ARTICLE

Check for updates

Driving STEM learning effectiveness: dropout prediction and intervention in MOOCs based on one novel behavioral data analysis approach

Xiaona Xia [1,2 ✉] & Wanxue Qi[1,2]

With the full application of MOOCs online learning, STEM multidisciplinary and knowledge structures have been achieved, but it has also resulted in a massive number of dropouts, seriously affected the learning sustainability of STEM education concepts, and made it difficult to achieve learning effectiveness. Based on the massive STEM learning behavior instances generated by MOOCs, as well as the entire learning periods, this study considers some key explicit and implicit features associated with learning behavior, and achieves the fusion of convolutional neural network and recurrent neural network through data-driven approaches, incorporates long short-term memory mechanism to develop dropout prediction methods and models. Based on the experimental results, we also discuss the relevant problems of dropouts related to STEM learning behavior, explore the key dropout temporal sequences of the learning process, identify related factors that have key impacts on learning behavior, and deduce intervention measures and early warning suggestions. The entire study can provide effective methods and decisions for researching the STEM learning behavior of MOOCs and has strong research feasibility and urgency.

[1] Faculty of Education, Qufu Normal University, 273165 Qufu, Shandong, China. [2] Chinese Academy of Education Big Data, Qufu Normal University, 273165 Qufu, Shandong, China. ✉email: xiaxn@sina.com; xiaxn@qfnu.edu.cn

Introduction

STEM is a broad term covering a range of interdisciplinary disciplines and skills that integrate science, technology, engineering, and mathematics (Flegr et al., 2023). Its objective is to help learners develop professional knowledge and abilities (Buckley et al., 2023), foster their conscious awareness of participating in interdisciplinary research, enhance their ability to discover new information, encourage critical thinking, promote theory and practice in related fields, improve scientific technology, and enhance talent competitiveness and adaptability (Evenhouse et al., 2023). Many countries have recognized and implemented STEM education, which has facilitated the development and application of interdisciplinary studies while gradually integrating STEM theories into various educational processes (Weston et al., 2023). MOOCs have accelerated the popularization and development of STEM education, enabling many learners to gain access to STEM learning opportunities and acquire essential learning behavior and effectiveness. Moreover, the abundant learning resources of MOOCs provide technical support for recommending and guiding interdisciplinary STEM knowledge, which makes MOOCs an essential mode of the online learning process (Dash et al., 2022).

MOOC is a newly emerging online course development model that originated from publishing resources, learning management systems, and integrating learning processes with more open network resources (Chen et al., 2021). MOOCs are large-scale online open courses that are distributed on the internet and organized by learners with a spirit of sharing and collaboration to enhance knowledge propagation. MOOCs and online courses have similarities in that they both achieve online education, forming a new, interactive, and open learning process. They give new meanings to teaching and learning, bring profound changes and impacts to education, and promote the optimization and updating of educational concepts, educational ideas, educational models, and teaching methods. But they also have differences. Firstly, learners who participate in online courses have little interaction with others. The teachers and peers in class have nothing to do with themselves. MOOCs allow learners to interact with teachers and peers and even form a learning social network. Secondly, the videos of online courses are usually quite long, and most of them record the complete classroom teaching process of the teacher, which is not conducive to the participation of learners with limited time. The duration of MOOC courses is relatively short, usually with a complete knowledge point as the teaching sub-objective, which is conducive to learners using fragmented time sufficiently. Thirdly, there is no requirement for the class time of online courses, as long as learners are willing to learn, they can study the course at any time; MOOCs have time requirements, with each course having a specific start time and corresponding periods. Fourthly, online courses do not have serious requirements for homework and exams, nor do they have related quizzes, which greatly require learners' self-control. But MOOCs are completely different, learners need to complete homework and projects, and participate in the verification of course learning, reflecting the phased learning quality through quizzes (Primario et al., 2022).

Therefore, for STEM education with obvious demands for interactivity and collaboration, Meanwhile, there are correlations and limitations between the knowledge of STEM courses, MOOCs are more suitable for providing effective support for their learning process and learning resources. STEM also needs to deploy relevant assessment methods to timely test learning effectiveness, identify problems in the learning process, and provide learners with appropriate guidance and intervention (Ramadhan et al., 2023). MOOCs provide valuable interdisciplinary learning resources and interactive online modes that are used to support personalized STEM learning behavior, but are still struggling with high dropout rates among learners (Rahimi, 2023). While MOOCs have expanded access to quality education and promoted equity and diversity in education (Khor and Darshan, 2022), the pedagogy and objectives of STEM education have not been effectively implemented through MOOCs. The associated learning resources fail to consider the relationships between disciplines and knowledge graphs during production and recommendation (Xia and Qi, 2022; Mourdi et al., 2023), and learning processes rely primarily on learners' needs (Cristea et al., 2023). However, if these needs are not achieved through the exploratory and effective tracking and analysis of learning behavior, learners may encounter some problems that remain unresolved, leading to negative emotions and eventual dropout (Anttila et al., 2023; Xia and Qi, 2023a). The integration of STEM and MOOCs has also formed some tracking methods and evaluation methods, and some researchers have provided some theoretical and applied results in this area of research (Primario et al., 2022).

In addition, learning behavior has a ripple effect and affects course selection if a learning course has low participation or a high dropout rate (Labrovic et al., 2023; Gupta et al., 2022). This is directly related to the organization of online learning content and the video quality of courses (Chu et al., 2022). Firstly, videos of courses rely heavily on teachers and cannot track learners' real-time progress, making it challenging to accommodate too many individual learning needs (Jansen et al., 2022). Secondly, learners cannot provide direct feedback to teachers during the learning process. When problems encountered by learners cannot be effectively addressed or communicated, it hinders their learning experiences (Santos et al., 2023). The source of high dropout rates have already explored in business (Ortiz-Lozano et al., 2023), applied linguistics (Cervantes-Soon et al., 2017), psychology (Boumparis et al., 2023), management (de Oliveira Marques et al., 2023), and literature (Alves et al., 2023), etc., resulting in low participation in corresponding courses, low learning enthusiasm and interest among learners, and a high risk of frustration. These courses are not overly relevant in terms of knowledge correlations, learners do not need to associate too many precursor courses when learning a new one, and they can solve many problems by exerting their own learning enthusiasm and exploration. However, there are still a large number of learners who have not completed the complete learning of a course and give up halfway. For STEM courses with strong knowledge correlations and conditional constraints, as well as the demand for strong interaction and collaboration in application practice, the potential high dropout rate is more worthy of attention. Its research presents greater difficulty and will also put forward higher requirements and goals for research methods (Gallagher and Lamb, 2023). Therefore, the dropout problems cannot be ignored for the STEM learning process in MOOCs. Due to the unique demands of STEM education, the reasons for dropout are bound to have their characteristics, making effective MOOC dropout prediction critical for STEM.

This study defines STEM-related courses and relationships, collects complete learning behavior instances from MOOCs, and explores dropout prediction methods and tracking processes for STEM learning behavior. It utilizes data-driven model design and behavior pattern analysis to locate the temporal sequences and key factors of dropout. Furthermore, based on the propagation theory of learning behavior (Xia and Qi, 2023b), this study proposes potential dropout risks and intervention measures for new STEM courses and learners, with the aim of providing key strategies for benign guidance.

Related work

Researchers have begun to explore the advantages of MOOCs for institutions, teachers, learners, and the learning process. MOOCs have a vast amount of learning resources and service modes, as well as a certain tracking mechanism, but this may not be applicable to all learners (Benoit et al., 2024). The resources, services, interactions, collaborations, and application practices that learners need may not necessarily be met by MOOCs, or when learners have misunderstood their intentions, interests, and preferences, and their learning goals cannot be effectively addressed through the existing resources and services of MOOCs (Taranto et al., 2021; Ramadhan et al., 2023). The learning experiences might be relatively negative, and even drop out directly. Although researchers have found that the resources and services of MOOCs depend more on resource providers and service implementers, resource sharing and service follow-up without or with little consideration for learners' own learning backgrounds and interests are inefficient or ineffective. But the key value of MOOCs is to provide high-quality education for those who cannot access it (Taranto et al., 2021). However, how to define "high-quality" open online education remains to be determined. This poses many learning risks for STEM education that uses MOOCs as the learning environment (Dhiman et al., 2023). Learners are unable to construct positive and effective learning behaviors for a STEM course in a short period, and they have psychologically abandoned the entire learning process before discovering their interests and problems. This is very bad.

An important teaching consideration for MOOCs is that the presence of teachers cannot be ignored (Mandari et al., 2023). Due to the fact that enrollment can easily reach thousands, the relationships between teachers and learners in MOOCs are clearly different from traditional offline classrooms. Many teachers are prone to reflecting their dissatisfaction and even disgust with the communication and collaboration process with learners (Benoit et al., 2024). It is difficult for teachers to translate classroom teaching practice into tacit communication and collaborative win-win among a large number of learners (Primario et al., 2022). MOOCs have not promoted the construction of positive and effective interpersonal social relationships between teachers and learners, as well as between learners and peers. The learning process is easily reduced to individual behaviors, which is definitely an important factor that makes STEM education prone to dropout (Dhiman et al., 2023). Both course developers and teachers have the opportunity to reconsider the presentation of course resources in order to effectively support the learners and learning processes. It is extremely necessary to improve the new mechanisms and tasks of teacher roles in MOOCs, in order to achieve effective online classroom management in STEM (Martinez and Ellis, 2023). To provide teachers with more learning process data, enabling them to develop new strategies and timely deploy the relationships and constraints of STEM knowledge in order to enhance their influence in teaching. This requires a deeper understanding of the key value and effective strategies that a single teacher can use to bring more MOOCs to learners (Gallagher and Lamb, 2023).

In addition, the educational aims and scopes of MOOCs vary, and the different learning objectives, methods, and tasks of STEM also determine the different management methods and service scheduling modes (Alrajhi et al., 2023; Ramadhan et al., 2023). In recent years, improving STEM education has been identified as a major goal by organizations such as the National Academy of Engineering and the National Science and Technology Commission in the United States. MOOCs have not ignored the promotion of STEM education. A literature review found that although researchers have begun to explore the sustainable tracking, feedback, and decision of STEM courses in MOOCs,

sufficient analysis and prediction of massive STEM learning behavior instances have not been argued through key and effective research, evaluation, and practice (Mandari et al., 2023). The purpose of this study is to take the complete learning process of STEM in MOOCs as the research scenario, take massive learning behavior instances as the research object, complete sufficient correlation analysis and reliable modeling of the associated attributes, features, parameters, structures, and relationships, explore and predict the key potential factors that affect the occurrence of STEM dropout, and design innovative and feasible methods and techniques to derive effective conclusions and suggestions (Benoit et al., 2024).

The accumulation of STEM learning behavior instances continues through the opening and participation in related courses. However, during the initial stages of learning behavior formation, the scale of learning behavior instances is small, making it difficult to obtain accurate and reliable analysis results. The analysis process also requires high computing resources and has many limitations, resulting in significant outliers (Zhang et al., 2022). While long-term accumulated STEM learning behavior instances can produce better dropout prediction results in small-scale data analysis, it directly impacts course selection and recommendation and significantly reduces the effectiveness of late-stage intervention in the learning process (Mubarak et al., 2022). Therefore, a suitable temporal sequence is necessary to implement dropout prediction for STEM learning behavior, as it produces better predictive results and avoids predicting too late, which is a crucial issue (Mubarak et al., 2022).

After reviewing and sorting out relevant references (Cara et al., 2022; Zhu et al., 2022), it is found that the dropout prediction process of STEM learning behavior mainly focuses on two aspects:

Firstly, the complete learning behavior instances of a certain STEM course are directly used for modeling and prediction. As learners are still participating in the course, it is difficult to identify which learners have negative emotions. Therefore, unsupervised machine learning methods are needed to cluster learning behavior (Xia and Qi, 2023c), thereby clearly dividing learners into two categories: dropout and non-dropout. However, this aspect ignores the relationships between different courses, as well as the feasibility and possibility of learners constantly changing goals between these courses. It is limited to only analyzing the independent data of a single course (Mccarthy et al., 2021).

Secondly, modeling STEM learning behavior instances from the entire learning platform is used for dropout prediction of related or similar courses (Parviainen et al., 2020). As different courses on MOOCs have similar learning behavior patterns, they belong to the same STEM field and share common feature descriptions, the dropout prediction model can be directly applied to multiple courses. Currently, the dropout prediction models used in STEM learning behavior mainly focus on this aspect. However, the features and relationships analyzed during the dropout prediction process are affected by the data definition and rule constraints of MOOCs, which is needed to improve the analysis of implicit features and latent relationships (Xia and Wang, 2022). At the same time, the negative propagation of learners' dropout tendencies requires the dropout tracking process to consider the temporal sequences associated with the learning process (Guo et al., 2023; Xing et al., 2016).

Based on the second aspect, this study first mines the learning behavior instances associated with courses and interactive learning activities that have a large number of registrations and then thoroughly trains and tests the dropout prediction model. Secondly, the start time of the learning process is recorded, and the STEM learning behavior instances required for setting the

predictive indicators are analyzed by sequential data for predicting dropout, which provides a reference for the study and application of dropout prediction models for MOOCs learners. On this basis, we consider the modeling and association of STEM-related learning behavior instances in order to effectively drive the dropout prediction model to be applied in MOOCs. The model is trained using historical learning behavior instances to predict new or associated courses, obtain the more feasible learning behavior routes, and then provide relevant decision strategies and intervention measures to explore the relationships between learning behavior instances and predictive indicators.

