Complex & Intelligent Systems (2022) 8:5143–5156
https://doi.org/10.1007/s40747-022-00731-8

ORIGINAL ARTICLE

Student achievement prediction using deep neural network from multi-source campus data

Xiaoyong Li[1,2] · Yong Zhang[1] · Huimin Cheng[2] · Mengran Li[1] · Baocai Yin[1]

Received: 1 November 2021 / Accepted: 27 March 2022 / Published online: 6 May 2022
© The Author(s) 2022

Abstract

Finding students at high risk of poor academic performance as early as possible plays an important role in improving education quality. To do so, most existing studies have used the traditional machine learning algorithms to predict students' achievement based on their behavior data, from which behavior features are extracted manually thanks to expert experience and knowledge. However, owing to an increase in the varieties and overall volume of behavioral data, it has become more and more challenging to identify high-quality handcrafted features. In this paper, we propose an end-to-end deep learning model that automatically extracts features from students' multi-source heterogeneous behavior data to predict academic performance. The key innovation of this model is that it uses long short-term memory networks to capture inherent time-series features for each type of behavior, and it takes two-dimensional convolutional networks to extract correlation features among different behaviors. We conducted experiments with four types of daily behavior data from students of the university in Beijing. The experimental results demonstrate that the proposed deep model method outperforms several machine learning algorithms.

Keywords Academic performance prediction · Time-series features · Correlation features · LSTM · 2DCNN

Introduction

Students' performance is a key indicator in measuring the quality of academic education and is also closely related to students' mental health. Related studies have shown that students with poor academic performance are prone to anxiety and depression [1], and their risk of suicide is much higher than that of students with excellent performance [2,3]. Achievement prediction aims to identify students with high academic risk in advance, which reminds administrators, teachers, and students themselves of taking timely targeted intervention actions to avoid poor performance, such as failing courses, dropping out, staying out, and so on. Therefore, student achievement prediction has been receiving extensive attention and research.

Factors affecting academic achievement are complex and diverse. To explore related factors, researchers in various fields have done lots of work. For example, literature [4] explored the relationship between cognitive abilities and academic performance. Literatures [5,6] expounded the correlation between "Big Five traits" (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism) and academic achievement. Studies [7–9] show that good sleep habits are helpful to improve academic performance. Literatures [10–13] conclude that moderate physical activity can facilitate the improvement of academic achievement. Literature [14] shows that binge eating and purging behaviors lead to relatively poor academic performance, and literature [15] shows that the influence in girls is higher than in boys. Literatures [16–19] show that bad habits of using social software and electronic devices affect academic achievement. These studies demonstrate the strong correlation between various behavior-related factors and academic performance, and provide guidance and suggestions

✉ Yong Zhang
zhangyong2010@bjut.edu.cn

Xiaoyong Li
lixiaoyong@bjut.edu.cn

Huimin Cheng
chenghm@bjut.edu.cn

[1] Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing Artificial Intelligence Institute, Faculty of Information Technology, Beijing University of Technology, 100 Pingleyuan, Chaoyang District, Beijing 100124, China

[2] Information Technology Support Center, Beijing University of Technology, 100 Pingleyuan, Chaoyang District, Beijing 100124, China

for managers and teachers to improve students' academic achievement. However, most data used in these studies were collected from questionnaires or self-reports, which usually suffer from small sample size and social desirability bias.

With the rapid development of digital campus in recent years, many information systems are deployed on campus, such as learning management system (LMS), smart card system, gateway system, access control system, and so on, which truly record various behavior data of students in the learning and living processes. Compared with data obtained from questionnaires, these data objectively reflect students' behavior patterns and cover a large number of samples, which provides an great opportunity for performance prediction. Because there is an obvious correlation between learning behavior and academic achievement, many studies [20–22] create predictive models by analyzing students' learning behavior patterns from LMS log files, such as video watching, homework submitting, and BBS discussion. Unfortunately, these learning behavior data are limited to specific courses, so the models trained on a specific course cannot be well generalized to other courses. Furthermore, many courses in the traditional face-to-face education are not taught through LMS, in which there are little available learning data to predict achievement.

Daily living behavior data are another important data source that describe students' campus behavior patterns, they include dining behavior, shopping behavior, library entry behavior, web page browsing behavior, and so on. Be different from learning behavior on LMS, living behavior can be recorded for every student living on campus, which provide a much broader and available data source for performance prediction. Based on them, related studies [23–28] artificially extracted features from raw behavioral data relying on expert knowledge, such as breakfast frequency, Internet time, orderliness, diligence, sleep pattern, and so on, and then constructed prediction models using machine learning algorithms. However, the following challenges are encountered when manually extracting features from massive multi-source living data: (1) the quality and number of features are directly influenced by expert knowledge, and it is difficult to extract high-quality features by understanding the overall distribution of massive data; (2) although some features such as orderliness express the regularity of behavior, they still cannot fully represent the temporal characteristics of time-series behavior data; (3) the correlation between multi-source behavior data need to be further mined.

To address the aforementioned challenges, we put forward a novel academic performance prediction method based on deep neural network (DNN), in which behavioral features are automatically learned instead of being extracted manually, Long-Short-Term Memory (LSTM) networks are applied to model the temporal characteristics of behavior data, and two-dimensional convolutional neural network (2DCNN) is used to capture the correlation among different behaviors. The general framework of our method is shown in Fig. 1. Each type of raw behavior data is fed into LSTM model to obtain their time-series features separately. Then, these features are converted into a feature tensor; based on it, 2DCNN is applied to capture correlation features among different behavior. Finally, fully connected layers are used to output the academic performance level by concatenating the time-series features, correlation features, and students' demographic information. Specifically, for behavior data in log format such as web page browsing behavior, we use embedding layer to obtain dense vectors of nominal attributes, and use one-dimensional convolutional neural network (1DCNN) to reduce their sequence length.

The main contributions of the paper are as follows:

1. An end-to-end deep neural network is proposed for academic performance prediction based on students' multi-source daily life behavior data, which can automatically extract features without relying on expert experience.
2. The time-series features of each type of behavior data are efficiently extracted using long-short-term memory network, embedding layer and one-dimensional convolution networks.
3. The correlation features among various types of behaviors are obtained using two-dimensional convolutions.
4. The experiments are conducted on a large-scale real data set, and their results show that our proposed method outperforms the traditional machine learning methods.

In this paper, "Related work" presents related work in achievement prediction. "Data set" introduces the data sets used in this study. "Deep neural network for achievement prediction" describes the proposed DNN method and its detailed configuration. "Experiments and results" shows the experimental design and results. The conclusions and future work are presented in "Conclusion and future work".

Related work

Many works have been conducted on student achievement prediction, elaborated from the perspectives of objectives, data, and methods in this section.

