DKT-TDRL	0.3186	0.9469	0.8776	0.5714
	CKT	0.4081	0.8241	0.7513	0.2972
	DKT-SDRL1	0.4496	0.7283	0.6754	0.1469
	DKT	0.4505	0.7269	0.6752	0.1432
	DKT-STDRRP	0.3194	0.9461	0.8768	0.5693
	DKT-STDRRJ	0.4716	0.6485	0.6417	0.0612
ASSISTchall	**DKT-STDRL**	**0.2716**	**0.9710**	**0.9078**	**0.6847**
	DKT-TDRL	0.2790	0.9675	0.9009	0.6671
	CKT	0.4503	0.7119	0.6846	0.1330
	DKT-SDRL1	0.4683	0.6431	0.6642	0.0627
	DKT	0.4682	0.6430	0.6648	0.0628
	DKT-STDRRP	0.2854	0.9633	0.8942	0.6518
	DKT-STDRRJ	0.4493	0.7170	0.6935	0.1371
Statics2011	**DKT-STDRL**	**0.2772**	**0.9499**	**0.9010**	**0.5621**
	DKT-TDRL	0.2891	0.9395	0.8956	0.5220
	CKT	0.3630	0.8241	0.8099	0.2496
	DKT-SDRL1	0.3724	0.7942	0.8042	0.2103
	DKT	0.3739	0.7891	0.8023	0.2036
	DKT-STDRRP	0.2974	0.9275	0.8827	0.4940
	DKT-STDRRJ	0.3756	0.7873	0.7995	0.1967

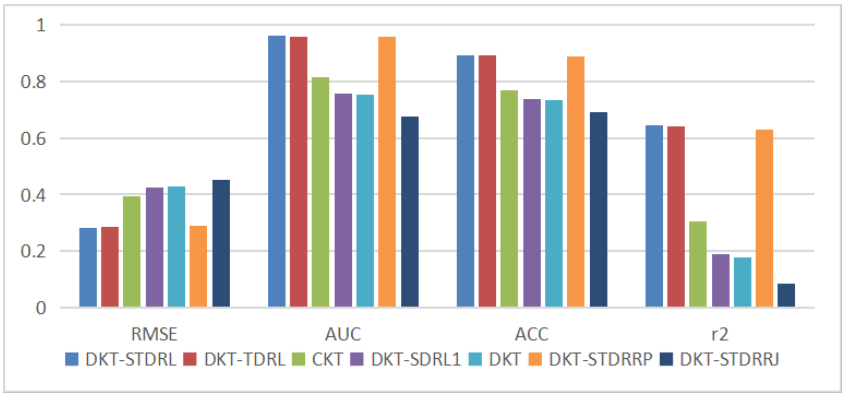

Figure 3. Prediction results of DKT-STDRL, CKT, DKT, and the variants of DKT-STDRL on the ASSISTments2009 dataset.

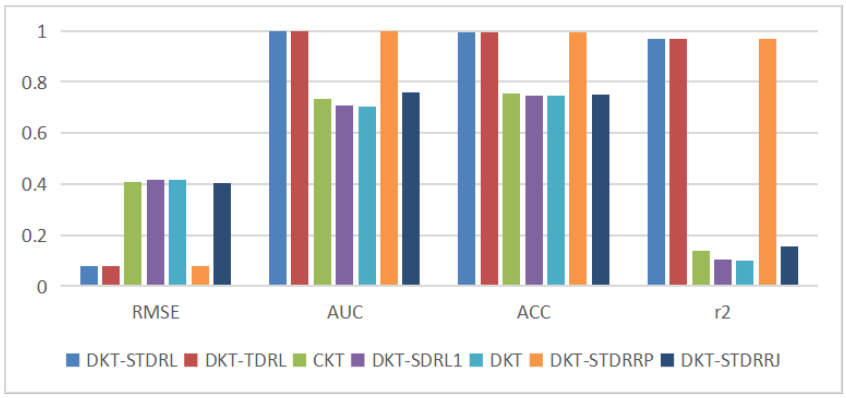

Figure 4. Prediction results of DKT-STDRL, CKT, DKT, and the variants of DKT-STDRL on the ASSISTments2015 dataset.

