2024 International Conference on Networking and Network Applications (NaNA) | 979-8-3503-7677-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/NaNA63151.2024.00067

Research on Big Data Analysis of Online Learning Behavior and Predictive of Learning Effectiveness

1st Lina Zhang
School of Computer
Baoji University of Arts and Sciences
Baoji, China
bjwllina@163.com

2nd Xueya Zhang
School of Computer
Baoji University of Arts and Sciences
Baoji, China
765808167@qq.com

3rd Xinjie Li
School of Computer
Baoji University of Arts and Sciences
Baoji, China
2913630168@qq.com

4th Kangrui Ma
School of Computer
Baoji University of Arts and Sciences
Baoji, China
2212763389@qq.com

***Abstract*—Data-driven artificial intelligence is transforming industries across the board, with AI combined with education big data shaping the trajectory of online education platforms. The project integrates big data analytics and deep learning technologies into online education, leveraging data analysis and visualization to offer personalized services to students, teachers, and educational institutions. This breaks geographical barriers, enabling universal access to high-quality educational resources.The platform empowers educational institutions to accurately grasp the needs of both students and teachers, optimize teaching resources, and devise tailored teaching and learning strategies. It mines data on online learning behaviors, constructs student profiles, and precisely recommends learning materials and customized learning approaches. This approach ensures that education is not only accessible but also adaptive and effective for all learners, thereby revolutionizing the way education is delivered and experienced in the digital age.**

***Index Terms*—data analytics, LSTM, big data, online learning, prediction**

I. Introduction

Under the rapid promotion of global science and technology, especially the wave of popularization of the Internet and mobile technology, digital education has gradually emerged as an important trend leading the development of global education [1]. In this new era of informationization and intelligence, the development of China's Moocs is particularly strong. According to the latest data, as of April 5, 2024, more than 76,800 courses have been put online on China's catechism, with a total of 454 million registered users, successfully serving the learning needs of 1.277 billion people nationwide [2]. This set of figures not only highlights the booming development of China's digital education, but also deeply reflects the public's eager expectation and strong demand for online education [3] [4].

In order to cope with these challenges, an online education big data analytics platform has been developed. The dataset selected for this paper is provided by Digital Rex, which is based on the practice dataset of Teddy's cloud classroom platform, including course chapters, course tasks, class members, class courses, log tables, user learning statistics, learning logs, test paper information, test paper results and other datasets, with its data volume of about 2,000w, and its content includes user ID, course ID, test scores and so on.

The project has designed a website that can evaluate the effectiveness of the course and analyze the quality of the video to provide feedback on the quality of the course for the instructor to further improve the teaching effect. The system analyzes user behavior and obtains information on user learning to help classroom teachers keep track of their students' learning in real time. It can predict the learning outcomes of each user and customize personalized learning paths for each individual, giving teachers suggestions for adjusting teaching strategies and educational resources.

II. Releted works

A. Data analytics technique

The scientific operation of transforming data into knowledge for making informed decisions is known as analytics. Data analytics is one of the main tools for obtaining valuable results from this data and is a technique that combines statistics, mathematics, specific application areas and other disciplines [5]. The rapid growth in the field of big data and analytics over the last decade has led to a growing interest in data analytics in education [6] [7].

Baoji Science and Technology Plan Project 2018JH-23

979-8-3503-7677-7/24/$31.00 ©2024 IEEE
DOI 10.1109/NaNA63151.2024.00067

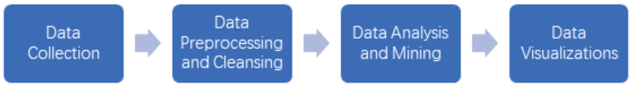


Fig. 1. Data analysis steps.

