JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 06/06 pp180−193
Volume 5, Number 2, June 2024
DOI: 10.23919/JSC.2024.0015

Online Learning Behavior Analysis and Prediction Based on Spiking Neural Networks

Yanjing Li, Xiaowei Wang, Fukun Chen, Bingxu Zhao, and Qiang Fu*

Abstract: The vast amount of data generated by large-scale open online course platforms provide a solid foundation for the analysis of learning behavior in the field of education. This study utilizes the historical and final learning behavior data of over 300 000 learners from 17 courses offered on the edX platform by Harvard University and the Massachusetts Institute of Technology during the 2012–2013 academic year. We have developed a spike neural network to predict learning outcomes, and analyzed the correlation between learning behavior and outcomes, aiming to identify key learning behaviors that significantly impact these outcomes. Our goal is to monitor learning progress, provide targeted references for evaluating and improving learning effectiveness, and implement intervention measures promptly. Experimental results demonstrate that the prediction model based on online learning behavior using spiking neural network achieves an impressive accuracy of 99.80%. The learning behaviors that predominantly affect learning effectiveness are found to be students’ academic performance and level of participation.

Key words: online learning; learning outcomes prediction; learning behavior analysis; spiking neural network

1 Introduction

As online courses and learning technology systems continue to evolve, online learning has become a prevalent part of various educational contexts. This shift has opened up new opportunities for comprehensive research into learner behavior. The data collected during the learning process reflect not only the demographic characteristics of learners, such as age, gender, educational background, and geographical location, but also their behavioral patterns. These patterns include frequency of video views, number of days participating in the course, quantity of discussion posts, and the progress of each student towards course completion. As students engage in autonomous and exploratory learning based on their personal preferences, their behavior naturally manifests online, providing a platform to record their implicit learning activities. These behaviors, often unconscious and fragmented, possess underlying correlation structures and subtle rules that more accurately represent the intricate logical relationships involved in learning than explicit behaviors do. Such behaviors present a genuine depiction of students’ learning experiences, which are often difficult for educators or assessment mechanisms to capture. Therefore, selecting appropriate methods to analyze and extract the hidden information within these behaviors can reveal influential factors in the

- Yanjing Li is with the Institute of Education Science Research, Heilongjiang University, Harbin 150080, China. E-mail: yanjing_li01@163.com.
- Xiaowei Wang is with the School of Cyberspace Security, Shandong University of Political Science and Law, Jinan 250014, China. E-mail: 002298@sdupsl.edu.cn.
- Fukun Chen and Bingxu Zhao are with the School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China. E-mail: fukun@hrbeu.edu.cn; tozhaobingxu@gmail.com.
- Bingxu Zhao is also with the Faculty of Electrical and Computer Engineering, University of Victoria, Victoria, V8P 5C2, Canada.
- Qiang Fu is with the Guangxi Key Lab of Brain-Inspired Computing and Intelligent Chips, School of Electronic and Information Engineering, Guangxi Normal University, Guilin 541004, China. E-mail: qiangfu@gxnu.edu.cn.

* To whom correspondence should be addressed.

Manuscript received: 2024-02-18 ; revised: 2024-06-20; accepted: 2024-06-21

educational process. This insight can then serve as a valuable reference for tracking, evaluating, and enhancing learning experiences.

Over recent years, research into learning behavior analysis and prediction has garnered increasing interest from scholars. Through predictive analysis, researchers can visualize learning behavior, discern the relationship between learning behavior and outcomes, and identify the behaviors that influence these outcomes. However, online education, lacking in face-to-face interaction and effective supervision between teachers and students, presents unique challenges not found in traditional offline education. Notably, online courses often suffer from high dropout rates and low completion rates, issues that have drawn significant attention and research efforts. Studies have shown that the average course completion and final certificate acquisition rates on platforms like edX are merely 5%[1, 2], indicating dropout rates typically range between 91%–93%[3]. Consequently, researchers are focusing on predicting learners' pass and dropout rates, aiming to accurately identify behaviors that contribute to successful course completion and implement corresponding interventions to boost completion rates.

In this study, we utilized an open dataset comprising over 300 000 individuals' historical learning behavior records, relevant background information, and final academic performance on the edX platform. We organized and analyzed this recorded information using spiking neural networks (SNNs) and classical data analysis techniques based on feature types. The neural networks were trained to predict whether learners would complete courses, identify learning behaviors that impact academic performance, and validate the model's effectiveness. This process demonstrated the accuracy of our predictive approach.

The SNN model can accurately predict learners' academic performance. By analyzing the key attributes within the prediction model, it provides insights for educators to implement targeted intervention strategies in subsequent teaching processes, thereby enhancing educational effectiveness. The primary sources of data for this study were extracted from Harvard University and the Massachusetts Institute of Technology. The paper is organized into sections including related work, proposals for utilizing spiking long short-term memory (LSTM) methods, experimental results, conclusions, and future work.

The main contents of this paper are as follows:

(1) We proposed an SNN model to identify learning behaviors that impact learners' final certification outcomes based on their historical behavior data and final learning scores. This model can predict whether a learner will obtain a certificate based on future learning behavior data. Furthermore, when the learner's final academic performance, certification status, and learning background information or behavior characteristics are known, the model can determine which actions influence the learner's final academic performance.

(2) We validated the effectiveness of our model. Training was conducted based on learners' relevant learning characteristics and final learning scores, and compared with the original dataset, the model achieved an accuracy of 99.8%. The model was benchmarked against traditional models, and experimental results were provided to verify its effectiveness. Utilizing behavior data is a critical factor in generating more accurate predictions for intelligent warnings and interventions.

(3) We employed the Pearson correlation coefficient to analyze the correlation between behavioral characteristics and learning outcomes. This method allowed us to identify key behavioral characteristics that affect students' academic completion. These findings can provide references for improving massive open online courses (MOOCs), learners, and teachers by identifying learning behaviors that influence learners' ability to obtain certificates.

This paper is organized as follows: Section 2 presents related work. Section 3 proposes the labelling method for LSTM. Experimental results are detailed in Section 4, Discussions are presented in Section 5, and conclusions and future work are outlined in Section 6.

