Research Article

MOOC Dropout Prediction Based on Multidimensional Time-Series Data

Zhaoyu Shou,1 Pan Chen,1 Hui Wen,1 Jinghua Liu,1 and Huibing Zhang2

1School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China
2School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China

Correspondence should be addressed to Hui Wen; huiwen@guet.edu.cn

Received 18 February 2022; Accepted 8 April 2022; Published 28 April 2022

Copyright © 2022 Zhaoyu Shou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Massive open online courses have attracted millions of learners worldwide with flexible learning options. However, online learning differs from offline education in that the lack of communicative feedback is a drawback that magnifies high dropout rates. The analysis and prediction of student’s online learning process can help teachers find the students with dropout tendencies in time and provide additional help. Previous studies have shown that analyzing learning behaviors at different time scales leads to different prediction results. In addition, noise in the time-series data of student behavior can also interfere with the prediction results. To address these issues, we propose a dropout prediction model that combines a multiscale fully convolutional network and a variational information bottleneck. The model extracts multiscale features of student behavior time-series data by constructing a multiscale full convolutional network and then uses a variational information bottleneck to suppress the effect of noise on the prediction results. This study conducted multiple cross-validation experiments on KDD CUP 2015 data set. The results showed that the proposed method achieved the best performance compared to the baseline method.

1. Introduction

With the support of big data technology and artificial intelligence, massive open online courses (MOOCs) have set off a new educational revolution and become a hot spot for cross research in education, psychology, information technology, and data science [1]. The MOOC breaks the limitations of traditional classroom teaching and provides an interactive internet-based platform for learners, teachers, and educational institutions, making learners more flexible in terms of time and place of learning. With the intensification of the COVID-19 epidemic since 2020, many universities have begun to use online learning platforms to continue teaching in response to the epidemic prevention policy of home isolation and reducing the flow of people. During the epidemic, the number of MOOC learners increased dramatically [2]. However, in the process of conducting online learning, the instructor and students are separated in time and space, and it is difficult for the instructor to grasp the learning dynamics of the learners, and most of the students who take online courses do not complete the course. The dropout rate of MOOC courses offered by prestigious universities, including MIT, Stanford University, and UC Berkeley, has been reported as high as 95% [3]. High dropout rates severely constrain the further development of MOOCs [4, 5]. Therefore, effectively reducing the dropout rate of courses is an issue that MOOC builders must address. By analyzing learner’s learning behavior and predicting whether they would drop out of classes, learners with the tendency to drop out can be found in advance. Teachers can implement targeted interventions to maintain student’s motivation and prevent them from dropping out of learning [6].

The MOOC platform stores a large amount of student behavior data, such as student clickstream data, homework submission, achievement information, forum discussion participation, and interactive information [7]. The researchers analyzed these student behavior data for critical factors that affect student motivation to predict whether students would drop out of the course. Clickstream data are
the most extensively used in dropout prediction due to cost and ease [8]. Until now, two main research methods are available for MOOC dropout prediction: traditional machine learning methods and deep learning methods. Traditional machine learning-based dropout prediction methods treat dropout prediction as a simple binary classification problem and use machine learning algorithms such as support vector machines (SVM), decision trees (DT), and logistic regression (LR) to make predictions [9–11]. Traditional machine learning methods rely on manual feature extraction and require significant time and human resource costs, which limit their use in today’s MOOC environment.

With the current development of deep learning technology, some deep learning algorithms have also been used for dropout prediction, such as CNN [12] and LSTM [13], which have greatly improved the prediction effect. Especially, CNN algorithms have become a current research hotspot in the field of deep learning and have achieved success in many fields such as image processing [14], natural language processing [15], and time series classification [16].

CNN algorithms can automatically extract abstract features from the original data and achieve better results without complex feature engineering. Due to its unique advantages, CNN algorithms also have a wide range of applications for MOOC dropout prediction.

However, the application of the traditional CNN method in the prediction also has certain limitations. Firstly, the student behavior time series is a special kind of multidimensional time-series data, which contains the long-term trend of student behavior and has short-term fluctuation characteristics. That is, it will show different characteristics in different time scales. The traditional CNN algorithm can only extract the features of a single time scale. It is difficult for the time series classification task to deal with long-term trends and short-term fluctuations on a single scale, and it is easy to lose important information, thus affecting the classification effect [17]. Secondly, students are easily interfered with by many factors in online learning. Time series data of student behavior contain a large amount of noise, bringing significant challenges to CNN feature extraction.

To overcome the above shortcomings, this study extracts multiscale features of student behavior time series by using multiple parallels Fully Convolutional Network (FCN). Then, the flow of noisy information in the network is suppressed by variational information bottlenecks (VIB), and only the most relevant features of student dropout behavior are retained. Finally, the probability of student dropout is predicted. Thus, a MOOC dropout prediction model based on a multiscalar full convolution network (MFCN) and variational information bottleneck (MFCN-VIB) is proposed to provide adequate support for the MOOC platform’s Sustainable development. The main contributions of this study can be summarized in the following three points.