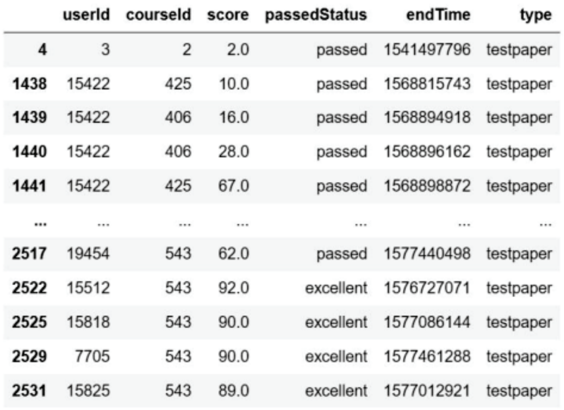
Several researchers have reviewed and analyzed the characteristics and applicability of big data and analytics techniques in the field of education.Andy Nguyen et al. outlined the current state of research and theoretical perspectives on data analytics in education.The purpose of data analytics techniques is to accurately interpret the data and to make precise inferences. In this context, data analytics techniques aim to accurately interpret data and make precise inferences to extract insightful information to optimize creative online learning [8]. In addition, data analytics technology enables instructors to identify risk factors associated with learner engagement and optimize the design of the learning environment [9] [10]. In short, data analytics technology is an umbrella term for any technology that deals with large amounts of data (e.g., text, audio, and video), including data capture, transmission, storage, organization, searching, analysis, and profiling. The main flow of this project is shown in Fig. 1.

B. RNN and LSTM

RNN (Recurrent Neural Network) is a special kind of neural network structure that is suitable for processing data with a time-series nature, such as natural language, audio signals, and so on. In traditional feed-forward neural networks, data is input in a static, unordered manner, while a RNN is able to deal with sequential dependencies between input data thanks to its recurrent structure - the network processes each element in the sequence and passes on the results. The RNN is able to handle sequential dependencies between the input data thanks to its recurrent structure - the network processes each element of the sequence and passes the result to the next time step, which allows the model to capture temporal dependencies in the data sequence [11].

Innovative applications of RNN in big data analytics further expand its potential in high-dimensional data processing, which is not only limited to dimensionality reduction and feature learning, but also delves into deeper applications of dimensionality reduction and feature learning for high-dimensional data, multimodal data analysis, anomaly detection and event prediction, and interactive learning and personalized recommendation [12].

LSTM (Long Short-Term Memory) is a special type of RNN, proposed by Hochreiter & Schmidhuber in 1997 [13], specifically designed to solve the long-term dependency problem. In standard RNNs, it is difficult for the model to learn the dependencies between inputs and outputs that are far away from each other due to the problem of gradient vanishing or gradient explosion.LSTM overcomes this problem by introducing a cell state and a series of control gate mechanisms, which include input gates, forgetting gates, and output gates, which are capable of selectively forgetting, storing, or exporting information. This design allows LSTMs to theoretically retain information over a very long range of time, thus excelling in a variety of sequence modeling tasks such as machine translation, speech recognition, text generation, and so on.



	userId	courseId	score	passedStatus	endTime	type
4	3	2	2.0	passed	1541497796	testpaper
1438	15422	425	10.0	passed	1568815743	testpaper
1439	15422	406	16.0	passed	1568894918	testpaper
1440	15422	406	28.0	passed	1568896162	testpaper
1441	15422	425	67.0	passed	1568898872	testpaper
...	...	...	...	...	...	...
2517	19454	543	62.0	passed	1577440498	testpaper
2522	15512	543	92.0	excellent	1576727071	testpaper
2525	15818	543	90.0	excellent	1577086144	testpaper
2529	7705	543	90.0	excellent	1577461288	testpaper
2531	15825	543	89.0	excellent	1577012921	testpaper

Fig. 2. Some raw data.

III. Data analysis

A. Data proprocessing

Before data analysis, the dataset needs to be preprocessed, which consists of the following steps.

In the first step, two DataFrames, merged_data_1 and data_3, are in-joined based on the userId column using the merge function of pandas. The result will be saved in a new DataFrame named merged_data.

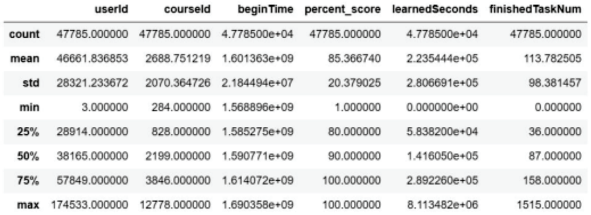
In the second step, select specific columns from merged_data: userId, courseId, beginTime, percent_score, learnedSeconds, and finishedTaskNum. once these columns are selected, we use them to create a new merged_data DataFrame to replace the original merged_data.