2 Related Work

MOOCs generate vast amounts of educational data and have garnered worldwide attention since their inception, becoming the platform with the highest number of learners globally. However, the effectiveness of MOOCs has been questioned, particularly given the extremely low course completion rates, a pressing issue that the MOOC platform currently faces. Existing methods for analyzing learners' behavior and course completion rates are largely based on data mining. These methods first extract features from the various educational data generated by learners during the learning process and then employ different mining algorithms for final prediction.

Many researchers have used educational data to analyze learning behavior and predict completion rates, investigating the relationship between different learning behaviors and the final course completion rates, as well as the link between learning behavior and academic effectiveness. For instance, some studies have analyzed the key behaviors that affect learners' learning outcomes from a fine-grained learning behavior perspective, concluding that learners exhibit significant differences in course selection and learning modes. It has also been observed that learners are influenced by their peers when obtaining certificates. If a learner has one or several friends who have obtained a course certificate, the probability of the learner obtaining the certificate will greatly increase.

Other studies have employed logistic regression (LR) as a classifier and used learning performance and social interaction data from the first week of MOOC courses to predict the probability of learners completing courses and obtaining certificates. Some researchers have used datasets, including learners' backgrounds, from the edX platform to analyze the relationship between learners' backgrounds, learning behaviors, and final grades. Further research has analyzed learners' discussion behavior on forums and found that learners' positive and constructive discussion behavior is an essential factor affecting learning.

In summary, previous research questions on the prediction and analysis of learning behavior can be categorized into three types. Research on MOOC learning behavior analysis and prediction mainly uses advanced data mining and machine learning techniques. With the data provided by MOOCs, researchers have developed new algorithms and models to identify behavior patterns that contribute to learning effectiveness and completion. Many studies have focused on predicting dropout rates, success rates, and learning effects. These studies have used data mining, machine learning, deep learning, and neural network models to analyze and predict learners' behaviors, combined student engagement data with demographic and past performance data for prediction, and used various experimental designs to test the efficacy of different models.

Moreover, researchers have investigated the impact of factors such as learners' age, gender, educational background, and learning behavior on completion rates to construct predictive models that can identify students at high risk of dropping out. These models assist educators in recognizing the critical factors that contribute to a student's success or failure in learning and help design interventions to improve student outcomes.

In conclusion, recent research on MOOC learning behavior analysis and prediction has shown the potential to improve learning outcomes in online education. Further research will undoubtedly continue to develop methods and models that enhance the usefulness of MOOCs as a tool for delivering education.

The main related works are as follows:

(1) Prediction and research on dropout rates

A research group from the Massachusetts Institute of Technology's Digital Learning Office and the Harvard Research Committee discovered that a majority of MOOC learners leave the course in its initial phase. On average, 50% of learners drop out within one or two weeks after registering for a course. The dropout rate then dramatically decreases to 16% in the course's second week[4]. The objective was to predict the dropout rate based on learners' behavior data in the approach of Ref. [5]. They employed a convolutional neural network to extract local features and proposed the contrastive learning based survival analysis (CLSA) model, which uses a kernel strategy to establish feature relationships. The model was successful, achieving an accuracy of 86.9%, a 1.6% improvement over the previous prediction accuracy. The goal of Ref. [6] was to understand the primary reasons why MOOC learners continue to learn and to elucidate how these reasons impact the learner's completion of the full course. The approach of Ref. [7] used the changes in learners' weekly historical data as a feature and applied a support vector machine to analyze learners' click sequences. This method determined the dropout rate, which yielded better results than the general feature. According to a time prediction mechanism, the approach of Ref. [8] used a deep learning algorithm to construct a dropout prediction model to predict the learners' dropout rate. This model outperformed previous baseline algorithms and achieved more accurate predictions. They also proposed a method for optimizing MOOC learners' learning behavior and intervening early. Starting from a network perspective, the approach of Ref. [9] used clickstream data in the learning process to create a network model of collective attention. They observed the learning behavior of learners who drop out during the course, improved the accuracy of predicting dropouts, and also proposed measures to prevent mid-course dropouts. The approach of Ref. [10] constructed a dropout prediction model for MOOCs based on a broad learning system (BLS). The output layer of this model was used for dropout prediction. Compared to other similar methods, this model significantly reduced BLS training time and could predict dropout rates with high accuracy. This research contributes to the development of more effective strategies for reducing dropout rates in MOOCs.

(2) Mining the relationship between learning behavior and learning outcomes

Research exploring the relationship between learning behavior and final outcomes primarily uses learners' historical learning behavior records to predict learners' ultimate academic performance, including whether they will obtain certificates. This approach also aids in determining the link between learning behavior and academic performance, and in identifying learners at risk of poor learning outcomes. The approach of Ref. [11] utilized a substitution algorithm to analyze the correlation between students' behavior and their final learning performance. Meanwhile, the approach of Ref. [12] introduced a time-series neural network that integrates extensive click video behavior data to forecast the learning performance of MOOC learners. The approach of Ref. [13] proposed a method that predicts participants' learning behaviors based on the learning behavior data of MOOC participants during their learning process. This approach selected 19 distinct learning behaviors. The model's construction consists of four parts: data collection, correlation analysis between learning behaviors, prediction of participants' course completion, and early intervention in learning behaviors. A rule-based genetic programming algorithm was also employed to predict the participants' behavior. The model achieved commendable prediction results, with a high prediction accuracy. The approach of Ref. [14] proposed a new simple feature matrix and a convolutional neural network (CNN) to predict whether learners obtain certificates and the dropout rate. This study deeply analyzed the characteristics of learners' learning behaviors. Finally, it is important to note that the correlation between learners' learning behaviors should not be overlooked. In a different vein, the approach of Ref. [15] analyzed the relationship between learners' MOOC learning data and learning outcomes from the perspective of students' learning motivation, attitude, and challenges. This comprehensive approach provides a holistic view of the factors influencing learning outcomes and paves the way for more effective learning strategies and interventions.

(3) Predictive analysis and research on pass rates

Various studies have utilized diverse online learning behaviors of learners as potential features to predict the final score[16]. Leveraging these potential features, they predicted the likelihood of learners passing the final examination. The approach of Ref. [17] delved into the influence of demographic characteristics such as age, gender, and geographical location on the course passing rate. Another study[18] amalgamated learning behavior data with text data, scrutinizing learners' online learning activities and the text generated on the platform's forum, and predicted the probability of learners completing the course.