1. A multiscale feature extraction framework based on student behavior time-series data is proposed, which uses multiscale full convolutional networks to extract features from student behavior time series, avoiding the shortcomings of traditional CNNs that use a single-sized convolutional kernel for feature extraction and increasing the diversity of features.

2. The variational information bottleneck is introduced into a multiscale full convolution network to suppress the influence of irrelevant noise on prediction results and improve the generalization ability of the model.

3. In this study, a comparative experiment was designed and the MFCN-VIB model was compared with other relevant research methods to prove the effectiveness of the proposed algorithm. In addition, we also use input data of different time lengths to verify the early prediction ability of the MFCN-VIB model.

The remainder of the paper is organized as follows: Section 2 presents the relevant research work. In Section 3, the dataset used for the experiments is presented. In Section 4, a MOOC dropout prediction model (MFCN-VIB) based on multiscalar fully convolutional networks and variational information bottlenecks is proposed. Section 5 presents the experimental procedure and the analysis of experimental results to evaluate the performance of the proposed algorithm. Finally, Section 6 concludes the paper and looks ahead.

2. Related Works

2.1. Machine Learning-Based Methods. Traditional MOOC dropout prediction methods are based on machine learning approaches. These methods transform the MOOC dropout problem into a dichotomous problem and use logistic regression, support vector machines, and decision trees to build classification models that can identify the dropout risk of online learners. For example, Goel et al. [18] used data mining techniques to mine the key factors leading to the increase in dropout rate from learner clickstream data and used logistic regression to predict whether learners would drop out of course. Liang et al. [19] used registration features, user features, and course features as classification features and used gradient-boosting decision trees (GBDT) to construct a dropout prediction model that predicted the probability of students dropping out in the next 10 days, of course.

Existing research shows that traditional machine learning methods need additional feature engineering [20]. As a result, the key to research is to figure out how to extract and identify practical feature sets from raw data. Many researchers proposed various feature engineering methods to extract helpful input feature sets based on specific problems. For example, Gelman et al. [21] proposed a nonnegative matrix factorization (NMF) based feature engineering method to extract essential features that affect student dropout behavior from MOOC learning behavior data and discovered that five types of learning behavior features have a significant impact on MOOC dropout behavior. Qiu et al. [22] proposed a MOOC dropout prediction
framework based on feature selection. This method first extracts fine-grained features from clickstream log data, then selects practical features through the integrated feature selection method and inputs them into the logistic regression classification model for dropout prediction. Bote-Lorenzo and Gómez-Sánchez [23] proposed a new feature selection method when analyzing the influence of features of student behavior on student dropout behavior. This method uses similarity of features to explore the key features that lead to the dropout of the student.

### 2.2. Deep Learning-Based Methods

In recent years, with the rapid development of the MOOC platform, the number of online learners has multiplied. The learning behavior data also gradually present the characteristics of high dimension, dynamic, and nonlinear correlation, which brings new challenges to predict dropout. The traditional machine learning algorithm faces bottlenecks, and the prediction method based on deep learning has become a new research hotspot. For example, Wang et al. [24] constructed an end-to-end MOOC dropout prediction model taking advantage of CNN’s ability to automate feature extraction and combining it with recurrent neural networks. Experiments on publicly available datasets show that this method achieves comparable results to the feature engineering method. Feng et al. [25] proposed a CNN-based context-aware feature interaction network (CFIN) MOOC dropout model that incorporates user and course information into the modeling framework to improve dropout prediction performance.

In the MOOC platform, each click operation of the students is accompanied by a timestamp, so the MOOC clickstream data have typical time-series characteristics. Some researchers used student behavior time-series data to construct the dropout prediction model. For example, Wen et al. [26] used the matrix to save the feature information related to the learning behavior and then used CNN to extract the local correlation characteristics of learning behavior, to improve the accuracy of the prediction of MOOC dropouts. Wang and Wang [27] proposed an E-LSTM-based dropout prediction model, and the model weighted the original data according to the time interval so that the model can more effectively use time information for dropout prediction. Fu et al. [28] input the high-dimensional feature vector generated by CNN into the bidirectional long and short-term memory network and combine the static attention mechanism with improving prediction performance. Mubarak et al. [29] proposed a CONV-LSTM model based on convolutional neural networks and long short-term memory networks. The model can automatically extract features from raw data of MOOC and predict whether each student will drop out of course. At the same time, to obtain better prediction performance, the cost-sensitive technique is used in the loss function, which considers the various misclassification costs of false and missing reports. Compared with the benchmark model, their models show better performance.