Machine learning approaches

The studies based on machine learning algorithms usually define the problem of academic achievement prediction as a classification task or a regression task, aimed at predicting students' achievement level or ranking. Mingyu et al. extracted features from students' behavior data and demographic information. The positive correlation with greater

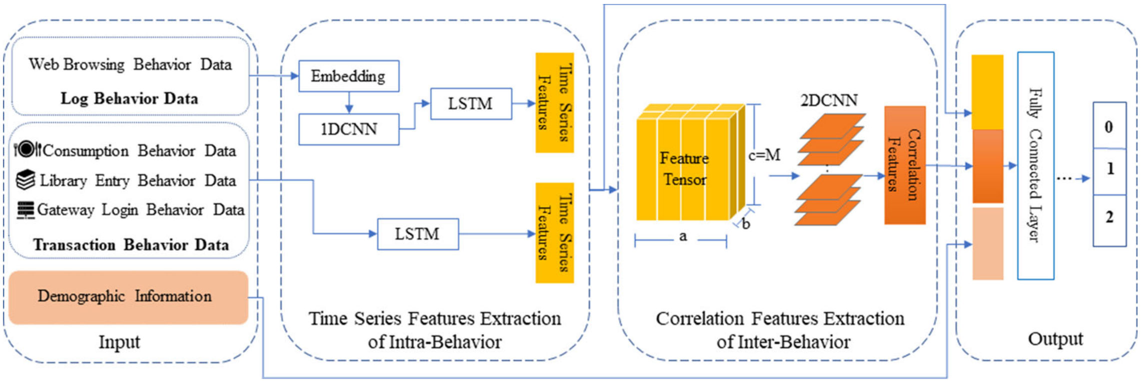

Web Browsing Behavior Data
Log Behavior Data
Consumption Behavior Data
Library Entry Behavior Data
Gateway Login Behavior Data
Transaction Behavior Data
Demographic Information
Input
Embedding
1DCNN
LSTM
Time Series Features
LSTM
Time Series Features
Time Series Features Extraction of Intra-Behavior
Feature Tensor
c=M
a
b
2DCNN
Correlation Features
Correlation Features Extraction of Inter-Behavior
Fully Connected Layer
0
1
2
Output

Fig. 1 Framework of deep network model for achievement prediction

weight between regular living habits and academic performance is verified, and the weighted grade point average (GPA) of students is predicted using the improved boosting algorithm Catboost [23]. Additionally, Cao et al. extracted two features (orderliness and diligence) from students' behavior data of eating, showering, entering the library, and fetching water on campus. The correlation between the two behavior features and academic performance is verified using Spearman's rank correlation coefficient, and students' academic ranking is predicted using a pair-wise learning to rank algorithm RankNet [24]. Yao et al., in an upgraded version of a study by Cao et al. [24], put forward sleep pattern features in addition to orderliness and diligence. Based on the three features, they analyzed the correlation of academic performance of the students with similar behaviors using social influence theory and built a multitask academic performance prediction framework using learning to rank algorithm [25]. Zhou et al. extracted the frequency and duration of students' visits to different types of web pages from Internet access logs. Frequency and duration were used as features to verify the close correlation between internet access and academic performance, and then six algorithms, namely Naïve Bayes, decision tree (DT), logistic regression (LR), support vector machine, neural network, and K-nearest neighbor, were used to predict students with high risk in academic performance [26]. Zhang et al. modeled the transformation mode of students' consumption behavior using the Hidden Markov Model and extracted the features expressing behavioral level, behavioral trend, behavioral regularity, and behavioral diversity. Based on this information, a regularized multitask learning model was built to simultaneously predict students' score for each course in the current semester [27]. Ghosh et al. used context-aware traj-graph to model the mobility patterns of students based on their GPS trajectories on campus, and uncovered the correlation of signature mobility patterns with the academic performance of the students [28].

These studies predict students' academic performance using machine learning algorithms. However, their performances are mainly reliant on the quality of features extracted manually based on expert knowledge, bringing a great challenge when faced with diverse data types and massive data.

Deep neural networks

DNN has the powerful nonlinear expression abilities by stacking multiple hidden layers, and its goal is to have the network automatically learn features for classification or regression tasks. In recent years, DNN has achieved great success in various fields, such as image recognition, speech recognition, text translation, and so on. Inspired by these great breakthroughs, the initial attempts are emerging to predict academic performance using DNN. For example, Yang et al. manually extracted the behavioral features from students' online learning behaviors, including the behaviors of visiting curriculum resources and participating in curriculum discussion, and then combined these features into an image tensor. Based on this tensor, convolution neural network was used to predict whether students could pass the curriculum exam [29]. Pu et al. built an undirected graph to express students' similarity by analyzing the correlation of their course achievements in the previous semesters and then used a graph convolution network to predict students' course achievements in the current semester [30]. Botelho et al. used LSTM to model students' behavior of doing homework and combined decision tree (DT) and logistic regression (LR) algorithms to discover "stop out" and "wheel spinning" behavior [31].

These studies show that DNN can better model students' behavior and obtains good prediction performance compared with the traditional machine learning algorithms. However, as far as we know, there are few studies that predict academic performance based on daily living behavior data using DNN.

Data set

At many Asian universities, the majority of undergraduates live on campus and take several courses each semester, receiving a score or grade for every course after completing them. Various types of behavior data are produced on campus, such as consumption behavior in canteen and web page browsing behavior. In the following subsection, we introduce the data set, preprocessing, and grading of academic achievements.

Data description

The data set in this study came from the university in Beijing in the spring semester over a period of 145 days. Four types of campus behaviors of 9000 students were collected from different databases using extract, transform, load (ETL) tools. They include consumption behavior, library entry behavior, gateway login behavior, and web page browsing behavior. The data samples of each kind of behavior are, respectively, shown in Tables 1, 2, 3, and 4, in which attributes and value domains can be clearly observed. For example, consumption behavior includes four attributes, data, time, location, and consumption amount; library entry behavior contains three attributes, which are data, time, and location; gateway login behavior has five attributes including data, time, location, access duration, and network traffic used; and web browsing behavior has four attributes, date, time, uniform resource locator (URL) domain, and location. Among those behaviors, consumption behavior was further refined into breakfast behavior, lunch behavior, dinner behavior, and shopping behavior. These behavioral data truly record students' activities on campus from different aspects. In addition, students' demographic information of gender, school, major, grade, and graduation middle school, as well as the course achievement information, were also collected. To protect students' privacy, all students' IDs were irreversibly anonymized in the collection process.

Because the goal of this paper is to predict academic performance based on students' campus behavior data, the student samples with fewer behavior records were filtered out by setting conditions. The specific conditions were that the number of students' breakfast behavior records, lunch behavior records, dinner behavior records, and gateway login records in a semester should not be fewer than 20 respectively, and the number of web page browsing behavior records should not be fewer than 1000. The filtered dataset contained 8228 student samples.

Data preprocessing

In original behavior data, behavior date in "yyyy-mm-dd" format could not clearly express the stage of the semester when the behavior occurred, and the values of the attributes of date and time could not be directly used as the input of the model. Therefore, it is necessary to preprocess the date and time. For the date attribute, its value was converted into an integer value starting from 1 by referring to the university calendar, that is, 1 represents the date corresponding to the first day of the calendar, and so on. Regarding the time attribute, in the first step, the 24 h in a day were evenly divided into K intervals according to the specified time interval τ, and the divided intervals were numbered as $1, 2, \ldots, K$. Then, the time of the recorded behavior was input as the number of the corresponding interval. In this study, τ was set to 4 h for web page browsing behavior, because students can visit the same web page repeatedly in a short time. A smaller τ value could have led to redundant behavior content. τ was set to 15 min for the other three types of behaviors.