In the third step, to gain more insight into the characteristics of these numeric columns, we execute the describe() method on the numeric columns in the new merged_data. This generates and displays descriptive statistics for these numeric columns, including mean, standard deviation, minimum, quartiles, and maximum, saved as a CSV file.

In the last step, the numeric columns (such as learnedSeconds and finishedTaskNum) are normalized,the raw data Fig.2 and dataset after processed is shown in Fig.3.

B. Models

By exploring and analyzing students' performance in previous exams, a complete learning profile can be built up. This file not only records the students' performance data, but also includes the students' performance in different subjects, different types of questions, as well as learning attitudes and habits and other aspects of information.Based on this information, our system is able to accurately predict a student's grade in the



	userId	courseId	beginTime	percent_score	learnedSeconds	finishedTaskNum
count	47785.000000	47785.000000	4.778500e+04	47785.000000	4.778500e+04	47785.000000
mean	46661.836853	2688.751219	1.601363e+09	85.366740	2.235444e+05	113.782505
std	28321.233672	2070.364726	2.184494e+07	20.379025	2.806691e+05	98.381457
min	3.000000	284.000000	1.568896e+09	1.000000	0.000000e+00	0.000000
25%	28914.000000	828.000000	1.585275e+09	80.000000	5.838200e+04	36.000000
50%	38165.000000	2199.000000	1.590771e+09	90.000000	1.416050e+05	87.000000
75%	57849.000000	3846.000000	1.614072e+09	100.000000	2.892260e+05	158.000000
max	174533.000000	12778.000000	1.690358e+09	100.000000	8.113482e+06	1515.000000

Fig. 3. Post-preprocessed data.

next exam. The prediction results are not only highly accurate, but also provide students with targeted learning suggestions.

A recurrent neural network model is defined and an LSTM layer is introduced. The model is designed to process the user's learning sequences and predict the corresponding labels. In the initialization phase of the model, we used a sequential model, which allowed us to add multi-layer networks to the model via the add method.

Then an LSTM layer, containing 64 neurons, was added and the layer was set to return only the output of the last timestep, since we are mainly interested in the final result of the whole sequence. In addition, a Dropout layer is added, which is a regularization technique that randomly ignores the output of 20% of the neurons during training, helping to prevent model overfitting and enhancing the generalization of the model.

The last fully connected layer (Dense Layer) has the number of output units matched to the number of exam levels, i.e., each prediction will output 3 results. The activation function of this layer uses softmax, which is able to convert the raw outputs into a probability distribution, ensuring that all output values sum to 1, and thus deriving the final prediction based on these probabilities.

IV. Results

A. Analysis of Student Behavior

The module focuses on students' behavioral trajectories and habits in the system, and through data analysis, it digs out users' learning preferences and needs, providing data support for personalized recommendations and precise marketing [14]. This not only improves students' learning experience, but also helps the platform better understand user needs and optimize service strategies.

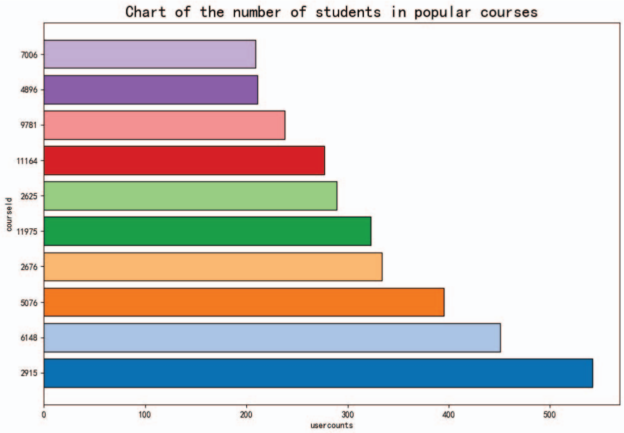


Fig. 4. Number of students in popular courses.