Further, the approach of Ref. [19] conducted a comprehensive investigation from a cognitive psychology perspective. They enhanced learners' enthusiasm by incorporating questions into MOOC course videos, validated the efficacy of this strategy in reducing drop-out rates, and analyzed the impact of various communication measures on the passing rate. Meanwhile, the approach of Ref. [20] employed a score prediction method, predicting whether learners could complete courses and obtain certificates by analyzing the learning activity trails generated by learners. The experimental results indicated that the proposed method improved prediction accuracy, allowing for more nuanced analysis of students and adjustments to the analysis to cater to a broader range of learners.

Given the notable low completion rate of MOOCs, prior research has employed various methods to scrutinize and investigate learner data. The diverse learning behavior characteristics have yielded varying prediction outcomes, leading to considerable fluctuations in prediction results and compromised accuracy. In essence, these preceding works have imparted valuable insights into the analysis and prediction of MOOC learning behaviors, underscoring the potential of leveraging sophisticated data mining and machine learning techniques to pinpoint the pivotal factors influencing MOOC learner behavior and ameliorate the learning outcomes of MOOCs. Against this backdrop, our paper puts forth the proposition of utilizing the SNN to predict the likelihood of learners securing certificates, predicated on their characteristics.

3 Proposed Method

In this section, we develop the SNN to predict students' academic performance and verify the efficacy of the proposed model. The process begins with data preprocessing, collection, and cleaning, which includes the removal of duplicate data, handling of missing values, and identification and treatment of outliers. We process and present relevant information about learners' characteristics, such as age, educational background, and gender, more directly using graphics and percentages.

Following this, we undertake feature selection, a step that involves identifying learning behavior characteristics that may potentially influence academic performance. This selection is grounded in both practical experience and theoretical considerations. The primary features under consideration include learning time, learning frequency, course engagement, forum participation, and completion of assignments.

3.1 Dataset

The dataset used in this study is the first large-scale MOOC open dataset, jointly released by Harvard University and the Massachusetts Institute of Technology on14 May 2014. This dataset includes open data for 17 courses on the edX platform for the 2012–2013 academic year. This article selects over 300 000 learning behavior records and conducts statistical analysis on learner characteristics. These data are summary records, each representing the corresponding characteristics of learners in the edX course (such as course ID, user ID, registration date, course registration date, last interaction date with the course, number of days students participated in course interaction, number of video events played in the course, and number of posts or discussions posted on the forum). The interpretation of each feature in the data is shown in Table 1.

Table 1 Detailed information of the dataset.

Information type	Attribute feature	Description
Course information	Course_id	Course name and semester
	User id_DI	User's registration ID
Learner type information	Registered	Registered learners of the course
	Viewed	Browse the learners of the course
	Explored	Active learners, anyone who has visited at least half of the chapters
	Grade	Score, the final score of the course, ranging from 0 to 1
	Certified	Anyone who gets a certificate. The certificate is awarded based on the course score, and deadline of the certificate will vary according to the course.
Learner information	final_Cc_Cname_DI	Country of the registered learner
	LoE	Educational background, the highest educational level of users, including “middle school”, “bachelor”, “master”, and “doctor”
	YoB	Provided by the user, year of birth. Example: “1980”
	Gender	Possible values: m (male), f (female)
Learner behavior information	start_time_DI	Date of course registration
	last_event	Date of the last interaction with the course. If there is no interaction, it is an interaction other than blank registration.
	nevents	Number of interactions that students participate in the course, which is recorded in the student learning activity tracker; if there is no registration, there is no interaction, and it is blank.
	ndays_act	Number of days students participate in the course interaction
	nplay_video	Number of video events played in the course
	nchapters	Number of chapters in which students participate in the interaction
	nforum_posts	Number of posts or discussions posted on the forum

Each row in the dataset represents the learning records of students enrolled in a particular course, while each of the 18 columns provides descriptive data about the learners or their behaviors. In this study, the main variables are divided into four categories: course information, learner's basic information, learner's type information, and learner's behavioral information. The basic information about learners includes their country, level of education, age, and gender. The course information includes the course ID and user ID. The types of learners are categorized as registrants, viewers, active learners, and certificate recipients. Learner behavior information includes course start date, last login date, number of chapters studied, number of forum posts, number of video playbacks, and scores. The original dataset contains the learning records of 338 224 learners, complete with course and user IDs.

Based on the edX open dataset, using the data analysis method, aiming at the problem of low probability of obtaining certificates for learners in the platform, characteristics such as learners' learning background and learners' learning behavior are selected as the input of SNN to predict the relationship between learners' behavior and the final acquisition of certificates. If the analysis results are not relevant, the behavior characteristics are discarded. If relevant, the behavior characteristics are retained as the eigenvalues of the algorithm model. The study used the Pearson correlation analysis method to analyze the variables of learning behavior and learning results. The pearson correlation coefficient is a kind of linear correlation coefficient, which is employed to calculated the linear correlation degree between two variables. The degree of linear correlation between two variables can be described by the correlation coefficient r . The calculation method is shown in Eq. (1).

$$r=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \tag{1}$$

Among them, the value of r is between −1 and +1. When r is positive, it means that the two variables are positively correlated; when r is negative, it means that the two variables are negatively correlated. When the value of r is 0, it means that there is no linear correlation between the two variables. The larger the absolute value of r, the stronger the linear correlation between the two variables.

(1) Learner category: According to the learning records on the platform, learners are divided into four categories: Only registered: after registering an account, they have not completed any learning tasks. Only viewed: after registering an account, the learning content is less than half of the course content. Active learners (only explored): After registering an account, the learning content exceeds half of the course content. Certified: the learner who finally obtains the certificate. Among them, only 117 114 learners registered, 193 856 registered and watched the course, 20 624 registered and actively studied, and 6570 finally obtained the certificate, as shown in Fig. 1.

(2) Nationality: The nationalities of learners include world learners from 26 countries, such as China, the United States, France, Canada, Japan, Germany, India, etc.