After transforming the date and time of behavior, redundant behavior data may occur; for example, two records of consumption behavior have the same date, time, and place values; two records of web page browsing behavior are identical. This phenomenon not only wastes computing resources but also does not facilitate the improvement of model performance. Therefore, it is necessary to remove these redundant data. For consumption behavior data, multiple records with the same date, time, and place are merged into a new record, of which the consumption amount is equal to the sum of the consumption amounts of the merged records. The same merge operation is also performed on gateway login behavior data, in which the network traffic and online duration of a new record are equal to the sum of that of the merged records. For the library entry behavior and web page browsing behavior, the duplication eliminating operation was carried out.

Besides behavior data, the attribute of graduation middle school was also preprocessed, because students may graduate from thousands of high schools, which makes one-hot encoding of the attribute produce sparse vectors. To solve this problem, this attribute is transformed into three related attributes, namely, the administrative level (provincial, municipal, and county level) of the city where schools are located; the nature of schools (public and private); and the teaching level of schools (national key, provincial key, municipal key, county key, and ordinary schools), and then, one-hot encoding of these three attributes gets a ten-dimensional vector.

Grading of academic achievements

Students' academic performance is usually measured by GPAs of continuous numerical values. In this paper, predicting academic performance is defined as a classification task, that is, predicting whether the performance of a student is excellent, good, or poor. Thus, we must divide GPA into discrete academic levels. The grading process is as follows:

Table 1 Samples of consumption behavior data

Student ID	Date	Time	Location	Amount (CNY)	Date index	Time index
17****01	2019-02-21	16:51:29	Canteen no. 1	6.0	4	68
17****01	2019-02-22	08:31:52	Canteen no. 3	3.5	5	35

Table 2 Samples of behavior data of entering library

Student ID	Date	Time	Location	Date index	Time index
17****01	2018-05-27	08:15:29	Library	91	34
17****01	2018-06-02	09:31:01	Library	97	39

Table 3 Samples of behavior data of logging into gateway

Student ID	Date	Time	Location	Duration (min)	Network traffic (MB)	Date index	Time index
17****01	2018-2-26	12:35:46	Library	54	42.95	1	51
17****01	2018-2-27	15:30:57	Dormitory	47	85.92	2	63

all students in the dataset are sorted by GPA in descending order, then the top $r\%$ of students' achievements were defined as excellent, the bottom $r\%$ of students' achievements were defined as poor, and other students' achievements were good. However, there is no unified standard for reference to set threshold r; authors of relevant studies have usually artificially set thresholds, different thresholds could generate different grading results that leads to limited comparability of model performance. To observe the performance of the model as comprehensively as possible, we set four thresholds—5%, 10%, 15%, and 20%—to grade academic achievement. The grading results are shown in Table 5, which shows the GPA interval and the number of students in each grade under different thresholds.

Deep neural network for achievement prediction

As shown in Fig. 1, the proposed prediction method consists of four phases; (1) Input: various types of time series behavior data produced by student on campus and students' static demographic information are the input of the DNN model without extracting features manually; (2) Time-series features extraction of intra-behavior: we mainly use LSTM to separately model each type of behavior to learn their time-series features; (3) Correlation features of inter-behavior: 2DCNN is applied to a tensor converted from time-series features of all behaviors to capture the correlation features among different behaviors; (4) Output: Students' behavioral features and demographic information are concatenated as the input of fully connected layers to output academic performance level.

Input of the model

The model input includes two kinds of data, one is the behavior data produced by a student on campus, including breakfast behavior data, lunch behavior data, dinner behavior data, shopping behavior data, library entry behavior data, gateway login behavior data, and web page browsing behavior data; the other is student's demographic information, such as gender, major, graduation middle school, and so on. The attributes of these data are described in "Data description".

All types of behavior data are classical time-series data, in which each record contains a timestamp, but distinct behaviors have different attributes, and the length of the same behavior varies from student to student. $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{iN})$ denotes the N types of multi-source behavior data of student i, matrix $X_{ij} = [x_{ij}^1, \ldots, x_{ij}^t, \ldots, x_{ij}^{T_{ij}}]$ expresses the jth behavior data of student i, $x_{ij}^t (1 \le t \le T_{ij})$ is a vector representing one event record information at time t such as one consumption record, one gateway login record, in which T_{ij} is the length of the jth behavior of the ith student. After simple preprocessing such date trans-

Table 4 Samples of behavior data of browsing web

Student ID	Date	Time	URL domain	Location	Date index	Time index
17****01	2018-02-26	02:12:53	http://www.youku.com	Dormitory	1	1
17****01	2018-02-26	23:25:12	http://r6.mo.baidu.com	classroom	1	6

Table 5 Grading results of academic achievements

Academic	5%		10%		15%		20%	
	GPA	Students	GPA	Students	GPA	Students	GPA	Students
Excellent	[4, 4]	426	[3.92, 4]	828	[3.9, 4]	1249	[3.8, 4]	1682
Good	(2.5, 4)	7287	(2.8, 3.92)	6572	(3.0, 3.9)	5672	(3.1, 3.8)	4855
Poor	[0, 2.5]	515	[0, 2.8]	828	[0, 3.0]	1307	[0, 3.1]	1691

formation, time transformation, and redundant deletion as stated in "Data preprocessing" and normalization, they can be directly used as inputs to the model.

Time-series features extraction of intra-behavior

LSTM is a classic type of recurrent network that is specialized for processing a sequence of values. It can effectively reduce the difficulty of learning long-term dependencies to scale on much longer sequences. To automatically learn features from students' behavior which are presented in a time-series manner, LSTM is used to extract features of campus behavior data.

Campus behavior data can be divided into transaction behavior data and log behavior data according to their generation mechanism. The former refers to behavior in which one activity event only produces one record. For example, consumption behavior data, library entry behavior data, and gateway login behavior data in the dataset fall into this category; they typically contain hundreds of records. These behavior data are directly input into LSTM to extract time series features after one-hot encoding or normalization of attributes. The latter, log behavior data, refers to the behavior in which one event produces hundreds or even thousands of records, such as web page browsing behavior. For log-type web page browsing behavior, there are two challenges when modeling using LSTM. First, the number of (URL) domains is huge, making the vector obtained via one-hot encoding of the URL domain attribute extremely sparse. Second, extremely long sequence leads to high resource consumption and slow convergence when modeling directly with LSTM. To solve the two problems, an embedding layer is applied to learn the dense vector of URL domain, and an one-dimensional convolutional network is utilized to reduce the length of the sequence before modeling it using LSTM.