### 2.3. Variational Information Bottleneck-Based Methods

Information bottleneck theory [30] is an information-theoretic-based approach that aims to reduce the loss of helpful information when compressing the input data. In recent years, the information bottleneck principle has been used for the theoretical understanding and analysis of deep neural networks, using the iterative Blahut-Arimoto algorithm for network optimization [31]. However, this approach is hard to apply in practice.

To solve the practical application problem of information bottleneck, Alemi et al. [32] developed a variational approximation of the information bottleneck target by variational inference, used a neural network to optimize the information bottleneck target, and pointed out that a neural network with variational information bottleneck showed less overfitting and stronger robustness. In recent years, the variational information bottleneck has developed rapidly in deep learning applications. For example, Karimi Mahabadi et al. [33] used a large-scale pretrained language model as a generalized feature extractor. They then used a variational information bottleneck for a specific task to suppress information that is not relevant to the task, thus improving the model’s generalization ability. Gu et al. [34] used a convolutional neural network and a bidirectional gated recurrent unit to extract rich features from the text. They then used a variational information bottleneck to compress the extracted features so that the model focuses on the essential features for sentiment analysis. Qian and Gechter [35] proposed a variational information bottleneck model for precise indoor positioning to solve the overfitting problem caused by the high data dimension in the use of Wifi fingerprints to identify the location of the user. The model uses the variational information bottleneck to compress the dimensionality of the input data to improve generalizability. Liao et al. [36] proposed a new convolutional neural network with variational information bottleneck for P300 EEG signal detection. Experiments show that this method can effectively remove redundant information from the P300 EEG data.

### 3. Dataset

#### 3.1. Problem Definition

In this study, learner’s behavior records over 40 days were used. The first 30 days of behavior records were used as input to the model, and the last 10 days of behavior records were used to determine if the learners had dropped out of the course. If the learner has no behavioral record of learning in the last 10 days, the learner is labeled as a dropout with a 1. Otherwise, the label is 0. Thus, the dropout prediction problem is transformed into a problem of classification.

#### 3.2. Introduction for the Dataset

In this paper, the MFCN-VIB dropout prediction model is validated using the KDD Cup 2015 dataset [12], which comes from the largest MOOC platform in China, “XuetangX.” The dataset records 120542 behavior logs of 79186 students in 39 courses from 2013 to 2014. Each behavior log contains seven student behavior features: Problem, Video, Access, Wiki, Discussion, Navigation, Page_close. The dataset includes behavioral records for 30 days and a label indicating whether learners have dropped out within the fourth ten days. If the learner does
not record learning behavior within the fourth ten days, the learner will be marked as dropping out of the course, labeled 1, otherwise labeled 0.

3.3. Dataset Preprocessing. The clickstream data in the KDD Cup 2015 dataset records student IDs, course IDs, enrollment ID, click events, and occurrence times in detail. The dataset contains two parts: training data and test data. The training data is labeled, while the test data is not labeled. This study randomly selected 40,000 data from the training data according to the students’ enrollment ID as the experimental data. To make full use of the characteristics of students’ behaviors at different time scales, we counted the frequency of students’ seven learning behaviors per day, and the missing items were filled with 0. The learning behavior data were processed into the form of a two-dimensional matrix. The student behavior timing matrix $S_i$ contains data on seven behavioral features of students from day 1 to day $t$, which can be expressed as follows:

$$S_i = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{17} \\ s_{21} & s_{22} & \cdots & s_{27} \\ \vdots & \vdots & \ddots & \vdots \\ s_{71} & s_{72} & \cdots & s_{77} \end{bmatrix},$$

(1)

$S_i$ performs Z-Score normalization before inputting the model, which makes the contour plot of the model’s loss function closer to circular and reduces oscillations in the gradient descent direction, thus speeding up the training process.

4. MFCN-VIB Model

4.1. MFCN-VIB Model Structure. The MOOC dropout prediction model based on MFCN-VIB is divided into two modules: multiscale full convolution network and variational information bottleneck. Firstly, the features of different time scales are automatically extracted from the learning behavior time-series data through multiple parallel full convolution networks. Then the features of different time scales are spliced to obtain the multiscale features of the student behavior time-series data. Then, the variational information bottleneck is used to remove the features unrelated to the dropout behavior from the multiscale features of student behavior and further improve the model’s performance. The structure of the MFCN-VIB model is shown in Figure 1.

4.2. Multiscale Full Convolutional Networks. This module aims to extract multiscale features from learning behavior time-series data through three parallel, fully convolutional networks. As shown in Figure 1, each full convolution network consists of a convolution layer and a global pooling layer, corresponding to a time scale. Specifically, the receptive fields of convolution kernels of different sizes are different. The larger convolution kernel can capture the long-term behavior trend of students, while the smaller convolution kernel is more sensitive to the short-term fluctuation of student’s behavior. Multiscale features of learning behavior data can be obtained by fusing the features extracted from the full convolution network with different kernels.