Data standardization and problem description
To achieve the effective dropout prediction of STEM learning behavior, this study selects the massive learning behavior dataset from the Open University UK (Download address: https://analyse.kmi.open.ac.uk/open_dataset), which has been desensitized and standardized, and anonymized for relevant courses, learners, and learning periods etc. However, it can be confirmed that DDD, EEE, and FFF are three STEM courses which have been described in this dataset. The dataset includes complete learning behavior instances for the entire periods, covering four learning periods in total, which are labeled as 2013B, 2013J, 2014B and 2014J, respectively, "B" is for the presentation starting in February and "J" is for the presentation starting in October. In the data analysis process of this study, we directly used the labels of some attributes and features so that readers can directly correspond to the relevant dataset, which helps understand and apply the meaning and relationships of learning behavior instances. The dataset is large in scale and fully records learners' relevant registration information. Based on the key descriptive features of learning behavior, more complex relationships are constructed. Learning behavior descriptions include learners, courses, interactive learning activities and click rates. Learners and courses might be directly obtained from the dataset, while interactive learning activities need to be explored based on relevant conditions, and then corresponding click rates will be calculated. Before analyzing and calculating the corresponding data for interactive learning activities, we might ensure the classification conditions to achieve the data association.

The complete learning behavior instances of four learning periods for courses DDD, EEE, and FFF are obtained, and the abnormal values are analyzed and distinguished. As each course has various interactive learning activities for each learning period, the corresponding descriptive data has multiple characteristics. Although learners' online learning methods are relatively personalized, the types of interactive learning activities for the different learning periods are different. In the 2013J and 2014J periods, Content and Resources are the common interactive learning activities for learners. For differential interactive learning activities, 2013J focuses on the Forum and Homepage, while learners tend to participate in the Subpage and URL in 2014.

Learners' profiles and Learning backgrounds, to some extent, influence learning behavior, which has been proved in previous academic research (Mccarthy et al., 2021). Based on the composition of the dataset and learners' descriptive attributes, the factors influencing dropout behavior in three STEM courses need to consider learners' profiles and Learning backgrounds. As learners can generate abnormal values in multiple interactive learning activities, i.e., interactive learning activities mutually influence each other, the validation of STEM learning behavior will define potential variables and observation variables for the tested problems based on the analysis of learners' profiles, Learning background, and learning behavior. The entire research framework is defined in Table 1. Observation variables related to

Table 1 Related factors of STEM learning behavior.

Latent variable	Observation variable	Description
Demographic information	GD	Gender
	RN	Region
	IMD	IMD band
	AGE	Age band
	DY	Disability
Learning accumulation	NPA	Number of previous attempts
	SC	Studied credits
	HE	Highest education
Learning behavior	AT	Activity type
Assessment	FR	Final result
	AST	Assessment type

Table 2 Activity distribution of learning behavior.

Activity_type	Description	DDD		EEE		FFF	
DtP	Data plus					√	20
DlP	Dual pane			√	3	√	17
xQz	External quiz	√	26				
Fdr	Folder					√	2
FrD	Forum discussion	√	53	√	15	√	25
Gly	Glossary	√	5			√	4
HPg	Home page	√	4	√	3	√	4
Hty	HTML activity					√	4
Cbt	Collaborate	√	14	√	7	√	47
Ctt	Content	√	45	√	139	√	419
Lnt	Illuminate					√	16
Wki	Wiki	√	16	√	7	√	26
WPg	Web page	√	1	√	11	√	86
Qtr	Questionnaire					√	57
Quz	Quiz			√	6	√	44
Rty	Repeated activity					√	5
Rsc	Resource	√	706	√	103	√	454
SSP	Shared sub-page						
SPG	Sub-page	√	611	√	21	√	215
URL	Universal resource locator	√	223	√	21	√	522
Distribution scale		1708		336		1967	

Demographic Information, Learning Accumulation, and Assessment can be directly obtained from the dataset. The relationships between interactive learning activities and courses, as well as their learning periods, require flexible establishment of observation variables. Since Learning Behavior involves many types of interactive learning activities, the distribution of related interactive learning activities for the three courses is statistically analyzed. As shown in Table 2, we can see that the distribution of interactive learning activities is different. The participation scale of some interactive learning activities is very large, such as Content.

Learning behavior of three courses is divided into two types: dropout and non-dropout, respectively, to test whether Demographic Information, Learning Accumulation, and corresponding observation variables can affect the two types of Learning Behavior. Whether the selection of interactive learning activities can affect the participation of other activities, whether interactive learning activities can affect learners' Learning Behavior and Assessment, these problems will directly or indirectly affect learners' dropout trends and will inevitably affect the predicted results of STEM dropouts.

The prediction process of STEM dropouts requires first demonstrating and constructing the potential relationships
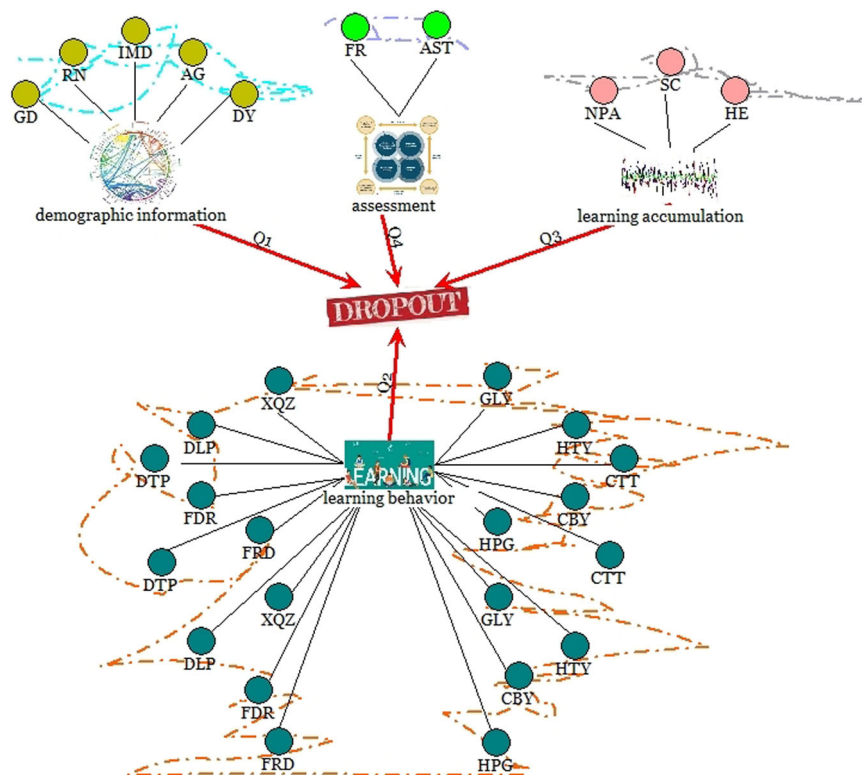
Fig. 1 Interrelationships and tested problems of influencing factors. The correlations are formed between Demographic Information, Learning Accumulation, Assessment, and Learning Behavior as latent variables and the observed variable of Dropout. Each correlation represents a problem to test the impacts of different latent variables on dropout trends.

between interactive learning activities and forming tested problems. The observation variables related to Demographic Information and Learning Accumulation form mutually constrained problems with Learning Behavior. As shown in Fig. 1, we construct the correlations between Demographic Information, Learning Accumulation, Assessment, and Learning Behavior as latent variables and the observed variable of Dropout. Each correlation represents a problem to test the impacts of different latent variables on dropout trends. At the same time, different latent variables are associated with many independent variables. The independent variables corresponding to Demographic Information, Learning Accumulation, and Assessment are consistent with Table 1, while the independent variables of Learning Behavior are refined into all interactive learning activities in Table 2. The value of the independent variable can be obtained through statistics or calculation, while the latent variables do not have specific values but might be described through the fusion of multiple independent variables. There may be potential relationships between independent variables associated with different latent variables, and we have marked them with dotted lines covering the independent variables in Fig. 1.

As the associated interactive learning activities are different for different courses, it is necessary to explore the strongly associated activities and construct potential behavioral paths. It is required to train data for different courses and learning periods, forming the topological structure of Learning Behavior, and then test whether it affects Assessment.

The four tested problems for STEM dropout prediction are:

P1: Whether learners' Demographic Information affects the dropout trends of STEM learning behavior.

P2: Whether the potential topological paths between interactive learning activities affect the dropout trends of STEM learning behavior.

P3: Whether Learning Accumulation affects the dropout trends of STEM learning behavior.

P4: Whether Assessment Results affect the dropout trends of new STEM learning behavior.

Methods

Based on previous academic research (Xing et al., 2016; Zhang et al., 2022), it is found that STEM learning behavior is composed of timestamps, interactive learning activities, and relationships. It is used to describe a series of continuous operations that one learner performs about a certain course in the temporal sequence. All learning behavior instances are stored in logs according to the temporal order. The accuracy of dropout prediction depends on the change patterns of learning behavior distributed in the temporal sequence (Hsu, 2022), as well as the data that can be mined from it, which can provide a basis for improving the STEM learning process (Borrella et al., 2022).

In order to more accurately predict the dropout trends in the learning process, the features described by Demographic Information, Learning Accumulation, and Assessment, as well as the associated independent variables, are defined. The features are special symbols or indicators that learners or learning behavior can recognize during the learning process. For example, by using the value of highest education, the features of learners can be directly divided into "higher educated" and "low educated", etc. At the same time, the different values of assessment results divide learners into different categories, the features of "Distinction", "Pass", "Fail", and "Withdrawn" are displayed because they can be directly defined by the final results of learners. These features can be directly determined by the descriptive nature of certain attributes, known as explicit features. Additionally, based on the participation of learners in the learning process, features of

learning behavior can be described as "positive" and "negative". This is the extended description of learning behavior, but it is obtained through the calculation of learner participation and interaction frequency, the features are not direct descriptive values of attributes, but they can be derived from several other attribute values or association values, which are the implicit features. In the process of data association analysis, it is necessary to calculate all explicit features and predict associate implicit features that are used to describe the learning interests or behavior trends. However, the relationships between explicit and implicit features are externalized as specific independent variables, and latent variables can be described as different feature categories based on the values of the associated independent variables.

This study will achieve the fusion of convolutional neural networks and recurrent neural networks in the method design process. The main focus is on two aspects: (1) Convolutional neural network uses a convolutional layer, pooling layer, and fully connected layer to achieve feature extraction and classification. The convolutional layer extracts local features through convolution operations, the pooling layer is used to reduce the dimensionality of the feature map, and the fully connected layer is used for classification. Its disadvantage is that it requires normalization of all data, making it difficult to train the mixed data with different lengths, and it lacks memory function, which is definitely not conducive to data analysis and prediction of continuous learning processes. It cannot track the explicit and implicit features of learning behavior before and after dropout. (2) Recurrent neural networks are deep learning structures capable of processing sequential data. It achieves modeling and prediction of sequence data through a combination of recursive and hidden layers. The recursive layer processes the temporal relationships of sequence data through recursive operations, while the hidden layer is used to learn the representation of sequence data. Recurrent neural networks are deep learning structures capable of processing sequential data. Its disadvantage is the training complexity, which requires a large amount of labeled data, and the computational process involving multi-data structures is extremely complex. The implicit features of the learning process are mainly derived through statistics and calculations and cannot be directly labeled, so the recursive neural network cannot be used directly. In order to address the shortcomings of convolutional neural networks and recursive neural networks and to solve the problems of Fig. 1, we consider the fusion of convolutional neural networks and recursive neural networks.

The fusion method of these two neural networks mainly includes the following three steps: Step 1. Feature extraction: It involves first extracting local features from the input data through the convolutional layer, applying long short-term memory network (LSTM), and outputting them through a fully connected layer; Step 2. Feature merging: The concatenate layer achieves the association and fusion of multiple features, merging key features, and the attention mechanism is used to determine the important information of the features; Srep 3. Results output: The feature analysis results obtained in the first two steps are inputted into the recursive layer, and LSTM processes the temporal sequences with relevant state information. Finally, the analysis results of learning behavior are identified and outputted.

For the early dropout prediction process of STEM learning behavior, this study model and analyzes it in a certain temporal sequence. The relevant method is named STEM_DP. Since the dataset we selected is collected on a daily basis, the basic unit of the temporal sequence is defined as one day, which can observe more details of behavioral changes. The entire analysis process of STEM_DP is divided into four steps: firstly, we predict and select key explicit features and realize feature scoring and ranking using

mutual information, random forest, and recursive feature elimination methods; secondly, we predict and mine the key implicit features of learner behavior, realize end-to-end feature tracking by constructing a convolutional neural network; thirdly, we predict and construct the topological structure of explicit and implicit features, improve the long-short-term memory mechanism of the recurrent neural network, realize the fusion with the convolutional neural network, then analyze and calculate the correlations between features and construct a learning path. Finally, combined with the analysis results of the above three steps, we derive the laws of changes in learning behavior. The analysis flow framework for dropout prediction is shown in Fig. 2.

The explicit and implicit features can be explored respectively by classical algorithms and convolutional neural networks. Regarding the topological structure of learning behavior, as STEM_DP combines convolutional neural network, recurrent neural network, and long-short-term memory mechanism, it needs to combine the distribution of explicit and implicit features, as well as the instance clustering in the learning process, to adopt a strategy of fusing key features and mining the strong correlation. The training process is as follows:

Step 1: The predicted results of explicit and implicit features are fused, and the related calculation formula is described as $L_T = L_E + L_I$ (Formula 1), where L_T is the loss function of the topological structure, L_E is the loss function of the explicit feature analysis process, and L_I is the loss function of the implicit feature analysis process. We use cross-entropy as the loss function, defined as $L = -\frac{1}{m}(\sum_{k=1}^{m} y_k \log \hat{y}_k + (1 - y_k)\log(1 - \hat{y}_k))$ (Formula 2). m is the size of the training batch, y_k is the expected output value of the kth training sample in each iteration process, and $\hat{y}_k$ is the predicted result of the kth training sample in each iteration process.