(3) Education background: Statistics are carried out at five levels, including secondary school degree or below, secondary school degree, bachelor's degree, master's degree, and doctoral degree, as shown in Fig. 2. Among them, there are 7929 learners with a secondary school degree or below, 93 537 with a secondary school degree, 119 511 with a bachelor's degree, 71 702 learners with a master's degrees, 9354 with a doctoral degree, and 36 189 without academic information. The average education level of learners is bachelor's degree, accounting for 35.3% of the number of learners. The learners of master's degree account for 21.2%. Doctoral degree learners account for 2.8% of the total number of learners, 27.7% of whom have a secondary school degree, and 2.3% of whom are in secondary school.

(4) Age: We divide the age of users into 7 categories, as shown in Fig. 3. Among them, there are 3697

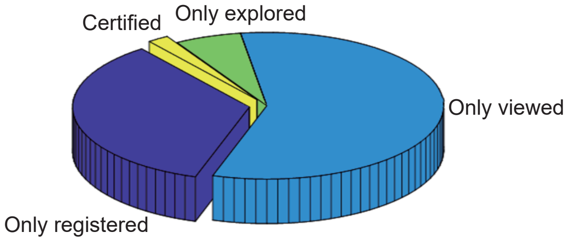

Fig. 1 Percentage of learners with different types of curriculum exploration.

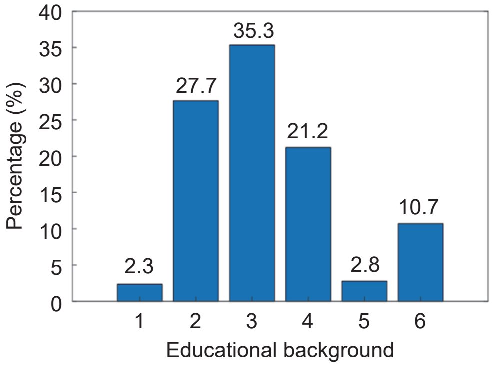

Fig. 2 Percentage of learners' educational background. (The values on the *X*-axis: 1: Secondary school degree or below; 2: Secondary school degree; 3: Bachelor's degree; 4: Master's degree; 5: Doctoral degree; 6: Without academic information.)

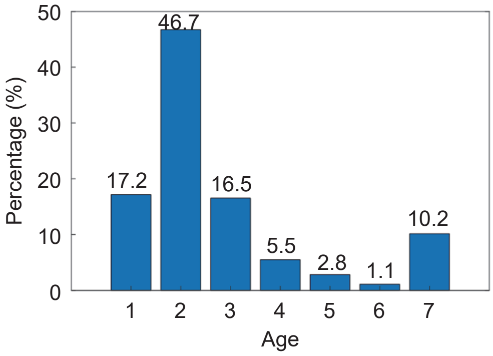

Fig. 3 Percentage of learners of different ages. (The values on the *X*-axis: 1: Aged under 20; 2: Aged 21–30; 3: Aged 31–40; 4: Aged 41–50; 5: Aged 51–60; 6: Aged over 60; 7: Without age information.)

learners aged over 60, 9517 learners aged 51–60, 18 597 learners aged 41–50, 55 878 learners aged 31–40, 158 083 learners aged 21–30, 58 091 learners under 20, and 34 359 learners lack age information. The average age of learners is 28 years old. The maximum age of learners is 82 years old, and the minimum age is 9 years old. Among them, the number of learners aged 21–30 is the largest, accounting for 46.7% of learners, followed by 52 887 learners under the age of 20, accounting for 17.2% of learners, and 16.5% of learners aged 31–40. Our choice of these categories was based on the objectives of our study and the nature of our data. Our goal was to study the overall trends in online learning behavior among learners of different age groups, rather than the detailed age differences. Furthermore, the age information in our dataset is only available at a granularity of ten years, so we were unable to make more detailed age divisions.

(5) Gender: The dataset includes 208 184 male learners and 97 061 female learners. 32 981 of gender information is missing from data, and 61.6% are male learners, as shown in Fig. 4.

(6) Learning time: It includes the start time of the course and the last event. A few weeks before the start of the course, there is a record of registering for the course, and more than half of the learners complete the registration before the beginning of the course. About 8% of learners register for the course at the end of the course. The learner's learning time in the text refers to the date of the learner's last learning record minus the registration time.

(7) nevents: The number of interactions between learners and courses can be obtained by analyzing the system log.

(8) ndays_act: This metric represents the count of unique days on which students interacted with the course.

(9) nplay_video: This metric represents the total number of videos watched by learners during the course.

(10) nchapters: This metric represents the number of complete chapters learners engaged with from the start to the end of the course.

(11) nforum_posts: Topics are initiated by learners on the learning forum about the course content, including answering questions raised by others.

(12) Grade: Score, the final score of the course, ranging from 0 to 1. Learners' scores are the output of their learning behavior, reflecting their participation and effort in the course. Although the score itself is a quantification of learning outcomes, it is considered a behavioral characteristic here because it is a direct result of learning behavior and can be used to predict whether learners can obtain a certificate. Therefore, we incorporate scores into learners' behavioral

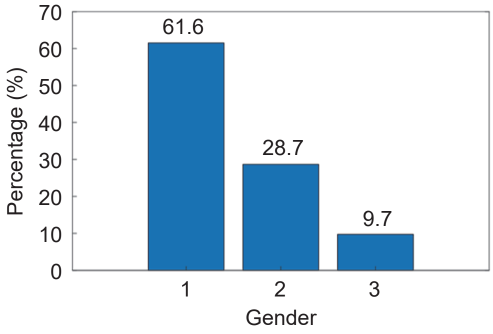

Fig. 4 Percentage of learners of different genders. (The values on the *X*-axis: 1: Male; 2: Female; 3: Without gender information.)

characteristics to facilitate a deeper study of the relationship between learning behavior and obtaining a certificate. Scores obtained range from 0 to 1. The number of learners who obtained scores is 20 964, that is, learners whose scores are not zero. From the dataset, we can see that learner with scores below 0.5, even though they obtained scores, did not get a certificate, and their number is 13 254. For learners with scores between 0.51 and 0.8, some obtained certificates. Among them, there are 520 learners with scores between 0.51 and 0.6. There are 1333 learners with scores between 0.61 and 0.7, and 1021 learners with scores between 0.71 and 0.8. All learners with scores above 0.81 obtained certificates, and their number is 4836.