The input data $S_i \in R^{m \times n}$ is a two-dimensional tensor, $t$ is the length of time, and $m$ denotes the kind of learned behavior in the input data. The $j$th convolutional kernel of the $i$th full convolutional network is gradually shifted along the time dimension to produce a new feature sequence $x_{i,j}$ computed by the following equation:

$$x_{i,j} = f(W \ast S_i + b).$$

(2)

In equation (2), $W$ is the parameter matrix of the convolution kernel, $b$ is the bias term, $f$ is the sigmoid activation function, and $\ast$ represents the 1D convolution operation.

Unlike ordinary CNNs that use several fully connected layers to obtain a fixed-length feature vector, FCNs use a global pooling layer to connect the high-dimensional feature matrix obtained from the convolutional layers. For the dropout prediction task, the trend of student behavior is the essential feature, so a global average pooling layer is used to process the feature sequence output from the convolutional layer. The global average pooling is denoted as follows:

$$\bar{x}_{i,j} = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}.$$  

(3)

In equation (3), $x_{i,j}^l$ denotes the $l$th feature value in the feature sequence $x_{i,j}$, $\bar{x}_{i,j}$ denotes the global average of the feature sequence $x_{i,j}$, and $n$ denotes the number of features in the feature sequence. Then the output features of the $k$ convolutional kernels of the $i$th full convolutional network are combined to obtain the output feature $x_i$ of the $i$th full convolutional network, expressed as follows:

$$x_i = [\bar{x}_{i,1}, \ldots, \bar{x}_{i,j}, \ldots, \bar{x}_{i,k}].$$

(4)

Finally, the features extracted by each full convolutional network are concatenated to obtain multiscale features of learning behavior timing data, expressed as follows:

$$X = [x_1, x_2, x_3].$$

(5)

Compared with the features learned from the learning behavior time-series data on a single scale, the multiscale features contain complementary and rich behavior features on multiple time scales, making the dropout prediction more accurate.

4.3. Variational Information Bottleneck. The primary function of this module is to retain helpful information and remove task-irrelevant noise by compressing the feature sequence $X$. The basic idea of the information bottleneck is that if the input data $x \in X$ and the output data $y \in Y$, assuming the intermediate variable $z \in Z$, the information related to $Y$ in $X$ is retained as much as possible in compressing $X$ into $Z$. To minimize the mutual information
\( I(X, Z) \) and maximize the mutual information \( I(Z, Y) \), the method can be reduced to a minimization problem by introducing the Lagrange multiplier \( \beta \), i.e.,
\[
\min -I(Z, Y) + \beta I(Z, X).
\]

The mutual information \( I(Z, Y) \) and \( I(Z, X) \) are calculated as follows:
\[
I(Z, Y) = \int p(z) \log \frac{p(y|z)}{p(y)} dz dy.
\]
\[
I(Z, X) = \int p(z, x) \log \frac{p(z|x)}{p(z)} dx dz.
\]

In practical applications, \( p(y|z) \) and \( p(z) \) in equations (7) and (8) are difficult to be calculated. So, the upper bound of equation (6) can be constructed by variational approximation [32]. Specifically, for \( I(Z, Y) \) in equation (7), let \( q(y|z) \) be the variational approximation of \( p(y|z) \). From the non-negativity of KL divergence, we can obtain the following:
\[
KL[p(y|z)\|q(y|z)] \geq 0 \Rightarrow \int p(y|z) \log q(y|z) dy dz \geq \int p(y|z) \log q(y|z) dy dz.
\]

Combining equation (9) with equation (7), we can obtain the following:
\[
I(Z, Y) \geq \int p(z) \log \frac{q(y|z)}{p(y)} dy dz
\]
\[
= \int p(z, y) \log q(y|z) dy dz - \int p(y) \log p(y) dy.
\]
\[
= \int p(z, y) \log q(y|z) dy dz + H(Y),
\]
where \( H(Y) = -\int p(y) \log p(y) dy \) is independent of the optimization, it can be neglected. According to the Markov hypothesis of joint distribution \( p(X, Y, Z) \) in reference [32], which is \( p(Z|X, Y) = P(Z|X) \), corresponding to the Markov chain \( Y \leftrightarrow X \leftrightarrow Z \), we have \( p(z, y) = \int p(x)p(y|x)p(z|x) dx \). Substituting equation (11) can get the lower bound of \( I(Z, Y) \):
\[
I(Z, Y) \geq \int p(x)p(y|x)p(z|x) \log q(y|z) dx dy dz.
\]

For \( I(Z, X) \) defined in equations (8), let \( r(z) \) be a variational approximation to \( p(z) \). Similarly, using the non-negativity of KL divergence, we can obtain the following:
\[
KL[p(z)\|r(z)] \geq 0 \Rightarrow \int p(z) \log p(z) dz \geq \int p(z) \log r(z) dz.
\]

Combining equation (13) with equation (8), the upper bound of \( I(Z, X) \) can be obtained as follows:
\[
I(Z, X) \leq \int p(x)KL[p(z|x)\|r(z)] dx.
\]