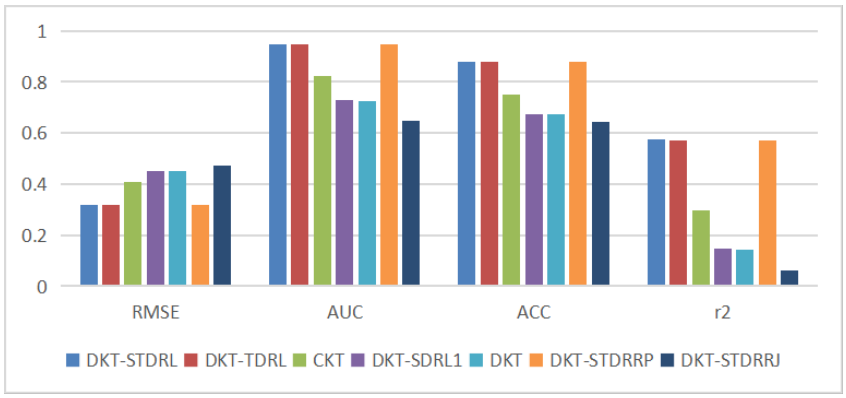

Figure 5. Prediction results of DKT-STDRL, CKT, DKT, and the variants of DKT-STDRL on the Synthetic-5 dataset.

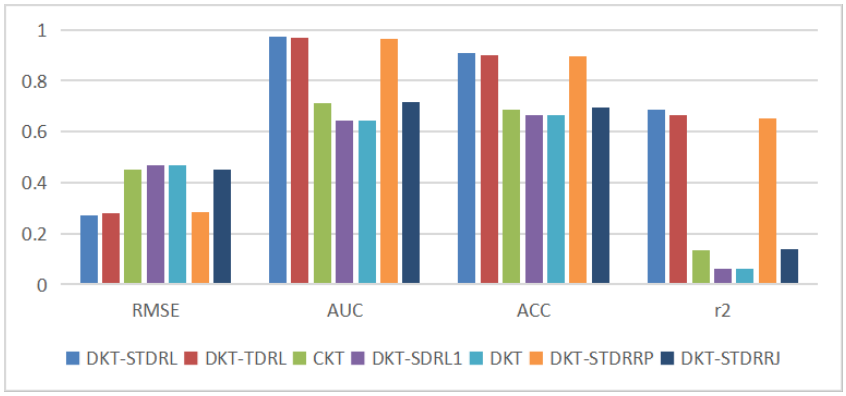

Figure 6. Prediction results of DKT-STDRL, CKT, DKT, and the variants of DKT-STDRL on the ASSISTchall dataset.

From the previous two types of comparison experiments, it can be found that our DKT-STDRL has better prediction accuracy than CKT [14] and DKT [12]. Analyzing the functions of the various parts of DKT-STDRL, the spatial features, temporal features, prior features, and joint features of DKT-STDRL all contribute to improving the prediction accuracy to varying degrees. In short, the success of DKT-STDRL in improving the prediction accuracy is inseparable from the comprehensive analysis of the spatial and temporal characteristics of students' practice sequences. In other words, for KT, a comprehensive consideration of students' personalized learning efficiency and characteristics of students' knowledge state change process is helpful to obtain more accurate prediction results.