URL domain embedding representation

To solve the vector sparsity problem caused by the huge number of domain names, Zhou et al. [26] labeled the URL domain as classes of learning, games, music, movies, and so on, followed by one-hot encoding of the domain classes. This method not only solves the problem of data sparsity but also facilitates the interpretability of academic performance prediction. However, it has the following disadvantages: (1) There are no general domain class database or labeling criteria, resulting in labeling the domain class manually, so this method is time-consuming and labor-intensive. (2) It is not conducive to protecting students' privacy, because the class of domain reveals the browsing content to a certain extent, especially when the domain is marked with fine-grained categories. These two disadvantages limit the scope of application of this method.

Inspired by the concept of word vector in natural language processing, words are mapped to a continuous low-dimensional vector space in which words with similar semantics are located in near positions to each other. This idea is introduced in this paper to find a dense vector for URL domain, attempting to overcome the two disadvantages of the labeling method stated above. Word2vec and embedding layer are two classical word vector learning methods. Among them, word2vec is an unsupervised learning method and uses the context of words to learn word vector; on the contrary, the embedding layer in deep neural network iteratively updates the word vector based on the labels of a task until the end of model training. Considering that the academic prediction task is a classification task that has label information, the embedding layer is introduced to learn URL domain vector. The specific process is as follows: (1) count the access frequency of all URL domains in the dataset; (2) construct a domain name index table, in which domains are sorted in the descending order according to access frequency and are assigned indexes in turn from 1; (3) filter high-frequency domain names from the domain name index table; (4) convert the domain names in the web browsing behavior sequence into the index value; (5) configure the embedding layer in the deep neural model.

Behavior sequence length reduction

Although LSTM can capture much longer information dependence than simple recurrent neural network, it still cannot efficiently model the extremely long sequence of web page browsing behaviors. To reduce the length of the behavior sequence, researchers usually use down-sampling technology to delete some records, possibly losing important information. In this study, one-dimensional convolutions are performed on the behavior sequence to extract the features

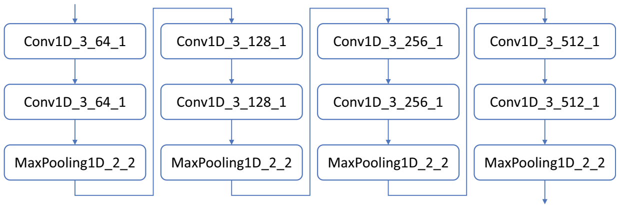


Fig. 2 Reducing behavior sequence length using one-dimensional convolution

in local time, and then, pooling layers are used to filter out redundant features. This method can efficiently reduce the sequence length while preserving important feature information within behavior data.

Inspired by VGG network [32], the sub-model for reducing sequence length is shown in Fig. 2, in which two consecutive convolution layers are followed by a pooling layer, "Conv1D_3_k_1" represents a convolution layer composed of k one-dimensional convolutions with a kernel size of 3 and a step size of 1, and the values of k are 64, 128, 256, and 512 in turn; "MaxPooling1D_2_2" indicates a one-dimensional maximum pooling layer with kernel size of 2 and step size of 2. The purpose of setting kernel size to 3 is to enhance the nonlinear expression ability of the network by increasing the depth under the condition of having the same receptive field as the large convolution kernel. Through this sub-model, the sequence length is greatly reduced from L to $(L-60)/16$.

Correlation features extraction of inter-behavior

Because multi-source behavior data are from the same student, there must be a correlation between different behaviors. It is a conventional idea to combine all types of behavior data into a unified data format as input to a deep neural network. However, as opposed to time-series data produced by sensors in industry fields, students' behaviors are actively triggered by the students themselves; behavior data are characterized by inconsistent sampling frequency. For example, Student A produces three records of having meals per day, while the records of page browsing behavior may number in the thousands. Thus, it is inefficient to convert these multi-source behavioral data into one tensor for extracting the correlation features between different Behaviors, which could lead a very sparse tensor under fine time granularity or lost lots of information under coarse granularity.

To capture the correlation features, we propose a tensor scheme to transform the time-series feature vector of each behavior into a three-dimensional tensor and then employ 2DCNN to analyze the relationship among local adjacent behaviors. 2DCNN is usually used to extract image features, in which an image is expressed by a tensor(w, h, c), where w, h, and c, respectively, represent the width, height, and number of channels of an image. As shown in Fig. 1, the time-series feature vectors of N types of behaviors are transformed into a 3-D tensor, where $w*h = N, c = M$, M indicates the dimension of time-series feature vector extracted in "Time-series features extraction of intra-behavior". Based on the tensor, 2DCNN is performed to obtain the correlation features.

Output of the model

In this paper, academic performance prediction is defined as a classification task, the output of the model should be $y \in \{0\text{: poor}, 1\text{: good}, 2\text{: excellent}\}$, and the detailed grading process is described in "Grading of academic achievements". Fully connected layers are applied to output the performance level, in which the time-series features of intra-behavior, the correlation features of inter-behavior, and student's demographic information are concatenated as its input. In addition, the dropout layer is used before each fully connected layer to prevent overfitting, weighted cross-entropy function described in "Weighted loss function for solving class imbalance problem" is used as the loss function, and adaptive moment estimation (Adam) is used as optimizer.

Detailed model configuration

The detailed configuration of the proposed model is shown in Table 6. The first layer represents the input composed of N types of behavior sequence, where T_i and F_i, respectively, represent the sequence length and feature number of the ith behavior data. It should be noted that the domain name vector learning and one-dimensional convolution operation need to be performed on the web page browsing behavior before it is input into the LSTM. The second layer performs LSTM modeling on each kind of behavior sequence and outputs a vector containing 32 features.

In the third, fourth, and fifth layers, the concatenation layer, reshape layer, and permute layer are adopted to convert the feature vectors of N kinds of behavior data into a tensor of $(w, h, 32)$. Owing to there being a few types of behaviors, more two-dimensional convolutional layers could lead to overfitting; therefore, only two convolutional layers are set up in the sixth and seventh layers, with the kernel numbers of 32 and 64, respectively, the kernel sizes of (2, 2), and the step sizes of (1, 1), and the sixth layer adopts filling mode to keep the dimension of tensor unchanged.

The tenth layer takes the basic information of students as input. In the eleventh layer, students' basic information, behavioral correlation features, and time-series features are concatenated as the input of the fully connected layer, where L_8, L_9, and L_{10} represent the lengths of the output vectors in the eighth, ninth, and tenth layers, respectively. The fully connected layer contains multiple layers, and the output units are 2048, 1024, and 512 respectively. A dropout layer is set in front of each fully connected layer to avoid overfitting, and the dropout rate is set to 0.5. For simplicity, not all fully connected layers and dropout layers are listed in the table, but are marked with $*$ after the 12th and 13th layers for illustration. The 14th layer is the output layer, 3 represents the levels of academic performance, and the activation function is softmax.

Experiments and results

In this section, we describe how to train the deep neural network and evaluate its performance.

Experimental design

Three key problems encountered during model training, including class imbalance problem, overfitting, and evaluating the deep model, are solved.