Combination equations (12) and (14) can get the upper bound of formula (6):
\[
-L - \beta I(Z, X) \leq L
\]
\[
L = -\int p(x)p(y|x)p(z|x) \log q(y|z) dx dy dz.
\]
\[-\beta \int p(x)KL[p(z|x)\|r(z)] dx.
\]

Thus, the final optimization objective translates into minimizing the upper bound \( L \). In the actual calculation, we can approximate \( p(x, y) = p(x)p(y|x) \) by the empirical data distribution \( p(x, y) = (1/N) \sum_{n=1}^{N} \delta_{x_n}(x)\delta_{y_n}(y) \), where \( N \) is the number of samples, and the upper bound \( L \) can be rewritten as follows:
\[
L = \frac{-1}{N} \sum_{n=1}^{N} \left[ -\int p(z|x_n) \log q(y_n|z) dz + \beta KL[p(z|x_n)\|r(z)] \right].
\]
The first term in equation (16) is the cross-entropy loss function, and minimizing this part can make the intermediate variable $Z$ contain more information related to $Y$. The second term can be regarded as a regularization term, $\beta$ is an adjustable hyperparameter, and minimizing this term can limit the information transmission between $X$ and the intermediate variable $Z$. $r(z)$ is a standard normal distribution. $p(z|x)$ is an encoder that converts $X$ to $Z$. Let the encoder be in the form of $p(z|x) = N(z|f^e(x), \Sigma^e)$, where $f^e$ is a fully connected network, and it outputs the mean $\mu$ and covariance matrix $\Sigma$ at the same time. We can use the reparameterization trick [37] to make $z = f(x, \epsilon)$, where $f(x, \epsilon)$ is the deterministic function of $x$ and Gaussian random variable $\epsilon$. Because $\epsilon$ is independent of the model parameters, it can be trained using gradient descent.

5. Experimental Results

5.1. Experimental Process. This study proposes a MOOC dropout rate prediction model based on MFCN-VIB. The prediction process based on this model is shown in Figure 2 and is divided into three parts: dataset preprocessing, model prediction, and model evaluation. First, we extracted the clickstream data from the KDD Cup 2015 dataset and processed it into a two-dimensional matrix form as the input data for the model. Then, the MFCN-VIB dropout prediction model automatically extracts the multiscale features of the learning behavior data for dropout prediction. Finally, the performance of the model is evaluated in terms of Precision, Recall, F1-Score, and AUC value.

5.2. Evaluation Indicators. To evaluate the performance of the MFCN-VIB prediction model proposed in this paper. In this study, AUC, Precision, Recall, and F1-Score are used as evaluation metrics, which are commonly used for dropout prediction models [12]. To calculate these evaluation metrics, we define the confusion matrix of the MOOC dropout prediction model as shown in Table 1.

From this, we can calculate the evaluation indexes of the model, and the specific formula of each evaluation index is as follows:

\[
\begin{align*}
\text{Precision} & = \frac{TP}{TP + FP} \\
\text{Recall} & = \frac{TP}{TP + FN} \\
F1 - \text{Score} & = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\end{align*}
\] (17)

Precision indicates the accuracy of samples predicted as dropouts, i.e., the proportion of samples predicted as dropouts that drop out, reflecting the model’s ability to identify dropouts correctly. The Recall is a coverage measure, i.e., the proportion of all dropout samples that are correctly predicted, reflecting the model’s ability to identify all dropouts. For the dropout prediction task, considering only Precision or Recall is not comprehensive. If the model predicts all samples as dropouts, Recall will be 100%, and Precision will be low; conversely, if the model only finds a small number of dropouts, Precision may be high, and Recall will be poor. F1-Score is the harmonic mean of Precision and Recall and is biased towards the smaller one, so F1-Score combines Precision and Recall and can better evaluate the model.

The receiver operating characteristic curve (ROC) took the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR) as the vertical axis. The ROC curve is created by changing the classification threshold, and the AUC value, or area under the curve, is used to assess the classifier’s performance. TPR and FPR have the following calculating formulas:

\[
\begin{align*}
\text{TPR} & = \frac{TP}{TP + FN} \\
\text{FPR} & = \frac{FP}{FP + TN}
\end{align*}
\] (18)

The larger the AUC value, the better the classification performance. The ratio of dropouts to nondropouts in the dataset used in this paper is about 4:1. It is not appropriate to use accuracy to measure the classification ability of the model in the case of category imbalance because accuracy is easily affected by the category ratio. In contrast, one of the most important features of AUC is that it is not affected by the category ratio of the data sample and can measure the performance of the model proposed in this paper more accurately.