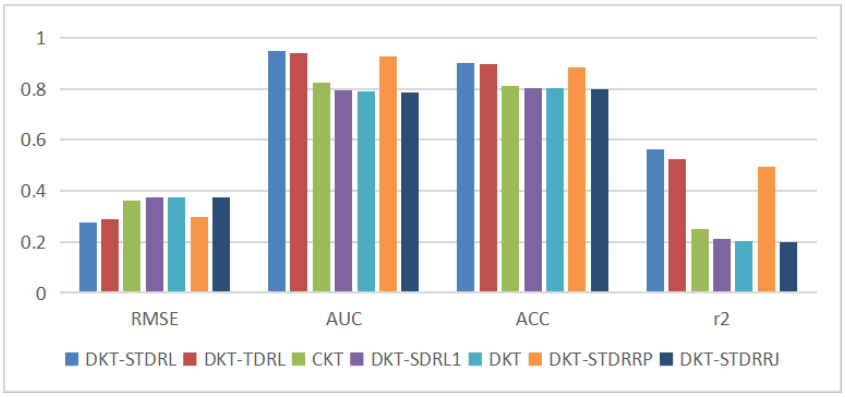

Figure 7. Prediction results of DKT-STDRL, CKT, DKT, and the variants of DKT-STDRL on the Statics2011 dataset.

Our model significantly improved the prediction accuracy of KT, and it had practical significance for different education stakeholders, such as students, LMS administrators, teachers, and instructional designers [2]. First, LMSs which use our method can analyze students' potential knowledge state based on the students' practice logs and make an accurate prediction of students' performance in the next stage . Therefore, our method can help the LMSs to obtain information about students' future learning performance. LMSs can generate reports for students or teachers to help system users better find the problems in learning. LMSs can also learn the knowledge points that students have not mastered from the prediction information of students' learning performance, so as to recommend content suitable for students' further learning. In this way, LMSs can send exercises that students really need without wasting too much time doing exercises they have grasped. Secondly, LMSs using our KT model can provide teachers and instructional designers with more accurate analysis reports on students' learning conditions based on the prediction of students' answer performance, so as to help teachers and instructional designers flexibly adjust teaching plans. Through the learning reports provided by the systems, teachers and instructional designers can prepare lessons efficiently around students' knowledge defects. With the continuous improvement of teaching methods, the student-centered outcome-based education (OBE) concept can be realized. Thirdly, the administrators of LMSs adopting our KT model can provide better consulting services for customers. Administrators can also divide classes according to the degree of knowledge mastery of students and assign special teachers to the classes for guidance, which is convenient for class management and improves the overall teaching effect for students. So, our KT model can promote the development of intelligent education.

5. Conclusions

For filling the gap of current KT models that express students' learning features insufficiently, Deep Knowledge Tracing Based on Spatial and Temporal Deep Representation Learning for Learning Performance Prediction (DKT-STDRL) was put forward. DKT-STDRL extracted the spatial features on the basis of students' exercise history, and then further extracted the temporal features of students' exercise sequence. Firstly, DKT-STDRL used CNN to extract the spatial feature information of students' exercise history. Then, the spatial features were connected with the exercise history features and input into the BiLSTM part as joint learning features. Finally, BiLSTM extracted the temporal feature from the joint learning features to obtain the prediction information of whether the students could answer correctly or not at the next time step. The prediction effect of the DKT-STDRL model was verified on five common public datasets. The experimentations demonstrated that the prediction of the DKT-STDRL outperformed the DKT and CKT. Therefore, DKT-STDRL is effective for promoting the prediction accuracy of the KT model on the basis of DL. Moreover, many experiments were conducted to compare the prediction performance of DKT-STDRL with CKT, DKT, and four variants of DKT-STDRL, which showed the different

impacts on the prediction effect from the aspects of spatial features, temporal features, prior features, and joint features.

Though we have succeeded in advancing the prediction accuracy, there are still some limitations of our work on the complexity and interpretability of the model. First, the structure of the model is complex and has too many parameters. Secondly, the parameters in the DL networks lack interpretability, which limits the significance of the model in practical applications. So, in prospective work, we strive to reduce the complexity of the model to save the computing resources and to improve the interpretability to provide more application value. In addition, in the future, because the model has good prediction accuracy, we can try to integrate the model into LMSs for a better recommendation, so that we obtain more intelligent and caring online learning systems. By using our improved KT algorithm, LMSs can obtain more accurate information about students' knowledge mastery or prediction results of answer performance. Based on this information, LMSs can intelligently recommend learning resources (such as lectures, documents, exercises, and quizzes) that better meet students' needs, thus helping students reduce their learning burden and improve their learning efficiency.