First of all, the number of student who have chosen the course is counted, and the selected data are courseId, userId and passedStatus data in the testpaper_result.csv file. Observing the data, it can be concluded that there are a lot of the same data in courseId and userId, so we need to carry out de-duplication operation, and then we use the drop_duplicates function to de-duplicate the userId and courseId. duplicates function to userId and courseId for de-duplication, and then we use the groupby function to courseId as the basis for userId counting statistics, and let the number of people to choose the number of courses from the largest to the smallest for sorting. Due to the large amount of data will be the course selection number of the top five courses for visual display, as shown in Fig. 4.

B. Analysis for Teacher

The rate of student progress in online learning is one of the most important topics of concern for teachers and parents. In this sub-module, the platform shows the stage-by-stage progress rate of students for teachers as one of the indicators for teachers to adjust their teaching methods.

The IDs of the five most popular courses are extracted and then the values in the courseId column are filtered using the is in method and then the rows with the IDs of the filtered values are filtered, the number of occurrences of each unique value in the column 'A' is calculated to find out the unique values with the number of occurrences of 1, the rows containing these unique values are deleted and then the indexes are reset and the original indexes are kept as the new columns.

After that we reset the index and keep the original index as a new column, then we get the progress rate of the course by using if function, if function conditions are mainly used to calculate the number of times the same user has progressed in two consecutive exams or tests in a particular course. Specific conditions are as follows: the first to ensure that the comparison is the same user's records. The second ensures that the comparison is between two consecutive examinations or tests in chronological order. Third determine if the second score is greater than or equal to the first score to determine if progress has been made. And we visualize the results as shown in Fig. 5 .

To display a student's overall assessment on the teacher's terminal, we will display the number of tasks reported by the user, the number of completed tasks, the task completion rate, the total number of logins, the total number of learning hours, which will help the teacher to get an overall picture of the student's learning in Fig.6.

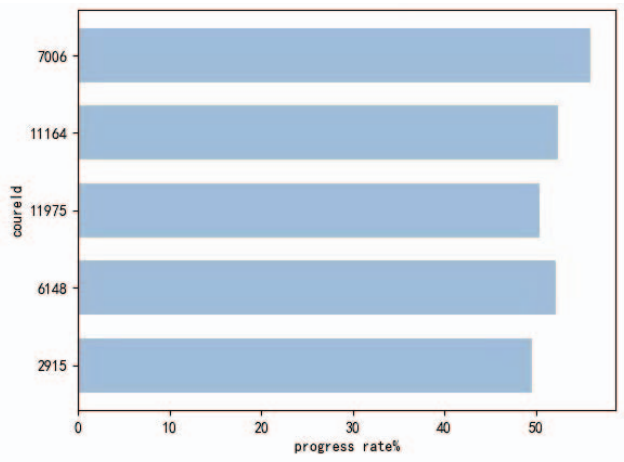


Fig. 5. Progress rate of students.

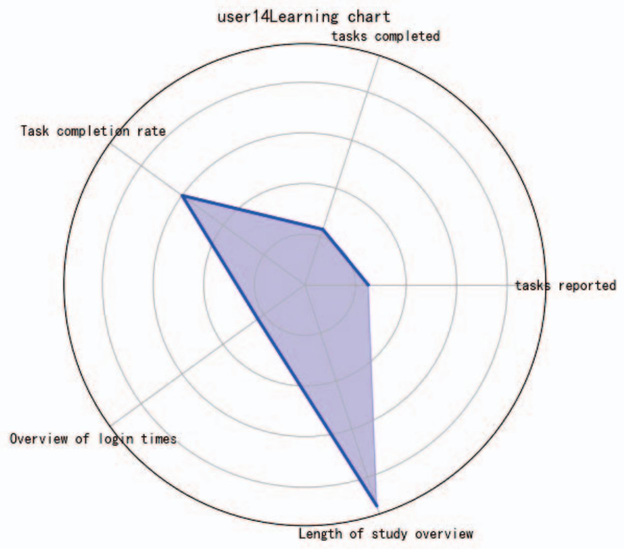


Fig. 6. Evaluation of students.