5.3. Baseline Model. To validate the advantages of the MFCN-VIB model, this study used five baseline models to perform comparison experiments with the MFCN-VIB model on the same dataset. These five baseline models include Classification And Regression Tree (CART), Naive Bayes (NB), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and Convolutional Neural Network (CNN) models. Among them, the CNN model has the same input data as the MFCN-VIB model, and the inputs of the remaining models are in the form of one-dimensional vectors.

5.4. Experimental Parameters. The experimental environment for this study is shown in Table 2.

In the MFCN-VIB model, the number of convolutional kernels in each of the three fully convolutional networks is 7, and the sizes are 8, 5, and 3. The number of neurons in the two fully connected layers in the variational information bottleneck is 21. The learning rate is 0.001, and the batch size is 128. Cross-entropy is used as the loss function, and the optimizer is Adam, trained 100 times.

The value of the hyperparameter $\beta$ determines the effect of VIB. Figure 3 shows the results of the variation of the AUC value of the model with $\beta$.

It can be seen from Figure 3 that when the value of $\beta$ is 0.001, the AUC value of the MFCN-VIB model achieves the optimal value.
5.5. Experimental Results and Analysis

5.5.1. Dropout Prediction Performance. To accurately describe the differences between the MFCN-VIB algorithm and the baseline algorithm, we apply a $t$-test to test whether there are significant differences ($\alpha = 0.05$) between MFCN-VIB and other baseline models. The $H_0$ hypothesis assumes that the results of the algorithms are equivalent. However, usually due to limited samples, when using experimental estimation methods such as cross-validation, there is a certain degree of overlap in the training sets of different rounds, which makes the test evaluation metrics not independent and can lead to overestimating the probability of the hypothesis establishment. To alleviate this problem, we used the $5 \times 2$ cross-validation method [38]. The mean of the five 2-fold cross-validation results is shown in Table 3, and the optimal values are in bold. Tables 4 and 5 show the results of the $t$-test for F1-Score and AUC value, respectively.

It can be seen from Table 3 that MFCN-VIB has the best values in Precision, F1-Score, and AUC values. Although the Recall of the LDA model is higher than that of the MFCN-VIB model, Precision is the lowest, which indicates that the LDA model is more likely to misclassify nondropout students as dropouts, so the F1 score is lower than MFCN-VIB. It can be seen from Table 4 that the SVM model and the CNN model have no significant difference in F1-Score ($H_0$ hypothesis cannot be rejected) due to the $p$ values $>\alpha$ of the $t$-test. The $t$-test $p$ values $<\alpha$ between other models indicate that the F1-Score of these models is significantly different ($H_0$ hypothesis is rejected). The above analysis shows that...
the \textit{F1}-Score of MFCN-VIB is better than other baseline models.

As can be seen from Table 5, the \textit{t}-test results on AUC show that there is no significant difference between SVM and CNN and between SVM and LDA, while there are significant differences between the rest of the algorithms. The ROC curves of each model are shown in Figure 4, and comparing the area under the ROC curves shows that the AUC value of the MFCN-VIB model is better than the others.

Through the analysis of the above experimental results, it can be concluded that compared with the machine learning method, the MFCN-VIB model can take a two-dimensional time sequence matrix as input, make full use of the time sequence information in the student behavior data, extract more useful features, and realize more accurate dropout prediction. Compared with CNN, MFCN-VIB can fuse features on multiple time scales and use the variational information bottleneck to filter out the most relevant features for student dropout behavior, thus achieving significantly better results than CNN.

From the training time of each algorithm in Table 3, CART, NB, LDA has the lowest time complexity and can complete the training in a few seconds. SVM has the highest time complexity, and CNN and MFCN-VIB are in the middle level. Overall, the time complexity of the six algorithms can meet the requirements of the dropout prediction task and can complete the training in a few minutes.

To further verify the effectiveness of the MFCN-VIB model, in the relevant literature published on MOOC dropout prediction in recent years, this study selected one MOOC dropout prediction model based on machine learning and four MOOC dropout prediction models based on deep learning to compare with MFCN-VIB. The comparison results are shown in Table 6. The five selected models are all based on KDD CUP 2015 data set for experiments.

As seen in Table 6, the F1-Score and AUC value of the MFCN-VIB model proposed in this paper outperformed the previously mentioned machine learning and deep learning methods. The results show that the MFCN-VIB model, which takes student behavior time-series data as model input, considers multiscalar features of student behavior, and combines variational information bottlenecks, can significantly improve the performance of MOOC dropout prediction.

5.5.2. Ablation Experiments. To explore the role of multiscalar feature extraction and variational information bottleneck in the MFCN-VIB model, we performed ablation experiments on the full convolution network (FCN), multiscalar full convolution network (MFCN), and MFCN-VIB model under the same data set and the same experimental parameters. The experimental results are shown in Table 7, and the \textit{t}-test results of the F1-Score and AUC value are shown in Tables 8 and 9.