Author Contributions: Methodology research, L.L. and Z.W.; model realization, L.L.; supervision Z.W.; writing and editing, L.L. and Z.W.; data collection, L.L., Z.W., H.Y., Z.Y., and Y.L.; model evaluation, L.L., Z.W., H.Y., Z.Y., and Y.L.; funding acquisition, L.L., Z.W., H.Y., Z.Y., and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Youth Science Foundation of Heilongjiang Institute of Technology (2021QJ07), the National Natural Science Foundation of China (Nos. 62177022, 61901165, 61501199), the Collaborative Innovation Center for Informatization and Balanced Development of K-12 Education by MOE and Hubei Province (No. xtzd2021-005), and Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU20ZT010), the Natural Science Foundation of Heilongjiang Province (LH2020F047), the Innovation Team Project of Heilongjiang Institute of Technology (2020CX07), the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province (UNPYSCT-2020052), and the Education and Teaching Reform Research Project of Heilongjiang Institute of Technology (JG202109).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The ASSISTment2009, ASSISTment2015, and ASSISTchall datasets are from the ASSISTments platform [32]. Synthetic-5 is a simulated dataset used by the DKT model, which can be obtained through the deep knowledge tracing paper [12]. The Statics2011 dataset originates from online college-level statistics lessons [34]. In the paper of CKT [14], these datasets have been organized and can be found by following the following link: https://github.com/bigdata-ustc/Convolutional-Knowledge-Tracing.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LMS	Learning Management System
KT	Knowledge Tracing
DKT-STDRL	Deep Knowledge Tracing Based on Spatial and Temporal Deep Representation Learning for Learning Performance Prediction
BKT	Bayesian Knowledge Tracing
HMM	Hidden Markov Model
KT-IDEM	Knowledge Tracing: Item Difficulty Effect Model
PC-BKT	Personalized Clustered BKT
RNN	Recurrent Neural Network
DL	Deep Learning
DKVMN	Dynamic Key-Value Memory Network

MANN	Memory-Augmented Neural Network
CNN	Convolutional Neural Network
CKT	Convolutional Knowledge Tracing
GKT	Graph-based Knowledge Tracing
GNN	Graph Neural Network
BiLSTM	Bidirectional Long Short-Term Memory
IRT	Item Response Theory
LSTM	Long Short-Term Memory
GLU	Gate Linear Unit
SKVMN	Sequential Key-Value Memory Networks
EERNN	Exercise-Enhanced Recurrent Neural Network
SAKT	Self-Attentive Knowledge Tracing
HRP	Historical Relevant Performance
CPC	Concept-wise Percent Correct
RMSE	Root Mean Squared Error
AUC	Area Under Curve
ACC	Accuracy
OBE	Outcome-Based Education