Weighted loss function for solving class imbalance problem

By observing the dataset in Table 5, a class imbalance problem is seen in the achievement prediction task. The solutions to this problem are generally divided into three categories: under-sampling technology, over-sampling technology, and weighted loss function. The first type of method is to randomly delete some student samples with good scores to make the number of students in each class similar. Based on the limited number of student samples in the dataset of this paper, under-sampling could further reduce samples, which is not feasible for the training of the deep neural network. The second type of method is to produce new student samples with poor scores and excellent scores to balance the three classes of students. Synthetic minority over-sampling technique (SMOTE) [33] and Borderline-SMOTE [34] are two classic over-sampling methods based on Euclidean distance, but they are computationally inefficient when synthesizing high-dimensional student samples expressed by various behaviors. The third type does not delete or produce samples; it only gives higher weight to students with poor scores and excellent scores when calculating the loss function, giving these samples greater influence on the loss function. Compared with over-sampling technologies, the weighted loss function requires fewer computing resources, so we used it to solve the class imbalance problem in this paper. The weighted cross entropy loss function is shown in Eq. (1), where w_i indicates the weight of class i, N is the number of total student samples, N_i is the number of student samples belonging to class i, M is the number of classes, y_i^k is the true score level of the kth student samples belonging to class i, and p_i^k is the predicted score level probability

$$\begin{aligned} w_i &= \frac{N}{M * N_i} \\ \text{loss} &= \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{M} w_i y_i^k \log(p_i^k). \end{aligned} \tag{1}$$

Data enhancement for preventing overfitting problem

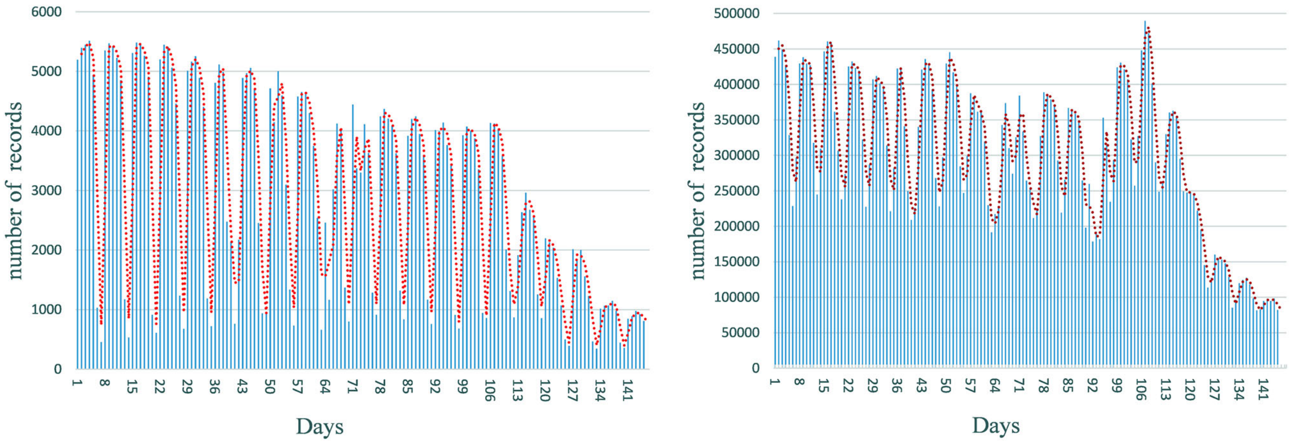
Solutions to prevent overfitting usually include data enhancement, early stopping, l_1 and l_2 regularization, and dropout. In this paper, the large gap between the number of model parameters and the number of student samples made the proposed model prone to overfitting. Owing to the limitation of experimental conditions, no more student samples could be collected. By analyzing the characteristics of students' behaviors, we found that students' campus behaviors have obvious periodicity in weeks, with a certain volatility at the same time, as shown in Fig. 3. Therefore, it is feasible to predict academic performance based on behavioral data within a time period. In this paper, the behavior data of 145 days were segmented into ten pieces, each containing behavior data of 2 weeks except for the last one. Each piece was taken as a new student sample, and its label was consistent with the one of the original student. This data enhancement method not only increased the sample size tenfold, but also enriched the sample distribution owing to the small fluctuation of behavior data in different time periods. In addition, early stopping and dropout were also used in the process of model training. When the loss value of the validation set was less than 0.001 for 30 consecutive rounds, the training was stopped, and the parameter setting of dropout was described in "Detailed model configuration".

Evaluation metrics

For classification tasks with unbalanced classes, precision rate and recall rate should be taken as more important evaluation metrics than accuracy. In the three-way classification achievement prediction task, $i = 0, 1,$ and 2 are used to indicate the poor, good, and excellent performance classes, respectively. P_i represents the precision rate of class i, R_i indicates the recall rate of class i captured by a model, F_β^i is a trade-off metric between P_i and R_i, and the relative importance of recall and precision in F_β^i metric is adjusted by setting β value. For evaluating the overall performance of the model, macro precision rate P, macro recall rate R, and macro F_β are calculated. The calculation of these metrics is shown in Eq. (2), where TP_i, FP_i, and FN_i represent the number of true-positive samples, false-positive samples, and false-negative samples of class i in the model, respectively.

Table 6 Detailed configuration of the proposed deep model

No.	Layer type	Kernel number	Kernel size	Step size	Output dimension	Activation	Conn. layer
1	Input				$(T_i, F_i) * N$		
2	LSTM				$(32, 1) * N$		1
3	Concatenate				$(32, N)$		2
4	Reshape				$(32, w, h)$		3
5	Permute				$(w, h, 32)$		4
6	Conv2D	32	(2, 2)	(1, 1)	$(w, h, 32)$	Relu	5
7	Conv2D	64	(2, 2)	(1, 1)	$(w - 1, h - 1, 64)$	Relu	6
8	Flatten				$((w - 1) * (h - 1) * 64)$		7
9	Flatten				$(32 * N)$		3
10	Input				(S)		
11	Concatenate				$(L_8 + L_9 + L_{10})$		8, 9, 10
12	Dropout(0.5)*				$(L_8 + L_9 + L_{10})$		11
13	Dense*				(2048)	Relu	12
14	Dense				(3)	Softmax	



(a) Number of having breakfast per day

(b) number of browsing webpages per day

Fig. 3 Statistics of behavior showing its periodicity and volatility

In our experiment, the value of β was set to 1

$$
\begin{aligned}
P_i &= \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \\
R_i &= \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \\
F_\beta^i &= \frac{(1 + \beta^2) * P_i * R_i}{(\beta^2 * P_i) + R_i} \\
P &= \frac{1}{3} \sum_{i=1}^{3} P_i \\
R &= \frac{1}{3} \sum_{i=1}^{3} R_i \\
F_\beta &= \frac{(1 + \beta^2) * P * R}{(\beta^2 * P) + R}.
\end{aligned}
\tag{2}
$$

Experimental results

To verify the performance of the proposed deep model, we compared it with the traditional machine learning algorithms and showed the advantages of the model based on multi-source heterogeneous behavior data.