From Table 8, it can be seen that there is a significant difference in F1-Score between FCN and MFCN-VIB and between MFCN and MFCN-VIB, while there is no significant difference in F1-Score between MFCN-VIB and MFCN. Combined with Table 7, it can be seen that the F1-Score of MFCN is better than those of FCN. MFCN-VIB does not have an advantage over MFCN in terms of F1-Score compared with MFCN. As seen in Table 9, the AUC values of all three models were significantly different. The ROC curves of these three models are shown in Figure 5, and comparing the area under the ROC curve of each model shows that the MFCN-VIB model has a better AUC value than the other two models.

From the above ablation experimental results, it can be found that MFCN can significantly improve the performance of model F1-Score and AUC value compared to FCN. This is due to the ability of MFCN to extract rich and complementary features from student behavioral time-series data using multiple FCN networks, thus enhancing model performance. The variational information bottleneck, on the other hand, can suppress the flow of task-irrelevant noise in the network by imposing a KL regular term on the loss function, allowing the MFCN-VIB model to focus on the features most relevant to the prediction task, further improving the classification performance of the model.

From the perspective of time complexity, the number of MFCN parameters is three times that of FCN, but the training time of MFCN does not increase by three times. This is because MFCN uses parallel FCN for multiscalar feature extraction, which can make full use of the parallel computing ability of GPU, thus accelerating the training process. The training time of the MFCN-VIB model is slightly longer than that of the MFCN model due to the reparameterization trick and the addition of KL regularization in the loss function.

5.5.3. Early Dropout Prediction. In order to test the early prediction ability of MFCN-VIB, this paper compresses the influence of learning behavior data of different time lengths on the prediction ability of MFCN-VIB models. Because MFCN-VIB uses full convolution to construct the model, MFCN-VIB can directly process data with different time lengths without modifying any parameters. We used the first ten days, 20 days, and 30 days of learning behavior data as model input data, and the results obtained are shown in Table 10, and the corresponding ROC curve is shown in Figure 6. It is worth noting that because learners on MOOC can freely choose the learning time, some learners did not produce any learning behavior in the first ten days or the first 20 days of the course. Therefore, when the learning behavior data of 10 days and 20 days are used as the model input, these students who have no learning behavior record are excluded.
As can be seen from Table 10, the MFCN-VIB model performs best when the learned behavior data from the previous 30 days are used as input data. The MFCN-VIB model has the worst classification performance when the time-series matrix of the previous 10 days is used as the input data. This is because there are fewer learning activities related to the course at the beginning of the course, and learners only generate a small amount of learning behavior data on the MOOC platform. The Recall at this time can still reach 0.937 and the Precision is 0.832, indicating that the MFCN-VIB model can still identify most of the dropouts. As can be seen in Figure 6, the area under the ROC curve is the smallest and the AUC value is the lowest at this time. As the course continued, the $F_1$-Score and AUC value increased by 2.8% and 5.1%, respectively, when using the time-series matrix of the first 20 days as input. The results indicate that MFCN-VIB can identify most dropouts even using data from only the early days of the course and can provide a valuable reference for teachers to implement targeted dropout prevention measures.

**Table 4**: Results of $t$-test for $F_1$-score of the MFCN-VIB model and baseline model ($\alpha = 0.05$).

<table>
<thead>
<tr>
<th>Model</th>
<th>MFCN-VIB</th>
<th>CNN</th>
<th>CART</th>
<th>NB</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>$7.453E-03$</td>
<td>$8.511E-06$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CART</td>
<td>$6.974E-06$</td>
<td></td>
<td>$1.724E-04$</td>
<td>$1.955E-03$</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>$2.074E-03$</td>
<td>$1.997E-01$</td>
<td>$1.688E-05$</td>
<td>$7.191E-05$</td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>$1.158E-03$</td>
<td>$3.792E-03$</td>
<td>$2.000E-05$</td>
<td>$3.771E-03$</td>
<td>$4.999E-02$</td>
</tr>
</tbody>
</table>

Bold indicates $p$ values $> \alpha$. $H_0$ hypothesis cannot be rejected.

**Table 5**: Results of $t$-test for AUC of the MFCN-VIB model and baseline model ($\alpha = 0.05$).

<table>
<thead>
<tr>
<th>Model</th>
<th>MFCN-VIB</th>
<th>CNN</th>
<th>CART</th>
<th>NB</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>$6.510E-05$</td>
<td></td>
<td>$2.040E-06$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CART</td>
<td>$1.390E-07$</td>
<td></td>
<td>$3.130E-08$</td>
<td>$1.309E-04$</td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>$8.730E-08$</td>
<td></td>
<td></td>
<td>$1.724E-04$</td>
<td>$1.955E-03$</td>
</tr>
<tr>
<td>SVM</td>
<td>$1.584E-03$</td>
<td>$7.500E-01$</td>
<td>$1.060E-06$</td>
<td>$1.850E-05$</td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>$1.080E-03$</td>
<td>$1.881E-02$</td>
<td>$1.040E-06$</td>
<td>$1.100E-06$</td>
<td>$1.041E-01$</td>
</tr>
</tbody>
</table>

Bold indicates $p$ values $> \alpha$. $H_0$ hypothesis cannot be rejected.