Step 2. The changes in the temporal sequence of explicit and implicit features are tracked. For the two hidden states h^t and s^t of the long short-term memory mechanism, we define the corresponding gradient values δ_h^t and δ_s^t. The calculation formulas are described, respectively, as $\delta_h^t = \frac{\partial L}{\partial h^t}$ (Formula 3) and $\delta_s^t = \frac{\partial L}{\partial s^t}$ (Formula 4), δ_h^t is jointly determined by the output gradient error for the corresponding convolution layer, i.e., $\delta_h^t = \frac{\partial L}{\partial h^t} = \frac{\partial l(t)}{\partial h^t} + \frac{\partial L(t+1)}{\partial h^t} \cdot \frac{\partial h^{t+1}}{\partial h^t} = V^{T(\hat{y}^t - y^t)} + \delta_h^{t+1} \cdot \frac{\partial h^{t+1}}{\partial h^t}$ (Formula 5). $l(t)$ represents the loss of the tth temporal sequence, $L(t + 1)$ represents the loss of the temporal sequence whose time index is greater than t, and V is the weight coefficient from the hidden state to the output.

Step 3. In the calculation process incorporating long short-term memory mechanism, the reverse gradient error of δ_s^t, denoted as δ_C^t, is jointly determined by the gradient error of δ_s^{t+1}, and the gradient error obtained from h^t in the corresponding convolution layer is described as $\delta_C^t = \frac{\partial L}{\partial s^{t+1}} \cdot \frac{\partial s^{t+1}}{\partial s^t} + \frac{\partial L}{\partial h^t} \cdot \frac{\partial h^t}{\partial s^t} = s^{t+1} \odot f^{t+1} + \delta_h^t \odot o^t \odot (1 - \tanh^2(s^t))$ (Formula 6), where f is the convolution function. The weight coefficients for learning route prediction can be calculated based on δ_h^t and δ_s^t.

Step 4. The forget gate weight coefficients of long short-term memory mechanism are defined as W_f. The gradient calculation formula is described as $\frac{\partial L}{\partial W_f} = \sum_{t=1}^{\tau} \frac{\partial L}{\partial s^t} \cdot \frac{\partial s^t}{\partial f^t} \cdot \frac{\partial f^t}{\partial W_f} = \sum_{t=1}^{\tau}[\delta_s^t \odot s^{t-1} \odot f^t \odot (1 - f^t)](h^{t-1})^T$ (Formula 7), where τ denotes the index of the last temporal sequence and is equivalent to the length of the entire complete temporal sequence.

This computational process can help the information processing system better adapt to complex temporal data, thereby improving processing efficiency and accuracy. In this process, explicit and implicit features are merged, modeled through a topology based on the convolutional neural network. Through continuous iterative training and optimization, the information
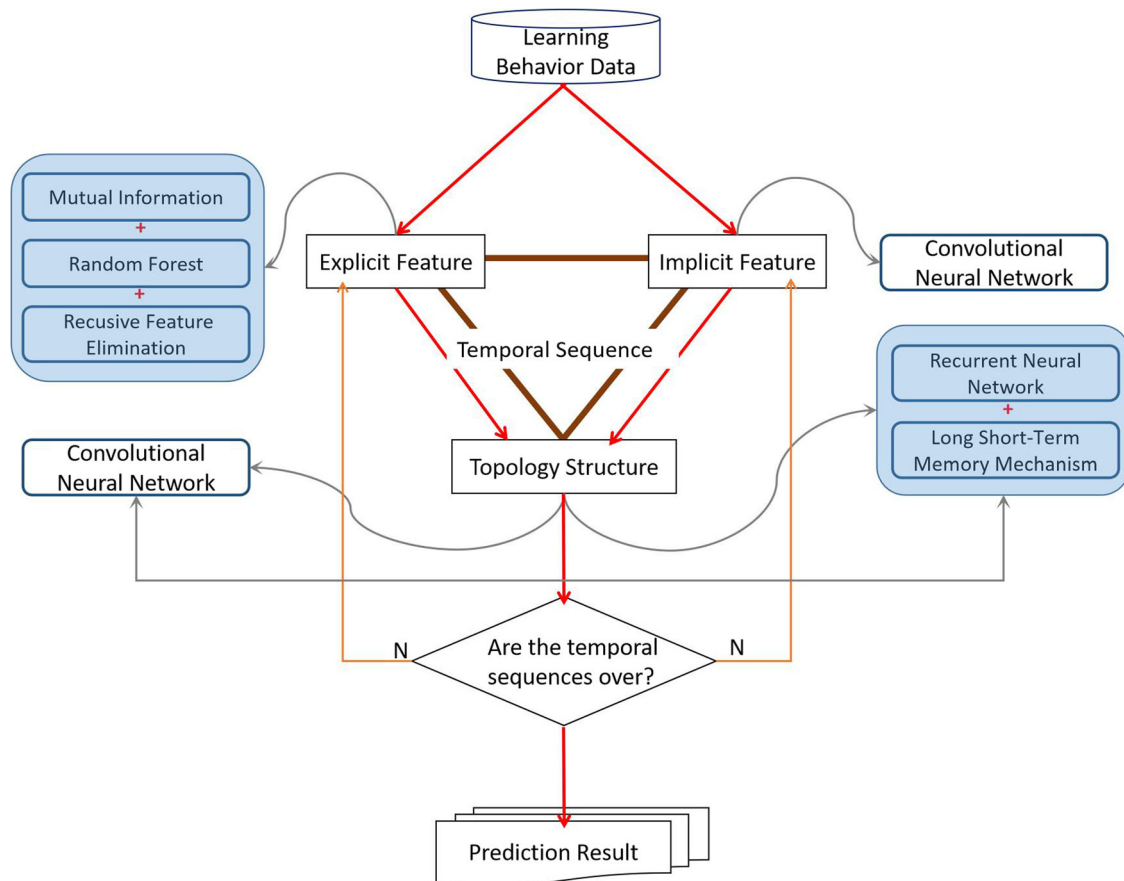
Fig. 2 Prediction and analysis process of STEM_DP. That is divided into four steps: Step 1. predict and select key explicit features and realize feature scoring and ranking; Step 2, predict and mine the key implicit features of learner behavior, realize end-to-end feature tracking; Step 3, predict and construct the topological structure of explicit and implicit features, analyze and calculate the correlations between features and construct a learning path; Step 4, derive the laws of changes in learning behavior.

processing system is able to automatically adjust the topological relationships based on actual circumstances and learn more accurate feature representations, thus possessing better adaptability in processing temporal data.

Experiments. Based on the three STEM courses and their corresponding learning behavior instances, STEM_DP is used to test the relevant problems proposed in the section "Data standardization and problem description" and evaluate performances in predicting dropout. In order to track learners' dropout trends, a comparative analysis of the evaluation indicators is analyzed to obtain the patterns that meet certain requirements. STEM_DP is iterated multiple times to select the optimal prediction results.

The dropout labeling for the learning behavior instances of 2013 and 2014 is as follows: Since assessment results of learners are described as four values, namely Distinction, Pass, Fail, and Withdrawn, learners labeled as Withdrawn are defined as dropouts and marked with "1", while learners labeled with other values are considered as non-dropouts and marked with "0". In the experimental process, the mini-batch stochastic gradient descent optimization algorithm helps to learn and select the suitable parameters, with a learning rate set at 0.001, a batch size is 256, and a total iteration set of 20,000 times. To complete the training and testing of STEM_DP, the dataset is randomly divided into training and testing sets in 8:2.

The learning behavior instances are modeled, and the four indicators of the test set are tracked and calculated. The changes in the indicator curves are visualized to explore the patterns of learning

behavior. Figures 3–5 illustrate the relationships between participation and different learning periods for three STEM courses. It can be seen that the learning behavior of DDD and FFF involves four learning periods, while EEE involves three learning periods. Even for the same course, the group trends of interactive learning activities among learners vary across different learning periods. Learners' participation is not only constrained by the courses but also by different learning periods. Therefore, dropout prediction should be implemented separately for each learning period, and performance indicators should be recorded to calculate the average values.

During the performance evaluation of STEM_DP, four indicators, Precision, Recall, F1 and AUC are selected. The dropouts of each learning period for the three courses are tracked and predicted. Thirty consecutive days are randomly selected from each period 10 times, and the average performance values of each period are calculated. Then the performance indicators of multiple learning periods for each course are averaged. Through sufficient data validation, Fig. 6 is obtained, it can be seen that the four performance indicators for the three courses are all above 0.900, indicating that the data analysis and prediction of STEM_DP have high reliability and accuracy. Among these three courses, FFF has the most types of interactive learning activities, the highest participation, and the largest scale for learning behavior. However, STEM_DP has the best data training effect, which is suitable for the associated calculation of multiple features and complex relationships, effectively tracking the temporal sequence of learning behavior, and achieving accurate classification and fusion. Since randomly 10 selected consecutive temporal sequences of
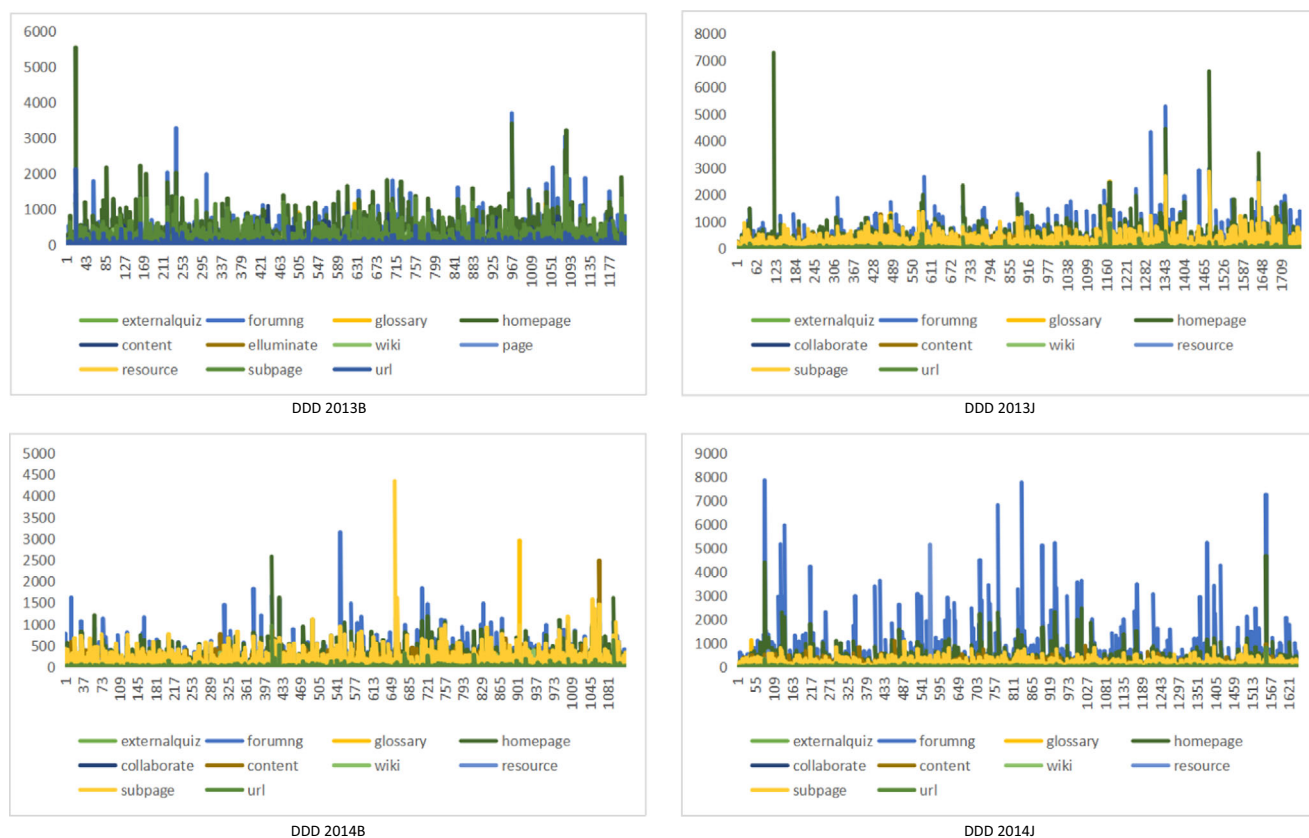
Fig. 3 Participation of DDD in interactive learning activities during the relevant learning period. That illustrates the relationships between participation and different learning periods (including 2013B, 2023J, 2014B and 2014J) for DDD.

30 days are taken as the basic duration, a comprehensive analysis of the full learning process of STEM_DP is achieved.

Furthermore, the performance change pattern of STEM_DP in the temporal sequence is tested, and the relatively optimal predictable temporal sequence is identified. Taking the learning behavior instances of three courses in 2014B as the analysis sample, 30 days are taken to form a continuous temporal sequence. STEM_DP analyzes and predicts the participation in interactive learning activities for each day, and the results of the four performance indicators are shown in Fig. 7.