C. Evaluation

It is also important for platform administrators to know how to assess the effectiveness of teachers' teaching through a number of metrics. The system assesses the popularity of teachers by the average number of people in the courses they teach (NormalizedPopularity in the table), the competence level of teachers by the average number of high grades in the courses they teach (NormalizedAbility in the table), and the conscientiousness of teachers by the number of rubrics such as assignments in the courses they have taught (Normalized-Dedication in the table) as shown in Fig.7.

D. Behavior prediction

Once the LSTM model was constructed, Adam foction was chosen as the optimizer, which is an adaptive learning rate method capable of dynamically adjusting the learning rate for each parameter based on the first-order moment estimate (mean) and second-order moment estimate (for the centered variance) of the gradient. Given that the labels have been transformed into a solo thermal coding format, categorical_crossentropy is used as the loss function, i.e., the cross-entropy loss, which is able to measure the difference between the probability distribution predicted by the model and the true probability distribution.

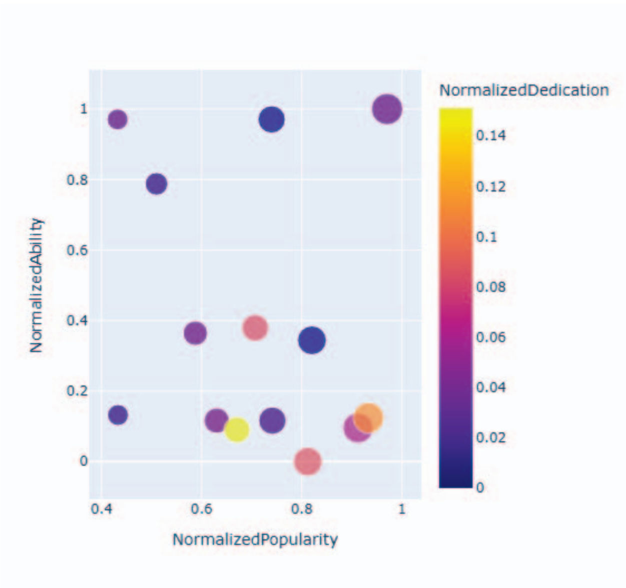
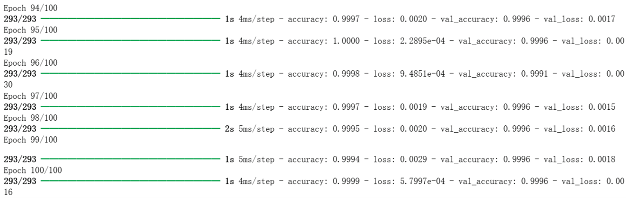


Fig. 7. Evaluation of teachers.



```
Epoch 94/100
293/293 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - accuracy: 0.9997 - loss: 0.0020 - val_accuracy: 0.9996 - val_loss: 0.0017
Epoch 95/100
293/293 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - accuracy: 1.0000 - loss: 2.2895e-04 - val_accuracy: 0.9996 - val_loss: 0.00
19
Epoch 96/100
293/293 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - accuracy: 0.9998 - loss: 9.4851e-04 - val_accuracy: 0.9991 - val_loss: 0.00
30
Epoch 97/100
293/293 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - accuracy: 0.9997 - loss: 0.0019 - val_accuracy: 0.9996 - val_loss: 0.0015
Epoch 98/100
293/293 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.9995 - loss: 0.0020 - val_accuracy: 0.9996 - val_loss: 0.0016
Epoch 99/100

293/293 ━━━━━━━━━━━━━━━━━━━━ 1s 5ms/step - accuracy: 0.9994 - loss: 0.0029 - val_accuracy: 0.9996 - val_loss: 0.0018
Epoch 100/100
293/293 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - accuracy: 0.9999 - loss: 5.7997e-04 - val_accuracy: 0.9996 - val_loss: 0.00
16
```

Fig. 8. Results.

The function grade_score is defined, which is used to convert the accepted user exam scores into grades according to the score bands, and the user's exam scores are converted into grade ABC.The extracted user learning sequences were populated and the user's labels were converted into classification format.