3.2 Key behavior

The primary objective of this section is to identify critical learning behavior characteristics that influence learning outcomes. For this purpose, we employed the Pearson correlation coefficient to calculate the correlation between each learning behavior feature and the acquisition of a certificate. We visualized the impact of 11 features on students' certificate attainment, and the results are illustrated in Fig. 5. The horizontal axis in Fig. 5 represents each attribute feature, while the vertical axis represents the corresponding weights of each attribute feature.

The Pearson correlation coefficient was used to quantify the relationship between various learning behavior features and the attainment of a certificate. Here is how each feature correlates:

Grades: There is a strong positive correlation between grades and certificate acquisition, indicating that as grades increase, the likelihood of certificate attainment also increases.

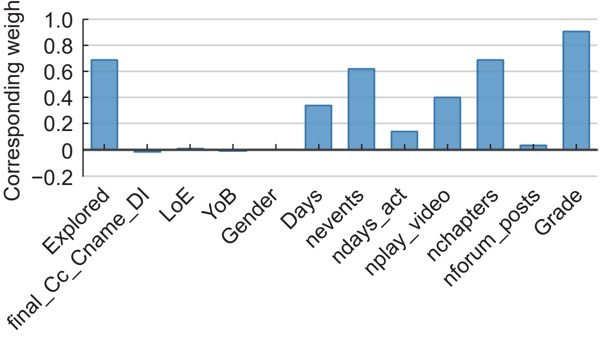

Fig. 5 Correlation analysis between labels and features (Pearson).

Exploration, event, and nchapters: These three features exhibit a moderate positive correlation with certificate attainment. This suggests that while they do impact certificate attainment, the effect is not as strong as grades.

Video interruptions, learning days, and forum discussions: These features show a weak positive correlation with certificate attainment. This suggests that while these factors do influence certificate attainment, the effect is relatively minor.

YoB and final_Cc_Cname_DI: These features show no correlation with certificate attainment, indicating that they do not influence the likelihood of a student earning a certificate.

In summary, the Pearson correlation coefficient was used to reveal that grades, exploration, event, and nchapters are the most influential factors in certificate acquisition, while video interruptions, learning days, and forum discussions have a lesser impact. Personal background information (YoB, final_Cc_Cname_DI) does not influence certificate attainment.

3.3 Spiking neural network

The SNN is considered as a special class of artificial neural networks. It mimics the dynamics and learning capabilities of the brain in a more bio-inspired way. The spiking neuron is the basic unit of SNNs, which communicated by firing spikes (action potentials). As shown in Fig. 6, spiking neurons interconnected by synapses, which simulate the information transmission mechanism of biological neurons. It models the process that the ion channel on the cell membrane is opened by neurons receiving stimulation, and then the charged ions inside and outside the cell membrane flow to generate an action potential. A spike will be generated if the action potential reaches a threshold. Then the

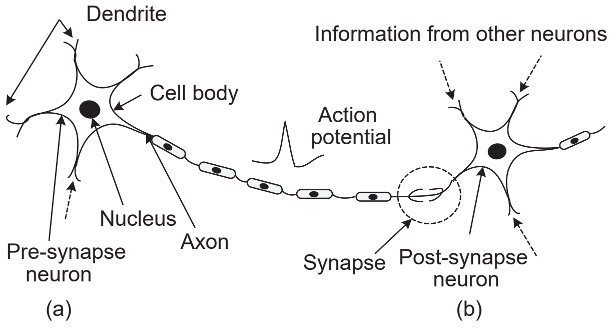

Fig. 6 Generation and transmission of spikes.

spike is transmitted along the axon to the nerve terminal. Finally, it is transmitted to the postsynaptic neuron through the synapse. The leaky integrate-and-fire (LIF)[21] model is one of the most fundamental neuron models in neuroscience, simulating the main characteristics of neurons, namely integrating input signals and generating spikes when a threshold is reached. In the LIF model, a neuron's membrane potential gradually decays (or "leaks") over time, while input signals cause the membrane potential to increase. When the membrane potential reaches a certain threshold, the neuron fires an action potential (or "spike"), and the membrane potential is immediately reset. In our research, we apply the principles of the LIF model to the LSTM SNN model. LSTM is a special type of recurrent neural network that effectively addresses the vanishing and exploding gradient problems when processing time series data. We implement LIF dynamics in the neurons of the LSTM, meaning that each neuron integrates its input signals and fires a pulse when the threshold is reached. This design allows our model to better simulate the behavior of biological neural systems while maintaining the advantages of LSTM. The connection between these two models lies in their attempt to simulate the behavior of biological neural systems. The LIF model provides a simple yet effective way to simulate the behavior of individual neurons, while the LSTM SNN model utilizes these behaviors to handle complex time series tasks. The simple circuit of the LIF model consists of one capacitor and one resistor in parallel. It can be calculated by

$$I(t) = C_m \frac{\mathrm{d}V_m}{\mathrm{d}t} + \frac{V_m}{R_m} \tag{2}$$

where C_m is the value of membrane capacitance, V_m is the value of membrane voltage, R_m is the value of the membrane resistance, and $I(t)$ is the total membrane current. $\tau = RC$ is the time constant of leakage current, which is calculated by

$$\tau \frac{\mathrm{d}V_m}{\mathrm{d}t} = -V_m(t) + RI(t) \tag{3}$$

The spiking LSTM networks is developed to perform the task of learning results prediction. Due to the non-differentiable characteristic of spiking neurons, the original SNN cannot be trained by the backpropagation algorithm. However, the spiking LSTM can be trained by the backpropagation algorithm. The LSTM architecture is shown in Fig. 7. The output value of a spiking neuron before applying the activation is called membrane potential. It can be denoted as $V_m(t)$ for a spiking neuron at time t. A unit of spiking LSTM consists of three interacting gates and associated "spike" functions. Generally, spike activations $\sigma_1(u)$ and $\sigma_2(u)$ are applied to each of their associated neurons individually. These functions take neurons' membrane potential $V_m(t)$ and output either a spike or null at each time step.