Due to the extremely imbalanced learning behavior instances associated with DDD, EEE, and FFF, the proportion of dropout types is about 75%. The predictive performance indicators reach at least 75%, and experimental results find that these four indicators all meet the requirements. The predicted Precision for each day of the three courses exceeds 89% and demonstrates an overall increasing trend. Recall, F1, and AUC show a fluctuating slow upward trend. Based on the distribution of the four indicators in Fig. 7, it is found that the predictive performances of STEM_DP are relatively stable around the first 20 days. As days went on, the interactive learning activities also increased, which further enhanced its credibility. Therefore, the 20-day should be defined as the left boundary of the temporal sequence, and dropout prediction should be implemented, the training parameters and optimization indicators of STEM_DP dynamically might be updated and adaptively adjusted to achieve the best predictive effect.

So the application of STEM_DP in predicting dropout for the STEM courses is feasible. It can accurately track the dropout trends, analyze the temporal sequence of dropout prediction, and discover the topological path of dropout behavior and possible intervention strategies. The data analysis results can effectively be applied to the problem analysis proposed in the section "Data standardization and problem description".

Results

STEM learners in MOOCs have obtained widespread recognition, data statistics show that they tend to make full use of online learning resources and assessment methods, breaking through geographic and time limitations in learning behavior construction, providing many conveniences for self-learning and personalized learning. Learners can retrieve suitable courses, knowledge videos and interactive learning activities according to their own learning needs and habits (Xia, 2020a). Compared with social science courses, STEM's online courses have been more fully applied and might form a stable and sustainable learning process, retaining massive amounts of learning behavior instances. However, many of these learning behavior instances are incomplete and cannot describe the entire learning period (Xia, 2021a), many learners terminate the learning process prematurely, resulting in dropout. This is directly related to the learning organization method of MOOCs, which gives learners sufficient autonomy and flexibility but may neglect the tracking and supervision of the learning process, fail to assess and analyze the learning process in a timely manner, and do not realize decision analysis from massive learning behavior instances in MOOCs. At the same time, there is a crucial issue of the interdisciplinary nature of STEM courses, namely, the complex knowledge structure and internal relationships, which determine the design of the STEM learning method, effective and feasible interaction and cooperation should be established among learners and teachers during the learning process. Some knowledge-related experiments and quizzes should not be just quantitative tests or submissions of
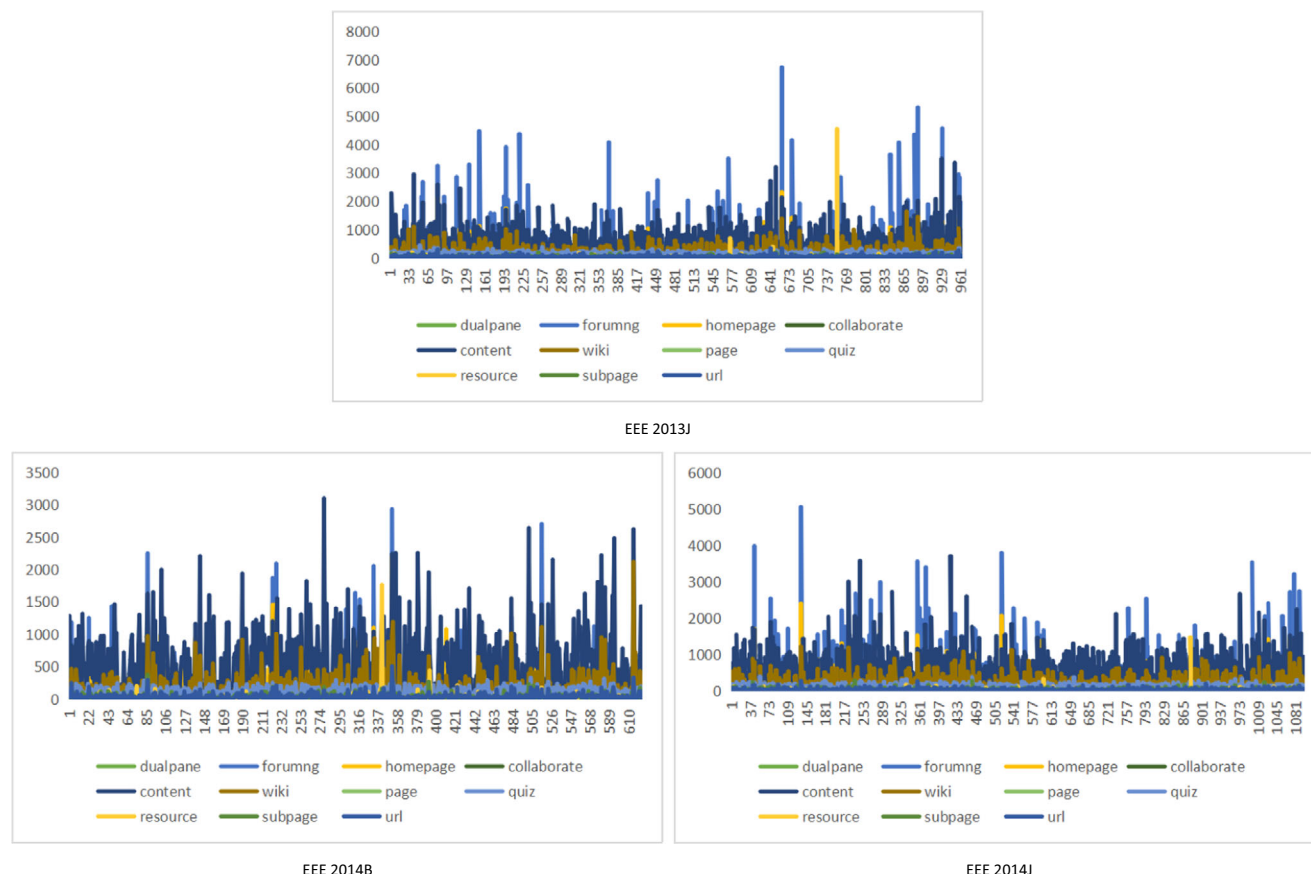
EEE 2013J

EEE 2014B

EEE 2014J

Fig. 4 Participation of EEE in interactive learning activities during the relevant learning period. That illustrates the relationships between participation and different learning periods (including 2013J, 2014B and 2014J) for EEE.

experimental results; they should be stepwise discussions and deductions based on principles and evidence. Therefore, the STEM learning process should not be a one-way propagation of learning videos; it requires communication, collaboration, and feedback among knowledge-driven learners, teachers, resource designers, etc. Thus, integrating the STEM learning process into MOOCs has its own features that are associated with the dropout prediction of the entire learning process and corresponding research topics, which have already been described in the section "Data standardization and problem description" and Fig. 1.

The learning behavior instances for DDD, EEE, and FFF are divided into two parts according to the final labels: dropout and non-dropout, and the distribution of relevant learning behavior instances are shown in Table 3, it can be seen that EEE has the largest proportion of dropout compared to non-dropout. Based on the analysis results of STEM_DP, we summarize the tested problems of STEM learning behavior shown in Fig. 1.

P1: Whether Demographic Information affects the dropout trends of STEM learning behavior.

Demographic Information of the learners described in Fig. 1 mainly involves five independent variables. Based on the data analysis results of STEM_DP, the dropouts of DDD, EEE, and FFF are investigated along the entire learning process from the 20-day to the end of the course assessment. Since Gender, Region, and Disability are non-quantitative values, coding is performed based on the distribution order of specific values. The results of the investigation are shown in Fig. 8. When defining Demographic Information as a latent variable, it did not produce a significant correlation with dropout. However, the variables have a significant impact on dropout. The results show that Age and IMD_band form a significantly negative correlation with dropout,

the younger the learners, the higher the dropout rate, and the lower the IMD_band, the higher the likelihood of dropout. Meanwhile, Gender and Disability might have no impact on learning effectiveness, but there is a dropout group of learners in different regions.

From Fig. 8, we can see that the dropout problem in these three different courses is consistent with the relationship between age and IMD_band. Younger learners have weaker learning focus and motivation to construct learning behavior, and the lower the IMD_band, the lower the learner's participation in the learning process. Learners have weak goals for passing the course assessment and might produce negative learning emotions within 20 days after starting learning. At the same time, there are similarities and overlaps in the three courses regarding regions with a higher probability of dropout. In the five regions of Fig. 8, the smaller the age and IMD_band, the significantly higher the risk of learners' dropout trends. So the probability of learners directly interrupting their learning behavior is extremely high.

P2: Whether the potential topological paths between interactive learning activities affect the dropout trends of STEM learning behavior.

In Fig. 1, there are many types of interactive learning activities associated with learning behavior, and there are differences in the interactive learning activities associated with DDD, EEE, and FFF. As shown in Table 2, not all interactive learning activities are participated by learners across the three courses. Regarding the analysis of the dropout trends of STEM learning behavior, it is necessary to conduct investigations for each course separately. The analysis results show that since learners tend to drop out around 20 days after starting learning, the interactive learning activities associated with dropout learners are not extensive or have a low
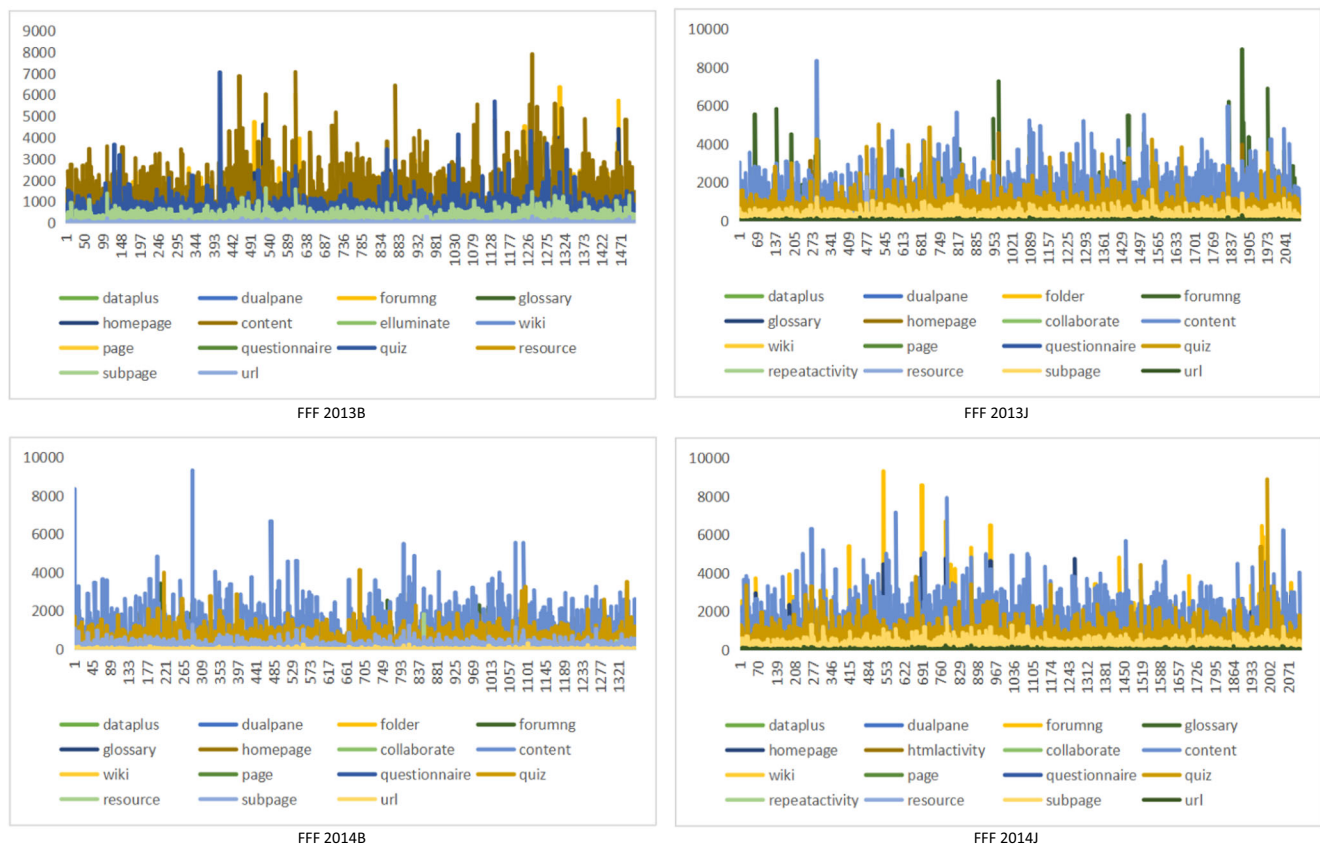
Fig. 5 Participation of FFF in interactive learning activities during the relevant learning period. That illustrates the relationships between participation and different learning periods (including 2013B, 2013J, 2014B and 2014J) for FFF.
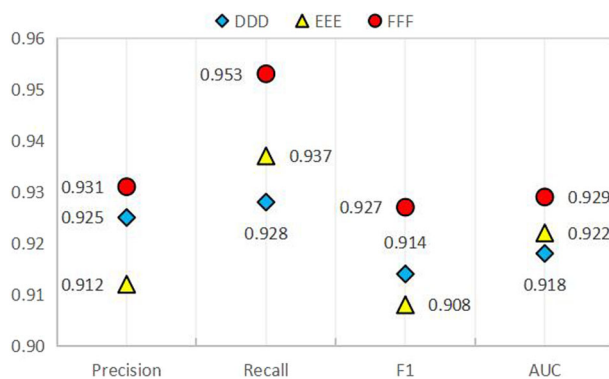

Fig. 6 Four indicators of STEM_DP. Four indicators "Precision", "Recall", "F1" and "AUC" are calculated and obtained for three courses "DDD", "EEE" and "FFF".

participation rate, directly leading to statistically insignificant dropout testing; therefore, this problem needs to explore the learning behavior instances of non-dropout learners and track the entire learning process. Through the construction of the potential topological path of benign learning behavior, we can deduce the possible problems when dropout occurs. Therefore, defining the interactive learning activities as independent variables and marking non-dropout as observation variables, based on the analysis and prediction of STEM_DP, we analyze the potential correlation between different interactive learning activities, build the effective path, and mine the feasible participation routes for learners. In the

investigation process, when a learner has a dropout label in a course, the probability of selecting and completing the course assessment again is very low. Therefore, we select interactive learning activities that have a significant and strong correlation, and the test results are shown in Fig. 9.