References

1. Khlaif, Z.N.; Salha, S.; Affouneh, S.; Rashed, H.; ElKimishy, L.A. The COVID-19 Epidemic: Teachers' Responses to School Closure in Developing Countries. *Technol. Pedagog. Educ.* **2021**, *30*, 95–109. https://doi.org/10.1080/1475939X.2020.1851752.
2. Cavus, N. Distance Learning and Learning Management Systems. *Procedia Soc. Behav. Sci.* **2015**, *191*, 872–877. https://doi.org/10.1016/j.sbspro.2015.04.611.
3. Pardos, Z.; Bergner, Y.; Seaton, D.; Pritchard, D. Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX. In Proceedings of the International Conference on Educational Data Mining, Memphis, TN, USA, 6–9 July 2013; Citeseer: Princeton, NJ, USA, 2013.
4. Heffernan, N.T.; Ostrow, K.S.; Kelly, K.; Selent, D.; Van Inwegen, E.G.; Xiong, X.; Williams, J.J. The Future of Adaptive Learning: Does the Crowd Hold the Key? *Int. J. Artif. Intell. Educ.* **2016**, *26*, 615–644. https://doi.org/10.1007/s40593-016-0094-z.
5. Gorshenin, A. Toward Modern Educational IT-ecosystems: From Learning Management Systems to Digital Platforms. In Proceedings of the 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Moscow, Russia, 5–9 November 2018; IEEE: Piscataway, NJ, USA, 2018, pp. 1–5. https://doi.org/10.1109/ICUMT.2018.8631229.
6. Hernández-Blanco, A.; Herrera-Flores, B.; Tomás, D.; Navarro-Colorado, B. A Systematic Review of Deep Learning Approaches to Educational Data Mining. *Complexity* **2019**, *2019*, 1306039. https://doi.org/10.1155/2019/1306039.
7. Teodorescu, O.; Popescu, P.; Mocanu, M.; Mihaescu, C. Continuous Student Knowledge Tracing Using SVD and Concept Maps. *Adv. Electr. Comp. Eng.* **2021**, *21*, 75–82. https://doi.org/10.4316/AECE.2021.01008.
8. Corbett, A.T.; Anderson, J.R. Knowledge tracing: Modeling the Acquisition of Procedural Knowledge. *User Model. User-Adapt. Interact.* **1995**, *4*, 253–278. https://doi.org/10.1007/BF01099821.
9. Rastrollo-Guerrero, J.L.; Gómez-Pulido, J.A.; Durán-Domínguez, A. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Appl. Sci.* **2020**, *10*, 1042. https://doi.org/10.3390/app10031042.
10. Pardos, Z.A.; Heffernan, N.T. KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. In Proceedings of the User Modeling, Adaption and Personalization, Girona, Spain, 11–15 July 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 243–254. https://doi.org/10.1007/978-3-642-22362-4_21.
11. Nedungadi, P.; Remya, M.S. Predicting Students' Performance on Intelligent Tutoring System — Personalized Clustered BKT (PC-BKT) Model. In Proceedings of the 2014 IEEE Frontiers in Education Conference (FIE) Proceedings, Madrid, Spain, 22–25 October 2014; IEEE Computer Society: Los Alamitos, CA, USA, 2014; pp. 1–6. https://doi.org/10.1109/FIE.2014.7044200.
12. Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.; Sohl-Dickstein, J. Deep Knowledge Tracing. In Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15, Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; Volume 1, pp. 505–513.
13. Zhang, J.; Shi, X.; King, I.; Yeung, D.Y. Dynamic Key-Value Memory Networks for Knowledge Tracing. In Proceedings of the 26th International Conference on World Wide Web, WWW '17, Perth, Australia, 3–7 April 2017; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2017; pp. 765–774. https://doi.org/10.1145/3038912.3052580.
14. Shen, S.; Liu, Q.; Chen, E.; Wu, H.; Huang, Z.; Zhao, W.; Su, Y.; Ma, H.; Wang, S. Convolutional Knowledge Tracing: Modeling Individualization in Student Learning Process. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 25–30 July 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1857–1860. https://doi.org/10.1145/3397271.3401288.
15. Nakagawa, H.; Iwasawa, Y.; Matsuo, Y. Graph-based Knowledge Tracing: Modeling Student Proficiency Using Graph Neural Network. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence WI '19, Thessaloniki, Greece,