4 Experimental Result

4.1 Experimental setting

Convert learners' birth dates to their actual age. The invalid information is removed in the characteristics of education background, gender, age, time, number of videos viewed, number of learning chapters, number of learning events, and number of days students participate in course interaction.

Among them, the education background characteristics, gender characteristics, and age characteristics are nonnumerical forms. For the convenience of calculation, we have numerically processed these characteristics and numerically processed three positives. The education background characteristics include less than secondary (secondary education or below), secondary (secondary education), bachelor's (bachelor's degree), master's (master's degree), and doctoral (doctoral degree). Assign several features to 1, 2, 3, 4, and 5. For gender characteristics, men are represented by 1 and women by 0. The learner age characteristics have 7 categories. Among them, there are learners aged over 60, learners aged 51–60,

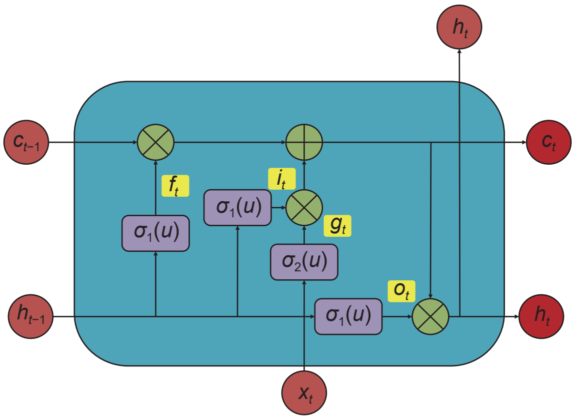

Fig. 7 Unit of spiking LSTM architecture.

learners aged 41–50, learners aged 31–40, learners aged 21–30, learners under 20, and learners who lack age information. The values are 7, 6, 5, 4, 3, 2, and 1, respectively. All experiments are conducted under a 10-fold cross-validation scheme. This method partitions the dataset into 10 subsets, and the model is trained and tested 10 times, each time using a different subset as the test set and the remaining ones as the training set. This ensures that our model is validated on different subsets of data, thereby preventing overfitting.

4.2 Experimental result

The experimental results of the network are shown on Fig. 8. The red curve represents the loss value, and the blue curve express the classification accuracy. As shown in Fig. 8, the loss value gradually decreases, and the classification accuracy can reach 99.80%. Therefore, spiking LSTM can be used to judge whether students can obtain certificates according to a given set of input characteristics.

Performance comparison of different methods on the online learning database is shown in Table 2. The research of Ref. [11] induced 19 behavior indicators in the online learning platform and randomly divided eigenvalue sets into 80% and 20%. It shows that 85.9% accuracy can be obtained by using the optimized logistic regression model. 89.2% accuracy can be obtained by using the LSTM[22], and 87.0% accuracy can be obtained by using the gated recurrent unit (GRU)[23]. The CNN with LSTM and the CNN with GRU can get higher accuracies than LSTM and GRU. The decision tree[24] and the random forest[25] achieve 87.7% and 92.1% accuracies, respectively. The result of our work is better than others on this dataset.

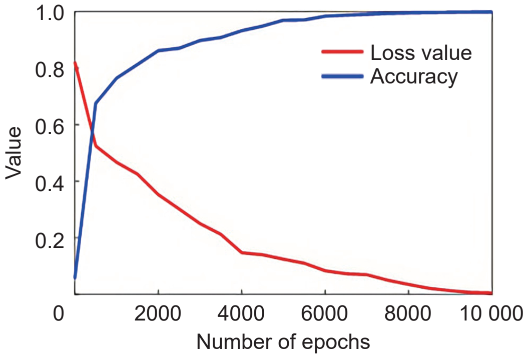

Fig. 8 Classification performance of the spiking LSTM.

Table 2 Performance comparison of several methods on the online learning database.

Method	Accuracy (%)
Logistic regression[11]	85.9
LSTM[22]	89.2
GRU[23]	87.0
CNN with LSTM	94.6
CNN with GRU	94.7
Decision tree[24]	87.7
Random forest[25]	92.1
This work	99.8

5 Discussion

In the field of education, the application of prediction and analysis is closely interconnected, together forming a dynamic feedback system that aids in improving teaching methods and enhancing learning outcomes. Analysis is a process whose aim is to understand the historical data of student behavior and learning outcomes, thus revealing past trends and patterns. By delving into these data, we can identify the key factors that impact learning outcomes. For instance, in this study, we employed Pearson correlation coefficient analysis to explore the relationship between factors such as scores, learning exploration behavior, learning events, and the number of chapters with whether students can obtain a certificate.

Based on these analysis results, we can use predictive models (like the spiking LSTM model used in this study) to forecast students' future behavior and learning outcomes. These predictions can provide valuable information for teachers and educational institutions, assisting them in formulating teaching strategies in advance. For example, if the model predicts that a student might encounter difficulties in the process of obtaining a certificate, educational institutions can provide timely additional support and resources to improve the student's learning outcomes.

In summary, the application of analysis and prediction in the field of education is complementary. Analysis offers a deep understanding of the past and present, while prediction provides insights into the future. These two processes together constitute a feedback system, which helps us to more effectively improve the educational process, enhance learning outcomes, and thus achieve the ultimate goal of education: promoting the holistic development and long-term success of students. In our study, we chose to utilize the entire dataset without employing methods to address class imbalance for the following reasons: In the real world, data imbalance is quite common, and our goal was for the model to make accurate predictions even in these imperfect circumstances. Additionally, we noticed that, despite the class imbalance in our dataset, the number of samples for each class was sufficient for the model to learn the characteristics of each class. However, we acknowledge the potential benefits of addressing class imbalance in improving the model's predictive accuracy and applicability across different educational settings.