Figure 9 shows the learning behavior of each course in different learning periods and constructs key topological paths in the first 20 days of the learning process. After 20 days, interactive learning activities and routes that enable learning motivation and participation are promoted. it can be seen from Fig. 9 that DDD and EEE have similar interactive learning activities associated with learners in the first 20 days of the learning process, but different routes have been formed. After 20 days, External Quiz, Wiki, Resource, and Collaborate play an important role in driving DDD, while Quiz, Content, and Wiki receive strong participation from learners in EEE. The interactive learning activities and relationships of FFF are significantly richer than those of DDD and EEE. In the first 20 days, five interactive learning activities enable the construction of effective learning behavior, DataPlus stimulates learners' confidence about course learning. After 20 days, there are six key interactive learning activities for FFF, and the Questionnaire is given full play when timely tracking the learning process, promoting the propagation of learning materials and relevant data among learners. Regardless of any STEM course, interactive learning activities within the first 20 days all become effective learning behavior participation nodes run through the entire learning process. Overall, effective learning cannot be separated from collaboration, communication, and participation (Xia, 2021b). Forum in the early stages of all three courses has become a crucial starting node, driving the
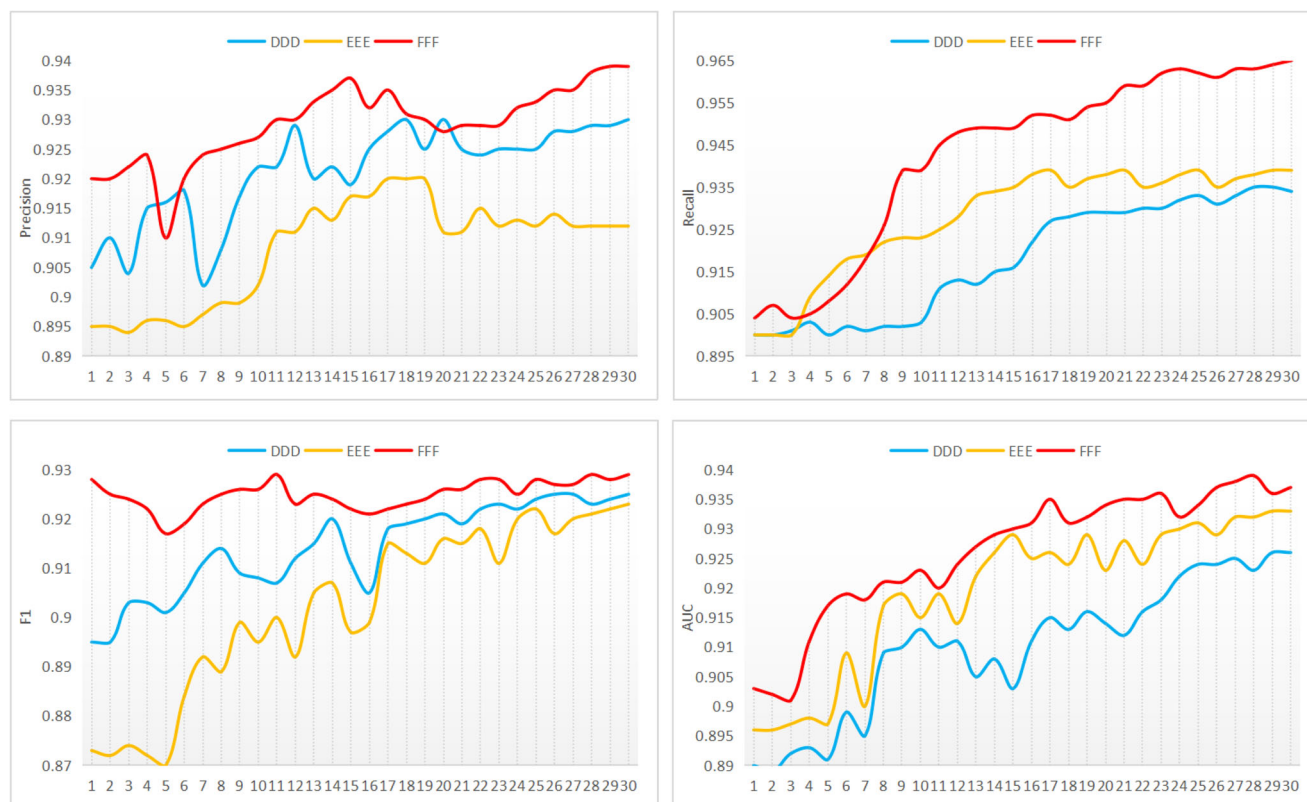
Fig. 7 Precision, Recall, F1 and AUC of STEM_DP about continuous temporal sequences. 30 days are taken to form a continuous temporal sequence. With the tracking and calculation of STEM_DP, the interactive learning activities for each day are analyzed and predicted correctly.

association and scheduling of other interactive learning activities. In addition, Quiz provides timely evaluation and feedback for testing and assessing different knowledge, helping learners discover learning problems and improving learning methods. The analysis results show that without effective guidance and construction of learning behavior within the first 20 days or without achieving effective interaction and cooperation throughout the entire learning process, there is a high dropout rate.

P3: Whether Learning Accumulation affects the dropout trends of STEM learning behavior.

Learning Accumulation described in Fig. 1 mainly involves three independent variables. Based on the data analysis results of STEM_DP, starting from the 20-day of each learning period until the end of course assessment, dropout datasets of DDD, EEE, and FFF are tested along the complete learning process. Since the Highest Education is a non-quantitative value, numerical encoding is performed for its value during testing. The test results are shown in Fig. 10. When defining Learning Accumulation as a latent variable, it forms a significant negative correlation with learners' dropout, and each independent variable has a negative significance with the dropout mark. The results show that the fewer studied credits, the lower the Highest education, or the fewer Num of previous attempts, the higher the dropout probability of learners. Learners with weaker subjectivity in constructing learning behavior are prone to group dropout trends shortly after starting learning.

From Fig. 10, it can be seen that the dropout problems of these three courses are consistent with the relationships between Learning Accumulation and related independent variables. The Studied credits of learners will significantly affect their learning attitudes and methods, leading to different subjective initiatives when facing new learning content. For new learners with little online learning experience, i.e., those with low values formed by

Table 3 Statistical results of dropouts and non-dropouts.

Course	Dropout number	Non-dropout number
DDD	2250	4022
EEE	722	2212
FFF	2403	5359

Num of previous attempts, relying solely on learners' consciousness and autonomy to organize the learning process can also easily lead to dropout. Therefore, the learning process should not be a completely personalized task organization mode (Xia, 2020b). Around the first twenty days of the learning process, STEM course content tends to increase in difficulty, logic and relevance with previous and upcoming knowledge. Once a learner does not have relevant Learning Accumulation, it will seriously affect the learning progress and hinder the understanding of knowledge. For young learners with weak self-discipline, the dropout is likely to occur.

P4: Whether assessment results affect the dropout trends of new STEM learning behavior.

The assessment described in Fig. 1 mainly involves two independent variables. Based on the analysis results of STEM_DP, starting from the 20-day of each learning period until the end of course assessment, dropouts of DDD, EEE, and FFF are tested. Since "final result" and "assessment type" are non-quantitative values, numerical encoding is performed. Because the vast majority of dropouts end their learning process early and do not participate in assessments, the assessment type is unfamiliar to those who have dropped out. Therefore, testing this problem with dropout behavior data does not carry statistical significance. The value distribution of Num of previous attempts in learners ranges
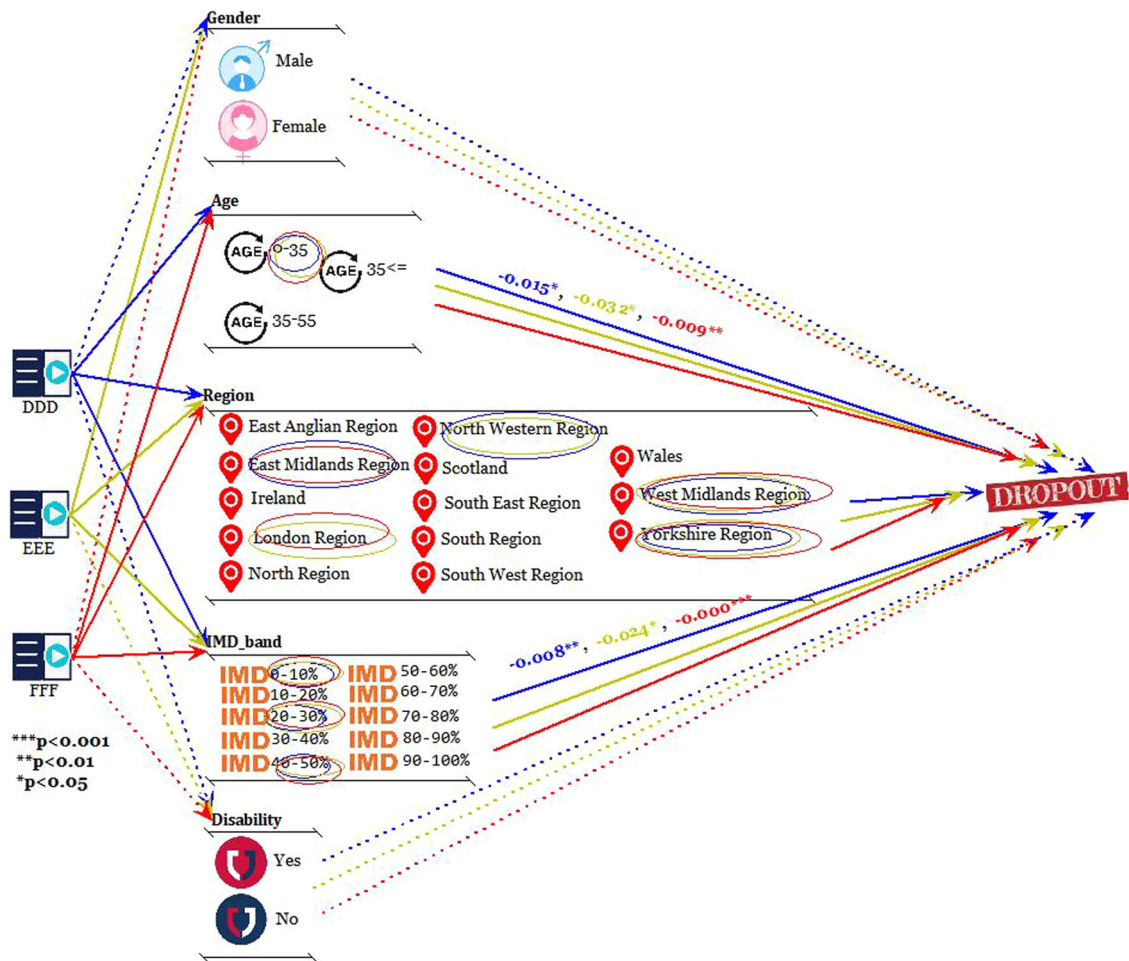
Fig. 8 Test results of demographic information for dropout. The dropouts of DDD, EEE, and FFF are investigated along the entire learning process from the 20-day to the end of the course assessment. There are five independent variables, that have a significant impact on dropout.

from 0 to 6, indicating that the same learner may have participated in online learning multiple times and completed the learning goals of multiple courses. To test the dropout rate after learners have completed at least one course's assessment, we use data from non-dropout learners. This is related to the prior assessment results and methods of learners and the test results are shown in Fig. 11. Except for Withdrawn, the Final result has three types. If learners' prior assessment results are "Fail", there will be a significant positive correlation with dropout. The more failures they experience, the stronger the correlation. When new learning courses cannot use the same assessment method as before, there is also a significantly positive correlation with dropout, so the learners who have participated frequently in online learning platforms tend to depend on the assessment method. They tend to choose previously used assessment methods, especially when their prior assessment results are Pass or Distinction. About the next course assessment, they are more likely to select the same assessment method as before.

Under certain conditions, the analysis results of dropout problems and assessment results for these three different courses are shown in Fig. 11. Regarding the impact of assessment methods, since the three courses are all STEM-related, learners tend to accept online testing more easily. However, this does not mean that learners do not have personalized preferences for assessment methods. When the assessment method of a new course is similar to the learner's previous learning experience, it is easier for learners to construct a complete learning process. These trends are clearly presented in the learning behavior instances of DDD

and EEE, but learners of FFF generally adapt to new assessment methods, and the assessment method is not a significant factor in causing dropout. With regards to the impact of assessment results, the actual assessment grades (Distinction, Pass, and Fail) cannot directly cause dropout. Therefore, the correlation of multiple learning processes is analyzed. If learners previously have one or more experiences here they do not pass a course assessment, they are more likely to drop out when selecting a new course. However, if their prior assessment results are Pass or Distinction, it would enhance the learner's confidence in participating in new courses, strengthen their enthusiasm for knowledge, and lead to a lower dropped rate.