14–17 October 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 156–163. https://doi.org/10.1145/3350546.3352513.
16. Graves, A.; Schmidhuber, J. Framewise Phoneme Classification with Bidirectional LSTM Networks. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July 2005–4 August 2005; IEEE: Piscataway, NJ, USA, 2005, Volume 4, pp. 2047–2052. https://doi.org/10.1109/IJCNN.2005.1556215.
17. Kim, B.H.; Vizitei, E.; Ganapathi, V. GritNet: Student Performance Prediction with Deep Learning. *arXiv* **2018**, arXiv:1804.07405.
18. Carmona, C.; Millán, E.; Pérez-de-la Cruz, J.L.; Trella, M.; Conejo, R. Introducing Prerequisite Relations in a Multi-layered Bayesian Student Model. In Proceedings of the User Modeling 2005, Edinburgh, UK, 24–29 July 2005; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; pp. 347–356. https://doi.org/10.1007/11527886_46.
19. Drasgow, F.; Hulin, C.L. Item Response Theory. In *Handbook of Industrial and Organizational Psychology*; Consulting Psychologists Press: Palo Alto, CA, USA, 1990; Volume 1, pp. 577–636.
20. De la Torre, J. DINA Model and Parameter Estimation: A Didactic. *J. Educ. Behav. Stat.* **2009**, *34*, 115–130. https://doi.org/10.3102/1076998607309474.
21. Zhang, L.; Xiong, X.; Zhao, S.; Botelho, A.; Heffernan, N.T. Incorporating Rich Features into Deep Knowledge Tracing. In Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale, L@S '17, Cambridge, MA, USA, 20–21 April 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 169–172. https://doi.org/10.1145/3051457.3053976.
22. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language Modeling with Gated Convolutional Networks. In Proceedings of the 34th International Conference on Machine Learning, ICML'17, Sydney, Australia, 6–11 August 2017; JMLR: MA, USA 2017, pp. 933–941.
23. Gori, M.; Monfardini, G.; Scarselli, F. A New Model for Learning in Graph Domains. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; Volume 2, pp. 729–734. https://doi.org/10.1109/IJCNN.2005.1555942.
24. Abdelrahman, G.; Wang, Q. Knowledge Tracing with Sequential Key-Value Memory Networks. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Paris, France, 21–25 July 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 175–184. https://doi.org/10.1145/3331184.3331195.
25. Liu, Q.; Huang, Z.; Yin, Y.; Chen, E.; Xiong, H.; Su, Y.; Hu, G. EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 100–115. https://doi.org/10.1109/TKDE.2019.2924374.
26. Pandey, S.; Temporali.; Karypis, G. A Self-Attentive model for Knowledge Tracing. *arXiv* **2019**, arXiv:1907.06837. https://doi.org/10.48550/arXiv.1907.06837.
27. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. https://doi.org/10.1109/5.726791.
28. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Gang, W. Recent Advances in Convolutional Neural Networks. *Pattern Recognit.* **2018**, *77*, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013.
29. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
30. Zeng, C.; Zhu, D.; Wang, Z.; Wu, M.; Xiong, W.; Zhao, N. Spatial and Temporal Learning Representation for End-to-End Recording Device Identification. *EURASIP J. Adv. Signal Process.* **2021**, *2021*, 41. https://doi.org/10.1186/s13634-021-00763-1.
31. Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Chen, Y.; Yin, Y.; Huang, Z.; Wang, S. Neural Cognitive Diagnosis for Intelligent Education Systems. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 6153–6161. https://doi.org/10.1609/aaai.v34i04.6080.
32. Feng, M.; Heffernan, N.; Koedinger, K. Addressing the Assessment Challenge with An Online System that Tutors as It Assesses. *User Model. User-Adapt. Interact.* **2009**, *19*, 243–266. https://doi.org/10.1007/s11257-009-9063-7.
33. Xiong, X.; Zhao, S.; Van Inwegen, E.G.; Beck, J.E. Going Deeper with Deep Knowledge Tracing. In Proceedings of the 9th International Conference on Educational Data Mining, Raleigh, NC, USA, 29 June–2 July 2016; pp. 545–550.
34. Koedinger, K.; Cunningham, K.; Skogsholm, A.; Leber, B.; Stamper, J. A Data Repository for the EDM Community. *Handb. Educ. Data Min.* **2010**, *43*, 43–56. https://doi.org/10.1201/b10274-6.