This study used the spiking LSTM model to predict certificate attainment by online learners. The model's predictions help teachers, schools, and policymakers understand learner behaviors and their impact on learning outcomes, allowing for the implementation of appropriate teaching strategies. This study offers valuable insights and guidance for enhancing the efficiency of learners' educational processes, increasing certification rates, and informing the development of online education policies. These models can predict learner behaviors, learning paths, and outcomes, enabling more effective teaching strategies and early intervention. Specifically:

(1) The application of predictive models, particularly in online educational environments, holds significant potential for understanding and improving the learning process. This is seen in various aspects of education; for example, teachers can use predictive models to anticipate student performance, thus allowing for timely adjustments in teaching strategies. Schools can leverage these models to predict graduation or dropout rates, enabling early interventions. Moreover, policymakers can use such models to forecast the long-term effects of educational policies, leading to more informed decision-making.

In online education, predictive models can forecast various aspects related to learners. For instance, they can predict learner behaviors and learning paths, such as whether a student will complete a course, participate in online discussions, or when they might require additional support. Such predictions enable teachers to strategize in advance to meet students' needs. Additionally, these models can predict learning outcomes, such as final grades or the ability to obtain a certificate. This insight can help students understand their current learning status and adjust their learning strategies accordingly. It also assists teachers in identifying students who may need additional support.

In this study, the spiking LSTM model is employed to predict whether online learners will obtain a certificate. This predictive model helps educational institutions understand learner behaviors and their impact on learning outcomes. Thus, institutions can implement suitable teaching strategies based on the model's predictions. For example, if a learner is predicted to struggle with obtaining a certificate, course designers can provide timely additional support and resources to improve their learning outcomes.

Furthermore, predictive models can forecast the most likely successful learning paths based on students' past behaviors and performance. This capability allows educational institutions to provide personalized learning resources and support, thereby enhancing student learning outcomes.

(2) The Pearson correlation coefficient analysis provided insights into the factors influencing student certification. Grades showed a strong correlation, emphasizing the importance of academic performance. Moderate correlation with exploration, event, and nchapters highlighted the role of student engagement and interaction with course material. Factors like video interruptions, learning days, and forum discussions had a weak correlation, indicating that they have an impact but are not as critical. Interestingly, personal background factors showed no correlation, supporting the principle of educational equity.

Grades: The strong correlation between grades and certificate acquisition underscores the importance of academic performance in educational outcomes. This suggests that efforts to enhance student understanding and mastery of course material, such as through effective teaching methods, tutoring, and self-study resources, can significantly increase the likelihood of certificate attainment.

Exploration, event, and nchapters: The moderate correlation of these factors with certificate acquisition indicates that student engagement and interaction with the course material also play a significant role in educational outcomes. This suggests that fostering an interactive and engaging learning environment, where students are encouraged to explore and interact with course content, can further enhance the likelihood of certificate attainment.

Video interruptions, learning days, and forum discussions: The weak correlation of these factors with certificate acquisition suggests that while these elements of the learning experience do have an impact, they are not as critical as the factors mentioned above. However, this does not mean that they should be overlooked. Efforts to minimize video interruptions, encourage regular study habits, and promote active participation in forum discussions can still contribute to improved educational outcomes.

YoB and final_Cc_Cname_DI: The lack of correlation between these personal background factors and certificate acquisition suggests that personal background does not directly influence educational outcomes. This finding supports the principle of educational equity, affirming that all students, regardless of personal background, have the potential to succeed academically.

In summary, these results underscore the importance of focusing on both academic performance and student engagement in the learning process to improve educational outcomes. At the same time, they affirm the principle of educational equity, emphasizing that all students have the potential to succeed.

6 Conclusion and Future Work

This study, leveraging 17 courses offered by Harvard University and Massachusetts Institute of Technology via the edX platform during the 2012–2013 academic year, conducts in-depth analysis and mining of behavioral data from over 300 000 learners. To further investigate the relationship between learning behavior and certificate acquisition, we selected typical behavioral features and applied the spiking LSTM model to predict learning outcomes. Experimental results demonstrate that spiking LSTM, based on input features (including learner background and behavioral data such as educational background, age, gender, learning time, nevents, nchapters, nplay_video, ndays_act, and nforum_posts), can predict whether students will obtain certificates. The accuracy of SNN in predicting whether learners will ultimately acquire certificates reaches 99.8%, validating the superiority of this method compared to existing prediction methods. Additionally, we further analyzed the relationship between online learners' behavioral characteristics and certificate acquisition. The research found a strong correlation between students' scores and certificate acquisition, which is primarily influenced by active participation and is related to learners' nationality and age. This study also verifies the practical value of pulse neural networks in online education.

However, the prediction and analysis of online learning behavior possess great potential in the field of education, but there are also challenges that need to be overcome. Firstly, the accuracy of learning behavior analysis and model prediction depends on the quality and quantity of data, which may require large-scale online learning environments and detailed learner behavior data. However, data might contain noise, missing values, or inconsistencies. Furthermore, some key learning behaviors, such as students' cognitive strategies or emotional states, may be difficult to directly obtain from the data. Secondly, the subjective factors of learners have not been fully considered. Although learning behavior data provide objective information, learners' course completion and certificate acquisition are also influenced by subjective factors such as course participation goals, desire for certificate acquisition, satisfaction with course content, social network participation skills, environmental influences, and learners' motivation, purpose, and emotions.

Lastly, collecting and utilizing learning behavior data may raise privacy and ethical issues. For instance, students may not want their learning behavior to be continuously monitored or used for decisions that may affect them, thus appropriate data management and protection measures are needed. Additionally, predictive models could be misused, for example, if the model predicts that a student may drop out, the school might choose to abandon the student instead of providing additional support. Finally, the model prediction results need to be presented to teachers and students in a manner that is easy to understand and use, which may require additional design and development work. In summary, future research needs to continue exploring how to better utilize analytical and predictive models to improve educational outcomes, while also paying attention to potential problems that may arise and finding appropriate solutions.

References

[1] M. H. Cho and D. Shen, Self-regulation in online learning,

Distance Educ., vol. 34, no. 3, pp. 290–301, 2013.

[2] S. Liu, T. Hu, H. Chai, Z. Su, and X. Peng, Learners' interaction patterns in asynchronous online discussions: An integration of the social and cognitive interactions, *Br. J. Educ. Technol.*, vol. 53, no. 1, pp. 23–40, 2022.