After tracking and statistical analysis of the entire learning process, the probability of learners who pass the assessment participating in other courses again is 92.22%, and their pass rate is as high as 95.47%. The probability of learners who fail the assessment participating in online learning again is 65.43%, which means that about one-third of learners have directly abandoned the learning process of MOOCs. When these learners who do not pass the assessment participate in online learning again, 78.19% of them leave MOOCs midway, indicating that these learners are still troubled by the problems they encountered during their previous learning and still develop negative emotions. It can be seen that learners who pass the assessment have a higher probability and pass rate of participating in MOOCs again, they might have more positive emotions compared to the learners who do not pass the assessment, while learners who do not pass the assessment have lower indicators and have produced a large number of dropouts.
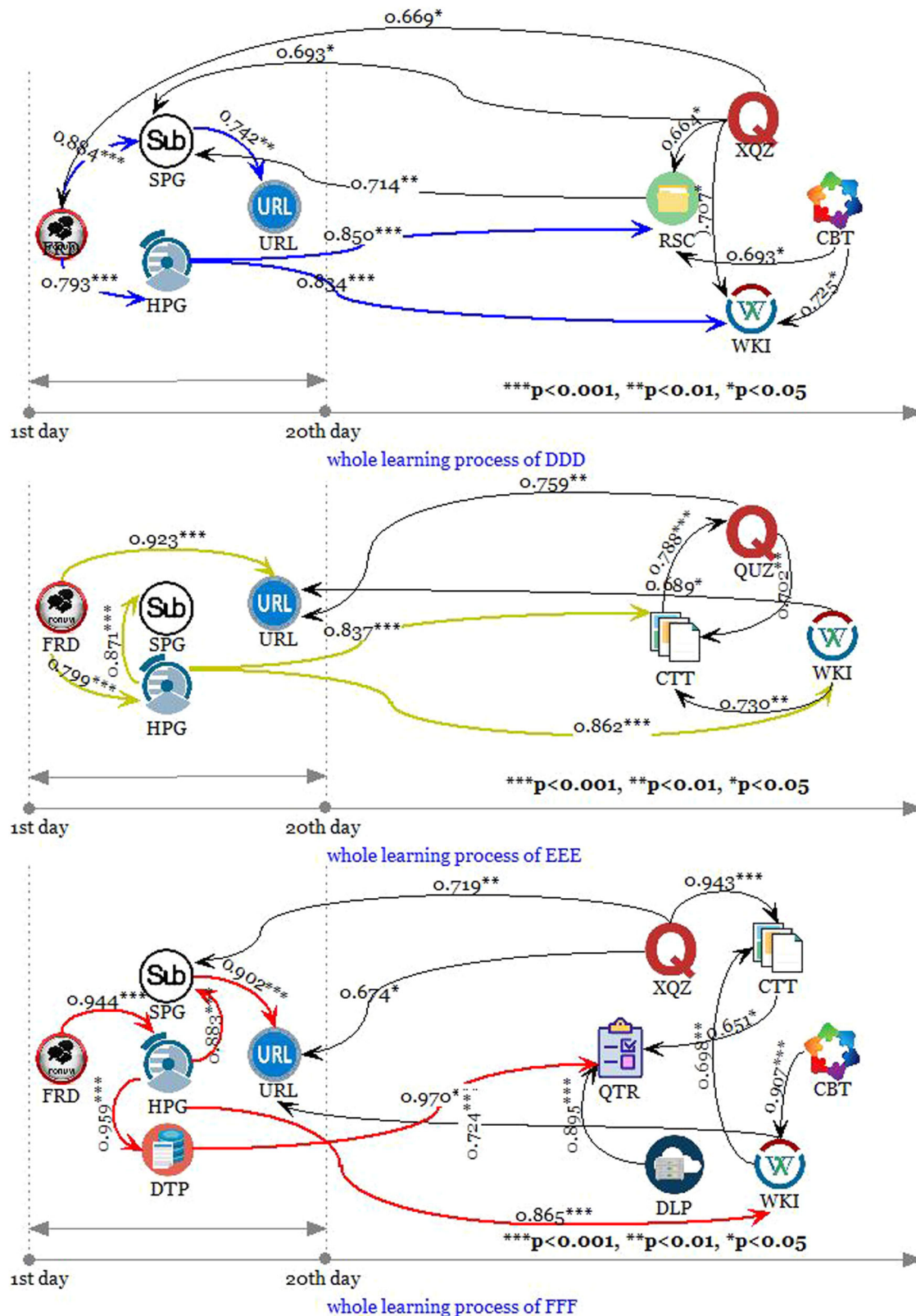
Fig. 9 Test results of interactive learning activities for dropout. That shows the learning behavior of DDD, EEE and FFF in different learning periods and constructs key topological paths in the first 20 days of the learning process.

Based on the significant test of the four above problems, it is found that for the dropout of STEM learning behavior in MOOCs, Demographic Information, interactive learning activities and relationships, Learning Accumulation, and assessment results all have direct or indirect impacts. Therefore, it is necessary to guide and recommend learners to select appropriate interactive learning activities and build suitable learning routes. The whole learning process cannot be separated from the
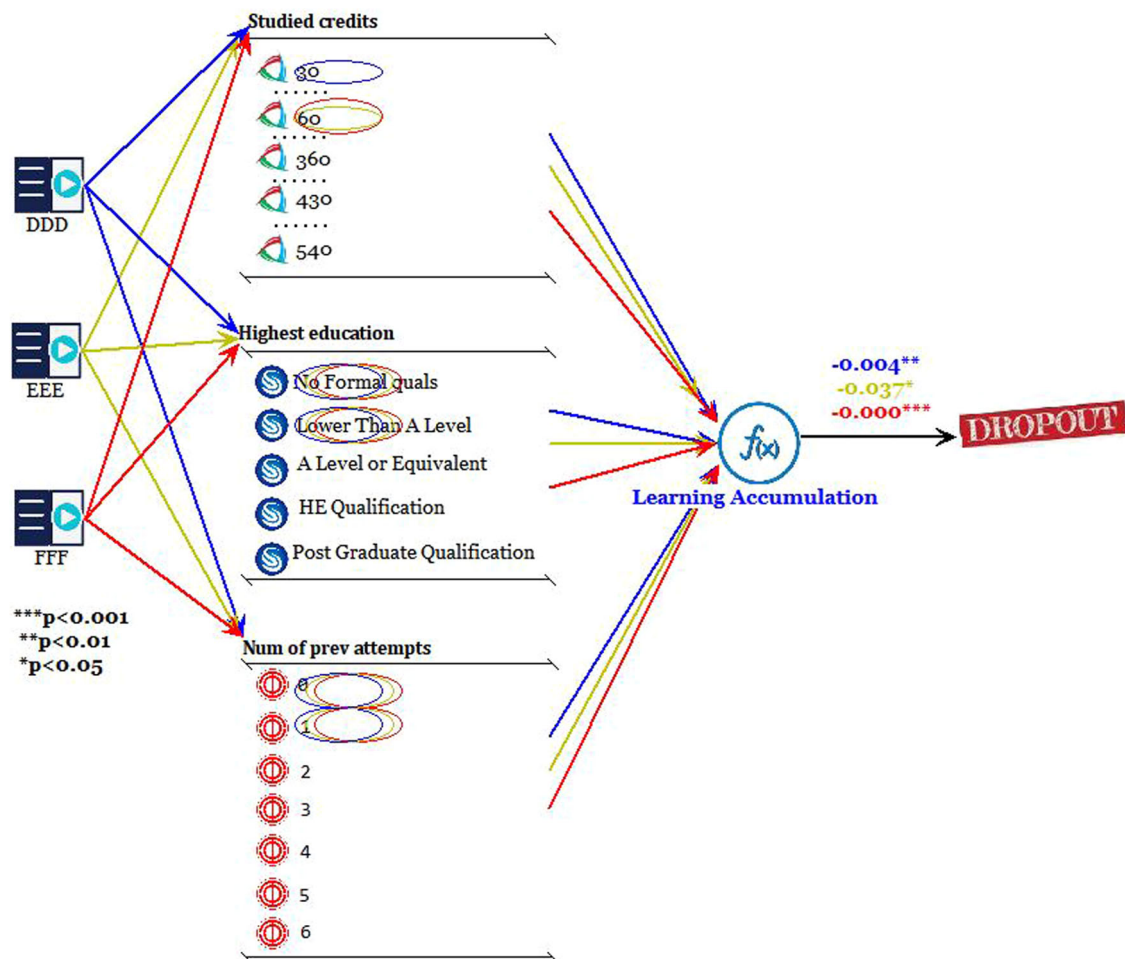
Fig. 10 Test results of learning accumulation for dropout. That shows that the fewer studied credits, the lower the Highest education, or the fewer Num of previous attempts, the higher the dropout rate. Learners might form group dropout trends.

tracking, supervision, and intervention of key temporal sequences.

Discussion
This study tracks the complete MOOC learning behavior of three STEM courses and mines massive instances. To solve the high dropout rate in the online learning process, a dropout prediction model is proposed to address the complex interaction and collaboration. Experimental results show that the model can improve the quality and effectiveness of STEM dropout prediction, a comprehensive correlation analysis and training of relevant factors are completed, and the key elements leading to dropout are summarized. In this section, we will discuss the findings and suggestions.

Findings. Based on the previous research results (Bañeres et al., 2023) and the analysis of complete learning behavior instances, STEM learning behavior is influenced by timestamps, interactive learning activities, and relationships. The effectiveness of constructing learning behavior can also be affected by learners' backgrounds and demographics, making the correlation analysis and tracking the effect of massive learning behavior instances directly related to the temporal sequence. Over time and with changes in knowledge structure, the topological paths and routing strategies of learning behavior will also undergo corresponding changes (Aldowah et al., 2020). Some learners may develop a learning burnout state or even drop out. Therefore, in the

dropout prediction process, learning behavior needs to be described as a series of related interactive learning activities in a continuous temporal sequence. By incorporating temporal sequence into the learning behavior analysis (Khoushehgir and Sulaimany, 2023), the potential values and rules might be explored, which has important guiding significance for the study of STEM Learning Behavior.

For achieving the prediction of dropout behavior in STEM learning, we might take into full consideration both explicit and implicit features of the complete learning process. The explicit features are defined items that already exist in the dataset, while the implicit features refer to the new descriptive items generated by learners during the learning process. The data analysis and prediction process of STEM_DP is mainly divided into four aspects.

(1) By tracking and analyzing the entire learning process, this study explores and perceives the key explicit features that influence dropout behavior. Quantitative indicators are calculated based on learner participation and preference, and features are selected and ranked accordingly. Three STEM courses that generated massive learning behavior instances are selected, and it is found that the key explicit features include Demographic Information, interactive learning activities, Learning Accumulation, and assessment results. This is consistent across all three courses. Prediction results are fused through a convolutional neural network and predicted algorithm and experimental results
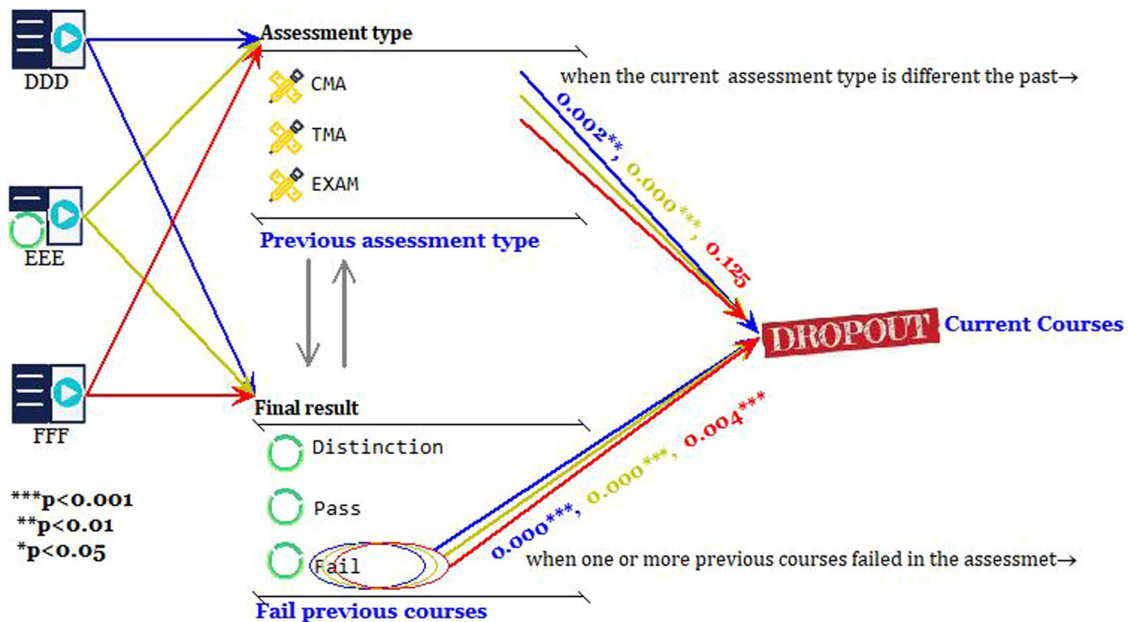
Fig. 11 Test results of assessment results for dropout. That shows that learners might choose previously used assessment methods, the actual assessment grades (Distinction, Pass, and Fail) cannot directly cause dropout.

demonstrate that these explicit features can significantly impact learning behavior paths.