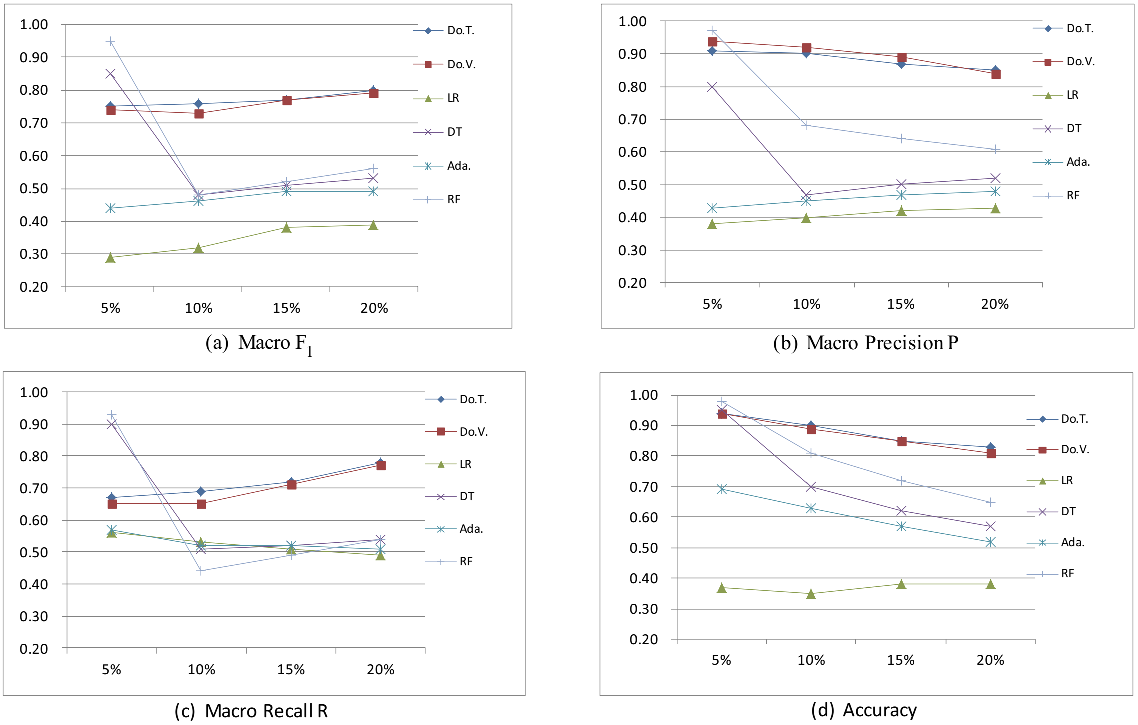
Performance comparison of related methods

LR, DT, AdaBoost (Ada.), and Random Forest (RF) are common machine learning algorithms for predicting academic performance, especially Ada. and RF, which improve the performance of the model by assembling the classification results of multiple base learners. In this paper, we manually extracted features for every kind of behavior using the method introduced in the literature [35]. These features

Table 7 Achievement prediction performance of different methods

Method	5%				10%				15%				20%			
	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1
LR	0.37	0.38	0.56	0.29	0.35	0.40	0.53	0.32	0.38	0.42	0.51	0.38	0.38	0.43	0.49	0.39
DT	0.95	0.80	0.90	0.85	0.7	0.47	0.51	0.48	0.62	0.50	0.52	0.51	0.57	0.52	0.54	0.53
Ada.	0.69	0.43	0.57	0.44	0.63	0.45	0.52	0.46	0.57	0.47	0.52	0.49	0.52	0.48	0.51	0.49
RF	**0.98**	**0.97**	**0.93**	**0.95**	0.81	0.68	0.44	0.48	0.72	0.64	0.49	0.52	0.65	0.61	0.54	0.56
Do.T	0.94	0.91	0.67	0.75	**0.9**	0.90	**0.69**	**0.76**	**0.85**	0.87	**0.72**	**0.77**	**0.83**	**0.85**	**0.78**	**0.80**
Do.V	0.94	0.94	0.65	0.74	0.89	**0.92**	0.65	0.73	**0.85**	**0.89**	0.71	**0.77**	0.81	0.84	0.77	0.79