The dataset was partitioned where 80% was used as a training set for model training and the remaining 20% was used as a test set for evaluating the performance of the model. By iteratively adjusting the network structure, optimizing the parameter settings, and introducing the softmax activation function.

Accuracy was chosen as the metric to assess the performance of the model on the validation set, which reflects the number of correctly predicted samples as a proportion of the total number of samples and provides us with an intuitive basis for performance assessment,as shown in Fig.8. As can be seen, the accuracy is higher than 95%, but there may be a risk of overfitting.

V. Conclusion

The analysis, design and development of educational data will positively impact all players in the education sector in a number of ways. These advanced analytical tools can empower students to learn on their own, improve academic performance, increase the productivity of teachers, and provide decision-making assistance to institutional administrators [15]. After validation and testing, the model proposed in this paper shows better performance in processing users' learning sequences and predicting the corresponding labels, which not only improves the prediction accuracy, but also reduces the risk of overfitting, and provides a basis for subsequent analysis of other online education data.

Next, we will explore more application scenarios. At present, our system mainly analyzes based on existing learning data, but with the development of technology and the abundance of data, we can try to apply the system to a wider range of subject areas to meet the needs of different users. We plan to strengthen the construction of the user feedback mechanism. User feedback is an important basis for improving the system functions and enhancing the user experience. We will establish a better user feedback channel, actively collect and analyze user data and feedback, and identify problems in time for improvement.

Acknowledgment

This work was supported by the Baoji Science and Technology Plan Project 2018JH-23.

References

[1] A. Alam, "Platform utilising blockchain technology for elearning and online education for open sharing of academic proficiency and progress records," in *Smart Data Intelligence: Proceedings of ICSMDI 2022*. Springer, 2022, pp. 307–320.

[2] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," *Computers and Electrical Engineering*, vol. 66, pp. 541–556, 2 2018.

[3] J. Sahni, "Is learning analytics the future of online education? assessing student engagement and academic performance in the online learning environment," *International Journal of Emerging Technologies in Learning*, vol. 18, pp. 33–49, 2023.

[4] A. I. M. Elfeky and M. Y. H. Elbyaly, "The use of data analytics technique in learning management system to develop fashion design skills and technology acceptance," *Interactive Learning Environments*, vol. 31, pp. 3810–3827, 2023.

[5] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 10, no. 3, p. e1355, 2020.

[6] M. Paneque, M. del Mar Roldán-García, and J. García-Nieto, "e-lion: Data integration semantic model to enhance predictive analytics in e-learning," *Expert Systems with Applications*, vol. 213, 3 2023.

[7] H. Luan, P. Geczy, H. Lai, J. Gobert, S. J. Yang, H. Ogata, J. Baltes, R. Guerra, P. Li, and C. C. Tsai, "Challenges and future directions of big data and artificial intelligence in education," 10 2020.

[8] M. A. Javed, S. Zeadally, and E. B. Hamida, "Data analytics for cooperative intelligent transport systems," *Vehicular communications*, vol. 15, pp. 63–72, 2019.

[9] B. K. Daniel, "Big data and data science: A critical review of issues for educational research," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 101–113, 2019.

[10] J. Lodge and L. Corrin, "What data and analytics can and do say about effective learning. npj science of learning, v. 2, n. 1," 2017.

[11] M. E. Dogan, T. Goru Dogan, and A. Bozkurt, "The use of artificial intelligence (ai) in online learning and distance education processes: A systematic review of empirical studies," *Applied Sciences*, vol. 13, no. 5, p. 3056, 2023.

[12] S. Caspari-Sadeghi, "Learning assessment in the age of big data: Learning analytics in higher education," *Cogent Education*, vol. 10, no. 1, p. 2162697, 2023.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] M. Maqableh and M. Alia, "Evaluation online learning of undergraduate students under lockdown amidst covid-19 pandemic: The online learning experience and students' satisfaction," *Children and youth services review*, vol. 128, p. 106160, 2021.

[15] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao, "Online learning: A comprehensive survey," *Neurocomputing*, vol. 459, pp. 249–289, 2021.