(2) By deeply correlating and calculating multiple learning periods, this study explores the implicit features related to key explicit features. By building a multi-layer neural convolutional network that is suitable for STEM learning behavior, implicit feature tracking for multiple learning periods is achieved. The different interactive learning activities and relationships are present in the same course at different learning periods, but the activities and relationships that have significant impacts are highly similar. These interactive learning activities and the topological paths play the key enabling roles in each course that constitute the key implicit features to drive changes in the learning process. By mining and describing the key implicit features of three courses, this work expands the descriptive factors for predicting dropout in learning behavior.

(3) This study achieves the fusion of key explicit and implicit features and examines their mutual influences. The experiment finds that explicit features can potentially influence the selection and construction of implicit features, and historical data can positively or negatively affect the construction and optimization of new learning behavior. Therefore, the memory mechanism for recurrent neural networks can perform correlated calculations of existing data and achieve the reliability of the analytical results, which might ensure multi-layer convolution operations.

(4) Based on the STEM_DP analysis process of the three aspects above, the relevant test results of STEM dropout prediction are deduced, and the patterns and issues of learning behavior are summarized. The analysis results show that different learning courses have different explicit and implicit features, and dropout is not due to learners directly giving up course assessment but is closely related to the temporal sequence of the entire process. At the early stage, learning behavior exhibits disorderliness and spontaneity, with no effective guidance and construction of effective learning behavior. This further illustrates the importance and necessity of introducing the temporal sequence participation in feature calculation.

Through the analysis of tested problems about dropout prediction for STEM learning behavior, the design of the model and the experiment of the full learning process, it has been demonstrated that STEM_DP can fully select and effectively calculate dropout factors, achieve the fusion of explicit and implicit features, improve the effectiveness of dropout prediction, and locate the key temporal sequence. Compared with the analysis of dropout in small-scale data and the use of baseline methods and tools with large limitations, STEM_DP has stronger applicability and completeness. This is an innovative design and implementation of a research method and experimental program driven by massive learning behavior instances.

Suggestion. The STEM courses have high requirements for practicality, applicability, and experimental skills. When these courses are presented to learners through MOOCs, it also requires matching auxiliary materials, interactive collaboration, and learning behavior. A complete learning process includes a continuous temporal sequence. When learners with different backgrounds and experiences study the same course together, personalized learning methods and differentiated learning outcomes will also be produced (Xia, 2021c). Due to the fully open learning mode of MOOCs, the construction of STEM learning behavior is not easy. The emergence of a large number of dropouts and the inadequate use of related learning resources have made learning behavior ineffective. Dropout issues are also key challenges and hot topics in STEM course learning. Based on the analysis of online learning behavior instances generated by STEM education, targeting the multi-features and complex relationships that lead to STEM dropouts and early warning signs, four tested problems are demonstrated. In addition, multiple types of problems and patterns that exist during the dropout process are uncovered. Based on these findings, several suggestions are proposed regarding the intervention and early warning for STEM.

Demographic Information is applied to predict and intervene in dropout trends. The dropout prediction process of STEM_DP shows that Age and IMD in Demographic Information can have a

negative significance on dropout, meaning that the older the learner is or the higher their Index of Multiple Deprivation (IMD), the lower the probability of dropout. Older learners have clearer learning goals and greater stability in the learning process. The IMD is used to describe the composite deprivation situation of one region, covering seven different deprivation dimensions such as income, employment, education, health, living environment, housing and service barriers, and crime. The greater this index, the greater the stability in learning. They create potential differences among learners. As the number of descriptive items in Demographic Information increases, it will also affect the learning process and subtly affect learners' learning enthusiasm and investment.

Therefore, MOOCs should make adaptive evaluations and judgment on learners' dropout trends based on their previous Demographic Information. The learning attitudes and methods are still easily influenced by their own learning experience. Learners also tend to use existing learning strategies to cope with new learning tasks, as this is their previous learning process experience, which may enable them to better meet their previous learning needs. Based on some psychological research findings (Lee et al., 2022; Weiss et al., 2023), they are willing to rely on such existing cases to enhance their understanding of new learning behavior. Whether it is success or failure, this is an intangible mindset that may help learners adapt to the new learning process as soon as possible. If the existing learning experience can be suitable for the current course, it may also bring some positive effects. If this learning experience is not suitable for the new learning goals, learners might feel frustrated or experience significant burnout, then negative emotions arise, and even the dropout occurs directly. If MOOCs can effectively evaluate learners' Demographic Information, identify the potential risks or advantages of their growth and learning experience, and provide effective guidance and intervention in combination with new courses, the dropout rate can be reduced to some extent. So suitable learning resources should be selected and utilized. Different learning behavior guidance strategies should also be generated according to the differences in Demographic Information.

Learning accumulation is applied to predict and intervene in dropout trends. The higher the completion rate and the more courses learners pass on the MOOCs platform, the lower the dropout rate. When learners have higher levels of education, they are also more likely to complete the entire learning process. Learners' attention and participation are also higher, enabling them to build efficient and effective learning behavior as soon as possible, which is proved by the dropout prediction results, although it may not apply to all learners. However, Learning Accumulation related factors have been proved that can significantly affect dropout trends. When learners have weak Learning Accumulation, especially new learners, if they select inappropriate learning resources or teaching methods or have unclear learning goals or weak motivation, the learning process will carry greater risks.

Therefore, before starting the STEM-related course learning on the MOOC platform, analysis and evaluation should be made for their potential knowledge gap and learning method based on their previous Learning Accumulation. The existing Learning Accumulation determines their knowledge foundation about a subject or major. If a learner systematically learns the precursor knowledge related to a certain course, it is easier for them to understand the new knowledge framework and related concepts for the new course. Data analysis results on MOOCs show that learners tend to make the related selection according to their own

professional skills, and of course, they also focus on their own academic interests and preferences, which are directly related to their existing Learning Accumulation (Rahimi, 2023; Wei et al., 2022; Zhu et al., 2022). However, there are also some learners who select courses that are not directly related to their existing Learning Accumulation, which achieves the interdisciplinary learning process. In this case, learners are more suitable to select the basic courses. At the same time, MOOCs should also provide learners with new knowledge context and more suitable learning needs to promote the learning process, helping them select the suitable courses instead of blindly selecting based on just one decision, it does not help the construction of effective learning behavior, and it is also difficult to efficiently advance the learning task. So suitable learning guidance plans and implementation measures should be associated with these findings. This also means that MOOCs' learning resources and learners should be appropriately evaluated and categorized to establish an effective mapping.

Assessment results are applied to predict and intervene in dropout trends. Assessment results cannot directly influence the dropout during the current learning process. Any learner who completes the course assessment has completed the entire learning process and is marked as non-dropout. However, regardless of the course, once the learners participate in the assessment, they face a situation where they need to complete other related courses on the same learning platform or not, whether different results will affect the learning interest and motivation, whether assessment methods and contents will affect their new course learning or even lead to dropout, etc. These issues can have significant impacts on STEM course learning due to the strong correlation. The analysis results of STEM_DP show that assessment results can affect the dropout trends in the new learning process, and it is also necessary to drive learner tracking and decision intervention after assessment results.

Therefore, after one learner completes the learning task of STEM courses on the MOOCs platform, the analysis and prediction of their potential learning trends and motivation should be made based on their previous assessment results, which will have a significant impact on their learning state to some extent. A learner who passes the assessment and performs well will have a greater enthusiasm for engaging in a new learning process. A good result is a driving force for stimulating suitable learning behavior. However, a learner whose assessment results are not ideal will have their learning state more or less affected. Some learners may have unwilling thoughts and restart the learning process until they achieve their learning goals. Some learners may develop certain negative emotions. Once learners lack direct supervision, tracking, and intervention from teachers, learners may lose confidence in MOOCs due to unsatisfactory assessment results (Xia and Qi, 2023a). In the subsequent new learning process, they might present a burnout state and drop out directly. It is necessary to evaluate the learners' potential learning engagement and participation in the next step based on assessment results, as well as predict the learning trends that learners may experience. The following learning process and demands of the learners should be tracked, analyzed, and judged using an adaptive temporal sequence approach, providing effective guidance for learning behavior and constructing learning strategies to reduce the dropout rate.

Key interactive learning activities and topological paths are applied to predict and intervene in dropout trends. The online learning mode of MOOCs provides learners with enough autonomy, making it easy for them to personalize their learning behavior

and form their own topological paths and behavioral routes during the learning process. At the same time, learners who take the same course unknowingly share certain commonalities in learning behavior to achieve similar learning goals. However, due to some potential factors, they may show differences in different learning periods. Courses belonging to the same major or category also share some similar features in learning methods and assessment modes. This study uses STEM_DP to mine the key implicit features and relationships of effective learning behavior in three courses and form topological paths. The data analysis results show that although learners study different courses, they can still have similar or identical structures in some key interactive learning activities and learning behavior routes. These findings can provide effective guidance for learners' gradual improvement of learning methods, as well as the selection of different courses.

Therefore, after complete learning behavior instances are generated for a certain course, a corresponding effective learning behavior topological path should be created. The learning process can be selected by one learner alone, but the construction of learning behavior requires various factors, and effective learning methods can be explored not only based on the learner's habits or interests. Although MOOCs might provide learners with sufficient autonomy and personalization in time and space, effective learning process and efficient learning behavior cannot be separated from applicable interactive learning activities and build a reliable or continuous learning path between them. MOOCs have constructed various interactive and collaborative activities for the online learning process, but for learners, how to quickly explore key interactive learning activities and feasible behavior routing strategies is the key issue. This requires MOOCs to derive potential feasible learning paths based on past learning behavior instances and also to provide timely guidance on adaptive learning routes based on learners' profiles. Based on effective learning behavior routing, associated relationships should be built across different courses. For new courses, related courses with associated knowledge should be explored based on the evaluation results of course content and relevant concepts. When learners participate in the online learning process, they should be recommended suitable interactive learning activities and guided on the feasible learning behavior route while also evaluating their learning experience and providing them with recommended learning content, reference resource, and effective guidance for learning behavior that is suitable and effective.

Conclusion

STEM is an educational concept that integrates multiple disciplines and knowledge structures. It differs from the traditional method of focusing on single disciplines and is related to many interdisciplinary concepts associated with applied practice. With the development and recognition of interdisciplinary advantages, as well as the continuous emergence of new technologies and new modes, improving the learning experience and learning effectiveness of STEM education has become a hot topic for both academia and industry. MOOCs have provided new ideas and attempts for the development and improvement of STEM but have not avoided the widespread problem of high dropout rates in online learning. While MOOCs have enabled the personalized learning process for learners, they might not achieve the goal of enhancing learners' abilities and effectively accumulating knowledge. This situation does not help to improve learners' skills and innovative practical solutions of STEM-related disciplines.

This study focuses on the massive learning behavior instances generated by STEM education in MOOCs, fully considering the entire learning period, analyzing and demonstrating the factors related to dropout, defining explicit and implicit features that can fully describe the learning process, and designing one predicted model STEM_DP to check the temporal sequence of dropout and provide an early warning in the learning process. Through multi-step iterative convolution operations and long–short-term memory mechanism, STEM_DP analyzes key issues related to dropout prediction and locates relevant temporal sequences and influencing factors that lead to dropout. Based on the propagation of STEM learning behavior within the same course and approximate learning behavior similarity between different courses, this study summarizes potential dropout risks and decision recommendations for courses and learners. The entire study aims to provide benign guidance strategies and improvement measures for STEM learning behavior in MOOCs, effective methodologies are used to optimize the supervision and tracking of the online learning process.

Through the analysis and argumentation, as well as the design and experimentation of innovative methods in this study, we gradually recommend MOOCs to enable the effectiveness of STEM education. Based on the mining and prediction of data value, on the one hand, it might help teachers more accurately guide and cultivate students' ability to think and solve problems, thereby helping learners better adapt to future work changes, discover new values, and generate innovation (Xia., 2022). On the other hand, it might enable to optimize the effective sharing of online resources and flexibly implement key service scheduling mechanisms in MOOCs, achieve more reliable analysis and prediction of learner preferences and needs, timely recommend applicable learning resources and efficient learning methods, help learners build positive and efficient learning behaviors in the shortest possible time, drive learners to self discover and solve problems, and explore the innovative ability of active learning and reflective learning. Furthermore, to provide more comprehensive support for interaction and collaboration among STEM learners during the learning process, with learning tasks as the goal, this study encourages learners to cooperate and share in teams, cultivate their teamwork spirits, and form a proactive learner social circle in MOOCs. This helps learners realize each other's skills and knowledge, fully leverage their personal strengths and improve their teamwork abilities. Finally, the analytical conclusions of this study can provide some key basis for the integration of MOOCs and STEM education. While fully exploring and predicting the value of existing learning behavior instances, it can better achieve feedback and early warning for subsequent learning processes, optimize learning experiences, improve learning quality, reduce negative emotions in the learning process, cut down dropout rate, and enhance the reliability and feasibility of MOOCs effectively assisting STEM education, the whole research has strong practical significance.

In the following work, the data description and analysis will be further optimized based on changes and substitutions in factors that describe learning behavior. The topological relationships of learning behavior routing will also be enriched, and the robustness and accuracy of the STEM_DP will be expanded to provide more technical decisions for the effective promotion of online STEM education and learning.

Data availability

The datasets used or analyzed during the current study are available from the corresponding author on reasonable request, relevant data can also be obtained and downloaded from https://analyse.kmi.open.ac.uk/open_dataset.