The bold indicate the highest metric values



(a) Macro F_1

(b) Macro Precision P

(c) Macro Recall R

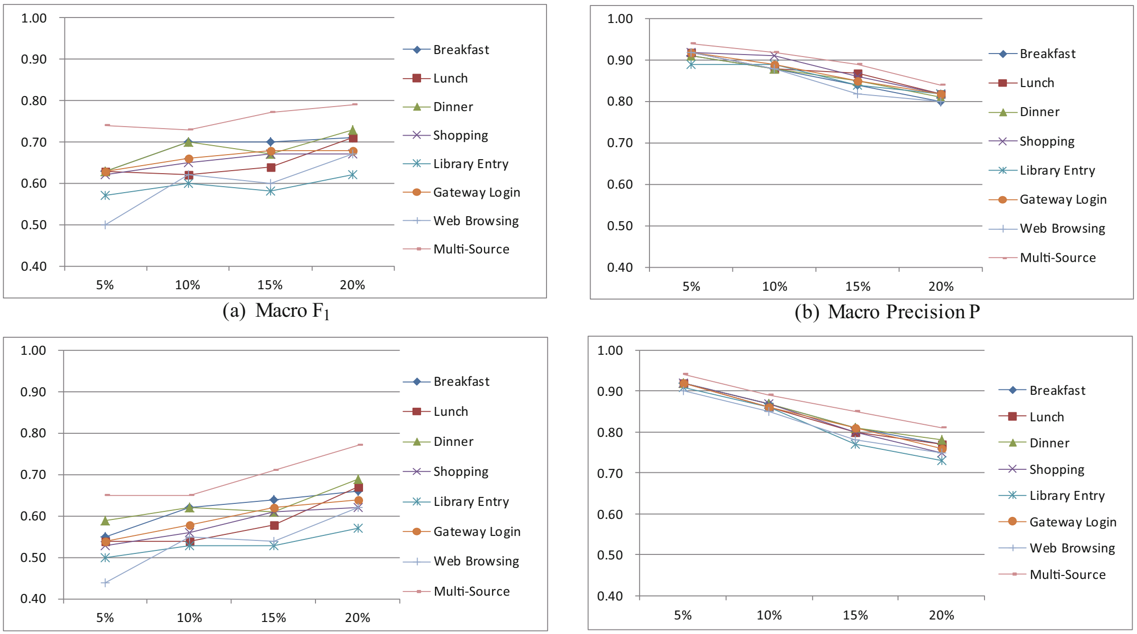
(d) Accuracy

Fig. 4 Performance comparison of achievement prediction-related methods

Table 8 Achievement prediction performance based on different behavior data

Behavior	5%				10%				15%				20%			
	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1
Breakfast	0.92	0.91	0.55	0.63	0.87	0.88	0.62	0.70	0.81	0.84	0.64	0.70	0.77	0.80	0.66	0.71
Lunch	0.92	0.92	0.54	0.63	0.86	0.88	0.54	0.62	0.80	0.87	0.58	0.64	0.77	0.82	0.67	0.71
Dinner	0.92	0.91	0.59	0.63	0.87	0.88	0.62	0.70	0.81	0.85	0.61	0.67	0.78	0.81	0.69	0.73
Shopping	0.92	0.92	0.53	0.62	0.87	0.91	0.56	0.65	0.80	0.86	0.61	0.67	0.75	0.82	0.62	0.67
Library	0.91	0.89	0.50	0.57	0.86	0.89	0.53	0.60	0.77	0.84	0.53	0.58	0.73	0.82	0.57	0.62
Gateway	0.92	0.92	0.54	0.63	0.86	0.89	0.58	0.66	0.81	0.85	0.62	0.68	0.76	0.82	0.64	0.68
Web	0.90	0.92	0.44	0.50	0.85	0.88	0.55	0.62	0.78	0.82	0.54	0.60	0.75	0.80	0.62	0.67
Multi-Source	**0.94**	**0.94**	**0.65**	**0.74**	**0.89**	**0.92**	**0.65**	**0.73**	**0.85**	**0.89**	**0.71**	**0.77**	**0.81**	**0.84**	**0.77**	**0.79**

The bold indicate the highest metric values



(a) Macro F_1

(b) Macro Precision P

(c) Macro Recall R

(d) Accuracy

Fig. 5 Performance comparison based on different behavior

express the concentration trend and dispersion degree of behavior distribution. Please note that, for the URL domain name attribute of web page browsing behavior, the number of times students visited various kinds of domain names every day was counted, and then using the method of literature [35] to extract features expressing the regularity, concentration, and dispersion characteristics of students visiting different domain names were described.

In addition to the aforementioned methods, we also compared the influence of different domain name processing methods in deep model on performance, namely, Domain Type (Do.T) labeling method and Domain Vector (Do.V) method. The difference between them is that the former labels the type of domain name of web page browsing behavior and then performs one-hot encoding, whereas the latter uses the embedding layer of the neural network to learn the dense vector of the domain name. In the Do.T method, domain names are marked into 20 categories; in the Do.V method, the top eight high-frequency domain names are filtered, and the size of domain name dictionary is set to 20,000. The performance of different methods is shown in Table 7, and to visually compare, Fig. 4 also shows the results using a line chart, where the horizontal axis represents the grading labels of score grades.

Figure 4a shows the macro F_1 values of different methods. It is found that the F_1 values of the two deep models with different domain name processing methods are greatly similar and steadily increase from 0.75 for 5% labels to 0.80 for 20% labels. Their F_1 values are much higher than that of the traditional machine learning algorithm, albeit slightly lower than the RF and DT methods on 5% labels.

Figure 4b shows the macro precision of different methods. The precision of the two deep models in this paper is obviously better than that of other algorithms. Although their precision decreases slowly with the increase of label ratio, they both remain above 0.84. The performance of RF algorithm on 5% labels is slightly better than that of the deep models, but drops sharply to 0.68 on 10% labels and continues to drop on 15% labels and 20% labels.

Figure 4c shows the macro recall of different methods. The values of the two deep models increase with the uptick of label proportion and reach 0.78 and 0.77, respectively, on 20% labels. Although the RF and DT algorithms outperform the two deep models on the 5% label set, their values drop rapidly and are lower than those of the deep models when the score label is equal to or greater than 10%.

Figure 4d shows the prediction accuracy of the methods. The accuracy of the two deep models are better than other methods in all cases albeit slightly lower than those of RF and DT algorithm on 5% label; the accuracy values under the four score labels are all higher than 0.8 and reach the highest score of 0.94 under 5% label.

By observing the four sub-figures, it is found that although the F1 value and recall of the deep network models are lower than that of the RF and DT algorithm on 5% label, the four metrics of the proposed model are obviously higher than those of traditional machine learning algorithms on 10%, 15%, and 20% labels. These results indicate that the deep network model is superior to the traditional machine learning algorithm. Meanwhile, the performance of the deep model with domain name vector is extremely close to that of the model with the domain type, which shows that the domain vector learning method can be used instead of the domain type labeling method to better protect students' privacy.

Performance comparison of DNN model based on different behavior data

To verify the importance of multi-source behavior for achievement prediction, in this section, we predict students' achievement based on each kind of behavior data separately using the proposed deep network model as the program framework, in which the model of behavior correlation feature extraction is shielded. Table 8 shows the performance of the models based on every type of behavior data. Figure 5 facilitates visual comparison of performance.

Through observation of Fig. 5, it is found that the four evaluation metrics of prediction performance based on multi-source behavior are all higher than that of any kind of single behavior data, especially in the macro recall metric, and the recall of the 20% label reaches 0.77. The difference between prediction performance based on single behavior is extremely little in macro precision and accuracy, but there are some differences in macro F_1 value and recall. Overall, prediction performance based on breakfast behavior and dinner behavior is the best, followed by gateway login behavior and shopping behavior, and then lunch behavior. The worst behaviors are web page browsing behavior and library entry behavior. These results have many implications for our observations of behavior. Students who often have breakfast are usually diligent and can get up early for class, and students who often have dinner can make full use of the time in the evening to review or preview their lessons. In general, predicting grades based on library entry behavior and web browsing behavior should achieve good results, but many students in this data set have few records of library entry behavior, and they may browse webpages through mobile phones rather than on the local network of the campus, making it impossible to fully analyze students' web browsing behavior. As a result, the prediction results based on these two behaviors are poor.

Conclusion and future work

In this paper, an end-to-end deep neural network model is proposed for predicting students' academic performance based on their daily living behavior data on campus. Our model addresses the challenges of extracting features manually from multi-source heterogeneous behavior data. Various types of behavior sequences can be directly input into our model after simple preprocessing. LSTM is applied separately on each type of behavior sequence to learn their time-series features, but for behavior sequence with extremely long length or the one that has nominal attributes with lots of values, it is necessary to use 1DCNN to reduce sequence length or use embedding layer to learn the dense vector of nominal attributes before using LSTM, which makes LSTM more effective. After extracting time-series features of every behavior, 2DCNN is applied to capture the correlation features between different behaviors. Finally, these two types of behavioral features are concatenated with students' demographic information as the input of fully connected layer to predict academic performance level. Experiments were conducted on the daily behavior data from 8228 students. Their results show that our model outperforms traditional machine learning algorithms. Meanwhile, our model has good scalability and versatility, and it can easily take new type of behavior data as input and be transferred to other application scenarios such as students' mental health diagnosis and employment choice consultation.

In practical applications, academic performance should be dynamically predicted over time, rather than based on the behavior data in a fixed period. In addition, we should not only identify students of high risk, but also know what causes poor performance. Therefore, we should collect more data to dynamically predict and enhance the interpretability of our deep model in the future work.