**Figure 4**: ROC curves of MFCN-VIB model vs. baseline model.
6. Conclusions and Future Work

This paper proposes the MFCN-VIB model to predict the behavior of MOOC dropout. This model combines the advantages of multiscale full convolution network and information bottleneck theory and is successfully applied to the field of MOOC dropout prediction. The performance of the MFCN-VIB model was verified by comparing it with the baseline model and ablation experiments. Furthermore, we used input data of different time lengths to test the early prediction ability of the MFCN-VIB model. The experimental results show that MFCN-VIB can also achieve a good prediction effect by using the early learning behavior data of the course.

This study takes the prediction problem of MOOC dropout as the research object. It constructs a prediction model of MOOC dropout based on a multiscale full convolution network and variational information bottleneck using time series data of learning behavior, which can effectively identify dropouts in the MOOC platform. We will improve our current model by using other information besides clickstream data in future research. In addition, we will also study the unsupervised MOOC dropout prediction.

Table 6: Comparison of the results with other models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR [39]</td>
<td>0.857</td>
<td>0.863</td>
<td>0.847</td>
<td>0.863</td>
</tr>
<tr>
<td>FWTS-CNN [40]</td>
<td>0.863</td>
<td>0.865</td>
<td>0.864</td>
<td>—</td>
</tr>
<tr>
<td>DP-CNN [12]</td>
<td>0.864</td>
<td>0.872</td>
<td>0.863</td>
<td>0.868</td>
</tr>
<tr>
<td>CLSA [28]</td>
<td>0.874</td>
<td>0.865</td>
<td>0.869</td>
<td>—</td>
</tr>
<tr>
<td>CONV-LSTM [29]</td>
<td>0.930</td>
<td>0.870</td>
<td>0.900</td>
<td>0.860</td>
</tr>
<tr>
<td>MFCN-VIB</td>
<td>0.887</td>
<td>0.960</td>
<td>0.922</td>
<td>0.872</td>
</tr>
</tbody>
</table>

Table 7: Results of ablation experiments.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>AUC</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>0.876</td>
<td>0.961</td>
<td>0.916</td>
<td>0.848</td>
<td>79</td>
</tr>
<tr>
<td>MFCN</td>
<td>0.883</td>
<td>0.960</td>
<td>0.922</td>
<td>0.862</td>
<td>111</td>
</tr>
<tr>
<td>MFCN-VIB</td>
<td>0.887</td>
<td>0.960</td>
<td>0.922</td>
<td>0.872</td>
<td>133</td>
</tr>
</tbody>
</table>

Table 8: Results of t-test for F1-score of ablation experiment ($\alpha = 0.05$).

<table>
<thead>
<tr>
<th>Model</th>
<th>MFCN-VIB</th>
<th>MFCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCN</td>
<td>9.857E − 02</td>
<td></td>
</tr>
<tr>
<td>FCN</td>
<td>2.090E − 03</td>
<td>1.543E − 02</td>
</tr>
</tbody>
</table>

Bold indicates $p$ values $>\alpha$. $H_0$ hypothesis cannot be rejected.

Table 9: Results of t-test for AUC of ablation experiment ($\alpha = 0.05$).

<table>
<thead>
<tr>
<th>Model</th>
<th>MFCN-VIB</th>
<th>MFCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCN</td>
<td>1.419E − 03</td>
<td></td>
</tr>
<tr>
<td>FCN</td>
<td>3.531E − 05</td>
<td>2.175E − 04</td>
</tr>
</tbody>
</table>

Figure 5: ROC curves of MFCN-VIB, MFCN, and FCN.

Figure 6: ROC curves of the MFCN-VIB model with different input data.

Table 10: Comparison of results with different input data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCN-VIB with 10 days data</td>
<td>0.832</td>
<td>0.937</td>
<td>0.881</td>
<td>0.798</td>
</tr>
<tr>
<td>MFCN-VIB with 20 days data</td>
<td>0.871</td>
<td>0.950</td>
<td>0.909</td>
<td>0.849</td>
</tr>
<tr>
<td>MFCN-VIB with 30 days data</td>
<td>0.887</td>
<td>0.960</td>
<td>0.922</td>
<td>0.872</td>
</tr>
</tbody>
</table>
method to achieve effective dropout prediction for new courses or courses on other platforms without training data.

**Data Availability**

The data used can be accessed at https://data-mining.philippe-fournier-viger.com/the-kddcup-2015-dataset-download-link/.

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**Acknowledgments**

This research was funded by the National Natural Science Foundation of China (62177012 and 61967005), Innovation Project of GUET Graduate Education (2020YCXS022 and 2021YCXS033), and The Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education (CRKL190107).

**References**