References

Aldowah H, Al-Samarraie H, Alzahrani AI et al. (2020) Factors affecting student dropout in MOOCs: a cause and effect decision-making model. J Comput High Educ 32:429–454. https://doi.org/10.1007/s12528-019-09241-y

Alrajhi L, Alamri A, Pereira FD et al (2023) Solving the imbalanced data issue: automatic urgency detection for instructor assistance in MOOC discussion forums. User Model User-Adapt Interact 2023(12):1–56. https://doi.org/10.1007/s11257-023-09381-y

Alves P, Lopes A, Cruz-Correia R et al (2023) The interrupted journey: factors and processes related to withdrawal, re-enrolment and dropout from doctoral education. High Educ 2023(10):1–18. https://doi.org/10.1007/s10734-023-01113-w

Anttila S, Lindfors H, Hirvonen R et al. (2023) Dropout intentions in secondary education: student temperament and achievement motivation as antecedents. J Adolesc 95(2):248–263. https://doi.org/10.1002/jad.12110

Bañeres D, Rodríguez-González DM, Guerrero-Roldán AE et al. (2023) An early warning system to identify and intervene online dropout learners. Int J Educ Technol High Educ 20(1):1–25. https://doi.org/10.1186/s41239-022-00371-5

Benoit DF, Tsang WK, Coussement K et al. (2024) High-stake student drop-out prediction using hidden Markov models in fully asynchronous subscription-based MOOCs. Technol Forecast Soc Change 198(1):123009. https://doi.org/10.1016/j.techfore.2023.123009

Borrella I, Caballero-Caballero S, Ponce-Cueto E (2022) Taking action to reduce dropout in MOOCs: tested interventions. Comput Educ 179:104412. https://doi.org/10.1016/j.compedu.2021.104412

Boumparis N, Baumgartner C, Malischnig D et al. (2023) Effectiveness of a web-based self-help tool to reduce problem gambling: a randomized controlled trial. J Behav Addictn 12(3):744–757. https://doi.org/10.1556/2006.2023.00045

Buckley J, Gumaelius L, Nyangweso M et al. (2023) The impact of country of schooling and gender on secondary school students' conceptions of and interest in becoming an engineer in Ireland, Kenya and Sweden. Int J STEM Educ 10(1):1–25. https://doi.org/10.1186/s40594-023-00416-9

Cara JA, Matthew LB, Mladen R et al (2022) Predicting student outcomes using digital logs of learning behaviors: review, current standards, and suggestions for future work. Behav Res Methods 2022(7):1–29. https://doi.org/10.3758/s13428-022-01939-9

Cervantes-Soon CG, Dorner L, Palmer D et al. (2017) Combating inequalities in two-way language immersion programs: toward critical consciousness in bilingual education spaces. Rev Res Educ 41(1):403–427. https://doi.org/10.3102/0091732X17690120

Chen X, Zou D, Xie H et al. (2021) Past, present, and future of smart learning: a topic-based bibliometric analysis. Int J Educ Technol High Educ 18(2):1–29. https://doi.org/10.1186/s41239-020-00239-6

Chu HC, Hwang GH, Tu YF et al. (2022) Roles and research trends of artificial intelligence in higher education: a systematic review of the top 50 most-cited articles. Australas J Educ Technol 38(3):22–42. https://doi.org/10.14742/ajet.7526

Cristea AI, Alamri A, Alshehri M et al (2023) The engage taxonomy: SDT-based measurable engagement indicators for MOOCs and their evaluation. User Model User-Adapt Interact 2023(5):1–52. https://doi.org/10.1007/s11257-023-09374-x

Dash R, Ranjan KR, Rossmann A (2022) Dropout management in online learning systems. Behav Inf Technol 41(9):1973–1987. https://doi.org/10.1080/0144929X.2021.1910730

de Oliveira Marques V, Kort-Kamp LM, de Souza MAN et al (2023) Oral health and behavioral management of children with autistic spectrum disorder: a 30-year retrospective study. J Autism Dev Disord 2023(12):1–8. https://doi.org/10.1007/s10803-023-06209-x

Dhiman NK, Salodkar SM, Gagandeep et al (2023) Advances in modeling and control of magnetorheological elastomers for engineering applications Arch Comput Methods Eng 2023(12):1–43. https://doi.org/10.1007/s11831-023-10031-0

Evenhouse D, Lee Y, Berger E et al (2023) Engineering student experience and self-direction in implementations of blended learning: a cross-institutional analysis. Int J STEM Educ 10(1):1–22. https://doi.org/10.1186/s40594-023-00406-x

Flegr S, Kuhn J, Scheiter K (2023) How to foster STEM learning during Covid-19 remote schooling: combining virtual and video experiments. Learn Instr 86(1):101778. https://doi.org/10.1016/j.learninstruc.2023.101778

Gallagher M, Lamb J (2023) Open education in closed-loop systems: enabling closures and open loops. Distance Educ 44(4):620–636. https://doi.org/10.1080/01587919.2023.2267475

Guo LM, Du JL, Zheng QH (2023) Understanding the evolution of cognitive engagement with interaction levels in online learning environments: insights from learning analytics and epistemic network analysis. J Comput Assisted Learn 2023(1):1–18. https://doi.org/10.1111/jcal.12781

Gupta A, Garg D, Kumar P (2022) Mining sequential learning trajectories with hidden Markov models for early prediction of at-risk students in e-learning environments. IEEE Trans Learn Technol 15(6):783–797. https://doi.org/10.1109/TLT.2022.3197486

Hsu LW (2022) EFL learners' self-determination and acceptance of LMOOCs: the UTAUT model Comput Assist Lang Learn 2022(5):1–20. https://doi.org/10.1080/09588221.2021.1976210

Jansen RS, Van Leeuwen A, Janssen J et al. (2022) Exploring the link between self-regulated learning and learner behaviour in a massive open online course. J Comput Assist Learn 38(4):993–1004. https://doi.org/10.1111/jcal.12675

Khor ET, Darshan D (2022) A learning analytics approach using social network analysis and binary classifiers on virtual resource interactions for learner performance prediction. Int Rev Res Open Distrib Learn 23(4):123–146. https://doi.org/10.19173/irrodl.v23i4.6445

Khoushehgir F, Sulaimany S (2023) Negative link prediction to reduce dropout in massive open online courses Educ Inf Technol 2023(1):1–20. https://doi.org/10.1007/s10639-023-11597-9

Labrovic JA, Petrovic N, Andelkovic J et al (2023) Patterns of behavioral engagement in an online English language course: cluster analysis J Comput High Educ 2023(8):1–26. https://doi.org/10.1007/s12528-023-09382-1

Lee WJ, Sok P, Mao S (2022) When and why does competitive psychological climate affect employee engagement and burnout? J Vocat Behav 139(1):103810. https://doi.org/10.1016/j.jvb.2022.103810

Mandari H, Koloseni D, Mahunnah M (2023) Self-regulation and age perspectives on MOOCS adoption in Tanzanian higher learning Institutions: the role of technology, user, and environmental factors. Educ Inf Technol 2023(12):1–18. https://doi.org/10.1007/s10639-023-12318-y

Martinez RR, Ellis JM (2023) A national study exploring factors promoting adolescent college readiness in math and science (STEM-CR). Educ Res 52(9):553–569. https://doi.org/10.3102/0013189X231193309

Mccarthy S, Kahma N, Lynch L et al. (2021) Open e-learning platforms and the design–reality gap: an affordance theory perspective. Inf Technol People 35(8):74–98. https://doi.org/10.1108/ITP-06-2021-0501

Mourdi Y, Sadgal M, Elabdallaoui HE et al. (2023) A recurrent neural networks based framework for at-risk learners' early prediction and MOOC tutor's decision support. Comput Appl Eng Educ 31(2):270–284. https://doi.org/10.1002/cae.22582

Mubarak AA, Cao H, Hezam IM et al. (2022) Modeling students' performance using graph convolutional networks. Complex Intell Syst 8(1):2183–2201. https://doi.org/10.1007/s40747-022-00647-3

Ortiz-Lozano JM, Aparicio-Chueca P, Triadó-Ivern XM et al. (2023) Early dropout predictors in social sciences and management degree students. High Educ 2023(9):1–14. https://doi.org/10.1080/03075079.2023.2264343

Parviainen M, Aunola K, Torppa M et al. (2020) Symptoms of psychological ill-being and school dropout intentions among upper secondary education students: a person-centered approach. Learn Individ Differ 80(1):101853. https://doi.org/10.1016/j.lindif.2020.101853

Primario S, Rippa P, Secundo G (2022) Rethinking entrepreneurial education: the role of digital technologies to assess entrepreneurial self-efficacy and intention of STEM students. IEEE Trans Eng Manag 2022(8):1–14. https://doi.org/10.1109/TEM.2022.3199709

Rahimi AR (2023) The role of EFL learners' L2 self-identities, and authenticity gap on their intention to continue LMOOCs: insights from an exploratory partial least approach. Comput Assist Lang Learn 2023(4):1–10. https://doi.org/10.1080/09588221.2023.2202215

Ramadhan A, Warnars HLHS, Razak FHA (2023) Combining intelligent tutoring systems and gamification: a systematic literature review Educ Inf Technol 2023(8):1–37. https://doi.org/10.1007/s10639-023-12092-x

Santos AC, Arriaga P, Daniel JR et al. (2023) Social and emotional competencies as predictors of student engagement in youth: a cross-cultural multilevel study. Stud High Educ 48(1):1–19. https://doi.org/10.1080/03075079.2022.2099370

Taranto E, Jablonski S, Recio T et al. (2021) Professional development in mathematics education—evaluation of a MOOC on outdoor mathematics. Mathematics 9(22):2975. https://doi.org/10.3390/math9222975

Wei X, Saab N, Admiraal WJ et al. (2022) Do learners share the same perceived learning outcomes in MOOCs? Identifying the role of motivation, perceived learning support, learning engagement, and self-regulated learning strategies. Internet High Educ 56(1):100880. https://doi.org/10.1016/j.iheduc.2022.100880

Weiss AK, Quinn SM, Danley AL et al. (2023) Burnout and perceptions of stigma and help-seeking behavior among pediatric fellows. Pediatrics 148(4):1–8. https://doi.org/10.1542/peds.2021-050393

Weston TJ, Laursen SL, Hayward CN (2023) Measures of success: characterizing teaching and teaching change with segmented and holistic observation data. Int J STEM Educ 10(1):1–21. https://doi.org/10.1186/s40594-023-00413-y

Xia XN (2020a) Learning behavior mining and decision recommendation based on association rules in interactive learning environment. Interact Learn Environ 2020(8):1–16. https://doi.org/10.1080/10494820.2020.1799028

Xia XN (2020b) Random field design and collaborative inference strategies for learning interaction activities. Interact Learn Environ 2020(12):1–25. https://doi.org/10.1080/10494820.2020.1863236

Xia XN (2021a) Sparse learning strategy and key feature selection in interactive learning environment. Interact Learn Environ 2021(11):1–25. https://doi.org/10.1080/10494820.2021.1998913

Xia XN (2021b) Decision application mechanism of regression analysis of multi-category learning behaviors in interactive learning environment. Interact Learn Environ 2021(4):1–14. https://doi.org/10.1080/10494820.2021.1916767

Xia XN (2021c) Interaction recognition and intervention based on context feature fusion of learning behaviors in interactive learning environments. Interact Learn Environ 2021(1):1–19. https://doi.org/10.1080/10494820.2021.1871632

Xia XN (2022) Application technology on collaborative training of interactive learning activities and tendency preference diversion. SAGE Open 12(2):1–15. https://doi.org/10.1177/21582440221093368

Xia XN, Qi WX (2022) Early warning mechanism of interactive learning process based on temporal memory enhancement model. Educ Inf Technol 2022(7):1–22. https://doi.org/10.1007/s10639-022-11206-1

Xia XN, Qi WX (2023a) Dropout prediction and decision feedback supported by multi temporal sequences of learning behavior in MOOCs. Int J Educ Technol High Educ 2023(6):1–24. https://doi.org/10.1186/s41239-023-00400-x

Xia XN, Qi WX (2023b) Learning behavior interest propagation strategy of MOOCs based on multi entity knowledge graph. Educ Inf Technol 2023(3):1–29. https://doi.org/10.1007/s10639-023-11719-3

Xia XN, Qi WX (2023c) Interpretable early warning recommendations in interactive learning environments: a deep-neural network approach based on learning behavior knowledge graph. Humanit Soc Sci Commun 10(1):258. https://doi.org/10.1057/s41599-023-01739-2

Xia XN, Wang TJ (2022) Multi objective evaluation between learning behavior and learning achievement. Asia-Pac Educ Res 2022(12):1–15. https://doi.org/10.1007/s40299-022-00703-z

Xing WL, Chen X, Stein J et al. (2016) Temporal predication of dropouts in MOOCs: reaching the low hanging fruit through stacking generalization. Comput Hum Behav 58(1):119–129. https://doi.org/10.1016/j.chb.2015.12.007

Zhang HJ, Dong JL, Lv C et al (2022) Visual analytics of potential dropout behavior patterns in online learning based on counterfactual explanation. J Vis 2022(11):1–19. https://doi.org/10.1007/s12650-022-00899-8

Zhu M, Sari AR, Lee MM (2022) Trends and issues in MOOC learning analytics empirical research: a systematic literature review (2011–2021). Educ Inf Technol 27(1):1–26. https://doi.org/10.1007/s10639-022-11031-6

Author contributions

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

Informed consent was not required as the study did not involve human participants.

Additional information

Correspondence and requests for materials should be addressed to Xiaona Xia.

Reprints and permission information is available at http://www.nature.com/reprints

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