Funding This study was funded by National Natural Science Foundation of China (Grant 62072015, Grant U19B2039, Grant U1811463), Beijing Natural Science Foundation (No.4202004), Beijing University of Technology (ER2020B088) and China Association of Higher Education (No.2020XXHYB16).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Sousa JMD, Moreira CA (2018) Anxiety, depression and academic performance: a study amongst Portuguese medical students versus non-medical students. Acta Medica Port 31(9):454–462. https://doi.org/10.20344/amp.9996
2. Wallin AS, Zeebari Z (2018) Suicide attempt predicted by academic performance and childhood IQ: a cohort study of 26 000 children. Acta Psychiatr Scand 137(4):277–286. https://doi.org/10.1111/acps.12817
3. Orozco R, Benjet C et al (2018) Association between attempted suicide and academic performance indicators among middle and high school students in Mexico: results from a national survey. Child Adolesc Psychiatry Ment Health. https://doi.org/10.1186/s13034-018-0215-6
4. Horn D, Kiss HJ (2018) Which preferences associate with school performance?—Lessons from an exploratory study with university students. PLoS One 13:e01901632. https://doi.org/10.1371/journal.pone.0190163
5. Cuadra-Peralta A, Veloso-Besio C et al (2015) Relationship between personality traits and academic performance in university students. Interciencia 40(10):690–695
6. Conard MA (2006) Aptitude is not enough: how personality and behavior predict academic performance. J Res Personal 40(3):339–346. https://doi.org/10.1016/j.jrp.2004.10.003
7. Eliasson AH, Eliasson CJL (2010) Early to bed, early to rise! sleep habits and academic performance in college students. Sleep Breath 14(1):71–75. https://doi.org/10.1007/s11325-009-0282-2
8. Wang G et al (2016) Sleep patterns and academic performance during preparation for college entrance exam in Chinese adolescents. J. School Health 86(4):298–306. https://doi.org/10.1111/josh.12379
9. Wernette MJ, Emory J (2017) Student bedtimes, academic performance, and health in a residential high school. J Sch Nurs 33(4):264–268. https://doi.org/10.1177/1059840516677323
10. Maher C, Lewis L et al (2016) The associations between physical activity, sedentary behaviour and academic performance. J Sci Med Sport 19(12):1004–1009. https://doi.org/10.1016/j.jsams.2016.02.010
11. Ishihara T, Morita N et al (2018) Direct and indirect relationships of physical fitness, weight status, and learning duration to academic performance in Japanese schoolchildren. Eur J Sport Sci 18(2):286–294. https://doi.org/10.1080/17461391.2017.1409273
12. Ansari WE, Suominen S, Draper S (2017) Correlates of achieving the guidelines of four forms of physical activity, and the relationship between guidelines achievement and academic performance: undergraduate students in Finland. Cent Eur J Public Health 25(2):87–95. https://doi.org/10.21101/cejph.a4387
13. Keating Deng X, Castelli et al (2013) Association of weekly strength exercise frequency and academic performance among students at a large university in the United States. J Strength Cond Res. https://doi.org/10.1519/JSC.0b013e318276bb4c
14. Serra R et al (2020) Binge eating and purging in first-year college students: prevalence, psychiatric comorbidity, and academic performance. Int J Eat Disord 53(3):339–348. https://doi.org/10.1002/eat.23211
15. Valladares M et al (2016) Association between eating behavior and academic performance in university students. J Am Coll Nutr 35(8):699-703
16. Whelan E, Islam AN, Brooks S (2020) Applying the SOBC paradigm to explain how social media overload affects academic performance. Comput Educ 143:103692. https://doi.org/10.1016/j.compedu.2019.103692
17. Yan H et al (2017) Associations among screen time and unhealthy behaviors, academic performance, and well-being in Chinese adolescents. Int J Environ Res Public Health. https://doi.org/10.3390/ijerph14060596
18. Nayak JK (2018) Relationship among smartphone usage, addiction, academic performance and the moderating role of gender: a study of higher education students in India. Comput Educ 123:164–173. https://doi.org/10.1016/j.compedu.2018.05.007
19. Busalim AH, Masrom M, Zakaria WNBW (2019) The impact of facebook addiction and self-esteem on students' academic performance: a multi-group analysis. Comput Educ 142:103651. https://doi.org/10.1016/j.compedu.2019.103651''5
20. Riestra-Gonzalez M, Del Puerto Paule-Ruiz M et al (2021) Massive LMS log data analysis for the early prediction of course-agnostic student performance. Comput Educ 163(1):104108. https://doi.org/10.1016/j.compedu.2020.104108
21. Conijn R, Snijders C et al (2017) Predicting student performance from LMS data: a comparison of 17 blended courses using Moodle LMS. IEEE Trans Learn Technol 10(1):17–29. https://doi.org/10.1109/TLT.2016.2616312
22. Phan T, McNeil SG et al (2016) Students' patterns of engagement and course performance in a massive open online course. Comput Educ 95:36–44. https://doi.org/10.1016/j.compedu.2015.11.015
23. Mingyu Z, Sutong W et al (2021) An interpretable prediction method for university student academic crisis warning. Complex Intell Syst. https://doi.org/10.1007/s40747-021-00383-0
24. Cao Y, Gao J et al (2018) Orderliness predicts academic performance: behavioural analysis on campus lifestyle. J R Soc Interface 15(146). https://doi.org/10.1098/rsif.2018.0210
25. Yao H, Lian D et al (2019) Predicting academic performance for college students: a campus behavior perspective. ACM Trans Intell Syst Technol. https://doi.org/10.1145/3299087
26. Zhou Q, Quan W et al (2018) Predicting high-risk students using internet access logs. Knowl Inf Syst 55(2):393–413. https://doi.org/10.1007/s10115-017-1086-5
27. Zhang X, Sun G et al (2018) Students performance modeling based on behavior pattern. J Ambient Intell Humaniz Comput 9(5SI):1659–1670. https://doi.org/10.1007/s12652-018-0864-6
28. Ghosh S, Ghosh SK et al (2018) Exploring the association between mobility behaviours and academic performances of students: a context-aware traj-graph (CTG) analysis. Prog Artif Intell 7(4):307–326. https://doi.org/10.1007/s13748-018-0164-6
29. Yang Z, Yang J et al (2020) Using convolutional neural network to recognize learning images for early warning of at-risk students. IEEE Trans Learn Technol 13(3):617–630. https://doi.org/10.1109/TLT.2020.2988253
30. Pu H et al (2021) Predicting academic performance of students in Chinese-foreign cooperation in running schools with graph convolutional network. Neural Comput Appl 33(2):637–645. https://doi.org/10.1007/s00521-020-05045-9
31. Botelho AF, Varatharaj A et al (2019) Developing early detectors of student attrition and wheel spinning using deep learning. IEEE Trans Learn Technol 12(2SI):158–170. https://doi.org/10.1109/TLT.2019.2912162

32. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/arXiv.1409.1556
33. Chawla NV, Bowyer KW et al (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953
34. Han H, Wang WY et al (2005) Borderline-smote: a new over-sampling method in imbalanced data sets learning. Lect Notes Comput Sci 3644:878-887
35. Li X, Zhang Y et al (2021) An unsupervised ensemble clustering approach for the analysis of student behavioral patterns. IEEE Access 9:7076–7091. https://doi.org/10.1109/ACCESS.2021.3049157

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.