[3] J. Goopio and C. Cheung, The MOOC dropout phenomenon and retention strategies, *J. Teach. Travel. Tour.*, vol. 21, no. 2, pp. 177–197, 2021.

[4] R. M. Latta, Notes in brief, *Environ. Psychol. Nonverbal Behav.*, vol. 2, no. 4, pp. 250–252, 1978.

[5] Q. Fu, Z. Gao, J. Zhou, and Y. Zheng, CLSA: A novel deep learning model for MOOC dropout prediction, *Comput. Electr. Eng.*, vol. 94, p. 107315, 2021.

[6] M. Badali, J. Hatami, S. K. Banihashem, E. Rahimi, O. Noroozi, and Z. Eslami, The role of motivation in MOOCs' retention rates: A systematic literature review, *Res. Pract. Technol. Enhanc. Learn.*, vol. 17, no. 1, p. 5, 2022.

[7] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, Predicting MOOC dropout over weeks using machine learning methods, in *Proc. EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, Doha, Qatar, 2014, pp. 60–65.

[8] W. Xing and D. Du, Dropout prediction in MOOCs: Using deep learning for personalized intervention, *J. Educ. Comput. Res.*, vol. 57, no. 3, pp. 547–570, 2019.

[9] J. Zhang, M. Gao, and J. Zhang, The learning behaviours of dropouts in MOOCs: A collective attention network perspective, *Comput. Educ.*, vol. 167, p. 104189, 2021.

[10] S. Lai, Y. Zhao, and Y. Yang, Broad learning system for predicting student dropout in massive open online courses, in *Proc. 2020 8th Int. Conf. Information and Education Technology*, Okayama, Japan, 2020, pp. 12–17.

[11] W. Zhang, X. Huang, S. Wang, J. Shu, H. Liu, and H. Chen, Student performance prediction via online learning behavior analytics, in *Proc. Int. Symp. on Educational Technology* (*ISET*), Hong Kong, China, 2017, pp. 153–157.

[12] T. Y. Yang, C. G. Brinton, C. Joe-Wong, and M. Chiang, Behaviour-based grade prediction for MOOCs via time series neural networks, *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 5, pp. 716–728, 2017.

[13] W. Zhang, Y. Zhou, and B. Yi, An interpretable online learner's performance prediction model based on learning analytics, in *Proc. 2019 11th Int. Conf. Education Technology and Computers*, Amsterdam, the Netherlands, 2019, pp. 148–154.

[14] Y. Wen, Y. Tian, B. Wen, Q. Zhou, G. Cai, and S. Liu, Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs, *Tsinghua Science and Technology*, vol. 25 , no. 3 , pp. 336–347, 2020.

[15] H. B. Shapiro, C. H. Lee, N. E. Wyman Roth, K. Li, M. Çetinkaya-Rundel, and D. A. Canelas, Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers, *Comput. Educ.*, vol. 110, pp. 35–50, 2017.

[16] A. Ramesh, D. Goldwasser, B. Huang, H. Daumé III, and L. Getoor, Modeling learner engagement in MOOCs using probabilistic soft logic, https://users.soe.ucsc.edu/~getoor/Papers/ramesh-nipsws13.pdf, 2013.

[17] J. Deboer, D. T. Seaton, and L. Breslow, Diversity in MOOCS students' backgrounds and behaviors in relationship to performance in 6.002x, in *Proc. 6th MIT's Learn. Int. Networks Consort. Conf.* (*LINC*), Cambridge, MA, USA, 2013, pp. 1–10.

[18] S. Crossley, L. Paquette, M. Dascalu, D. S. Mcnamara, and R. S. Baker, Combining click-Stream data with NLP tools to better understand MOOC completion, in *Proc. 6th Int. Conf. Learning Analytics & Knowledge*, Edinburgh, UK, 2016, pp. 6–14.

[19] S. Jiang, A. E. Williams, K. Schenke, M. Warschauer, and D. O. Dowd, Predicting MOOC performance with week 1 behaviour, in *Proc. 7th Int. Conf. Educational Data Mining* (*EDM*), London, UK, 2014, pp. 273–275.

[20] B. Xu and D. Yang, Motivation classification and grade prediction for MOOCs learners, *Comput. Intell. Neurosci.*, vol. 2016, p. 4, 2016.

[21] A. L. Rezaabad and S. Vishwanath, Long short-term memory spiking networks and their applications, in *Proc. Int. Conf. Neuromorphic Systems 2020*, Oak Ridge, TN, USA, 2020, pp. 1–9.

[22] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, LSTM: A search space odyssey, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2017.

[23] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, Qatar, 2014, pp. 1724–1734.

[24] H. Huang, Z. H. Yang, Z. W. Gu, M. Luo, and L. Xu, Decision tree model for predicting the overall survival of patients with diffused large B-cell lymphoma in the central nervous system, *World Neurosurg.*, vol. 166 , pp. e189–e198, 2022.

[25] L. V. Utkin and A. V. Konstantinov, Attention-based random forest and contamination model, *Neural Netw.*, vol. 154, pp. 346–359, 2022.

Fukun Chen received the BS degree from Liaoning University, China in 2015. He is currently pursuing the PhD degree in software engineering at the School of Computer Science and Technology, Harbin Engineering University, China. His research interests include recommendation algorithms, deep learning, and intelligent information processing.

Qiang Fu received the PhD degree in software engineering from Harbin Engineering University, Harbin, China in 2023. He is currently working at the School of Electronic and Information Engineering, Guangxi Normal University, Guilin, China. His research interest includes brain-inspired computing and intelligent information processing.

Bingxu Zhao is currently pursuing the PhD degree at the School of Computer Science and Technology, Harbin Engineering University, Harbin, China. His research interests are task allocation in spatiotemporal crowdsourcing, mobile computing, and artificial intelligence.

Xiaowei Wang received the BS degree in computer science and technology from Jining Medical College, China in 2014, and the PhD degree in software engineering from Harbin Engineering University, China in 2023. Since 2023, she has been at the School of Cyberspace Security, Shandong University of Political Science and Law, China. Her research interests include predictive algorithms, privacy protection, and machine learning.

Yanjing Li received the master degree in higher education from Heilongjiang University, Harbin, China in 2019. Her research interest includes educational data mining and principles of pedagogy.