IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 August 2023, accepted 2 October 2023, date of publication 10 October 2023, date of current version 17 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3323365

APPLIED RESEARCH

A University Student Performance Prediction Model and Experiment Based on Multi-Feature Fusion and Attention Mechanism

DAOZONG SUN, RONGXIN LUO, QI GUO, JIAXING XIE, HONGSHAN LIU, SHILEI LYU, XIUYUN XUE, ZHEN LI, AND SHURAN SONG
School of Electronic Engineering, College of Artificial Intelligence, South China Agricultural University, Guangzhou 510642, China

Corresponding author: Shuran Song (songshuran@scau.edu.cn)

This work was supported in part by the Guangdong Province Curriculum Ideological and Political Reform Demonstration Project (Yue Jiao Gao Han [2021] No. 21; Yue Jiao Gao Han [2023] No. 14).

ABSTRACT Predicting student performance is a crucial research area in educational data mining. Student grades are influenced by various factors such as past academic performance, family background, and personal achievements. However, it is difficult to collect many relevant factors other than past academic performance, and it is not ideal to predict student performance using past academic records alone. Therefore, reducing the difficulty of data collection while maintaining timeliness and prediction accuracy remains a pressing issue. To address this challenge, this study proposes a novel model for predicting university student performance based on multi-feature fusion and attention mechanisms. The proposed model focuses on analyzing historical academic grades from multiple dimensions among university students to extract features that reveal relationships between courses and students, among different students themselves, or even between courses themselves. Additionally, an attention mechanism is introduced to explore the relationship between different dimensional features. This study collected a triplet set of related courses and students' real historical grades, proved the correlation between courses through data analysis, and verified the effectiveness of different dimensional features. Experimental results show that, compared with traditional machine learning methods, our proposed method achieves better prediction accuracy with a precision rate of 72.5%.

INDEX TERMS Education data mining, student performance prediction, multi-feature fusion, attention mechanism.

I. INTRODUCTION

With the increasing emphasis on education by the nation and the widespread application of large-scale databases in the field of education, many researchers have applied data mining and machine-learning techniques to the realm of education. Educational Data Mining (EDM) is a research area of data mining aimed at extracting correlations between different features from massive amounts of data generated by teaching activities [1], [2], [3], [4] and uncovering patterns, trends, and associations within it. This endeavor seeks to provide valuable insights for enhancing educational decision making, student learning outcomes, and instructional effectiveness.

Grade prediction is a key area of focus in the EDM. Grades are a direct indicator of student learning and have a direct impact on students' ability to successfully complete their studies. Simultaneously, teachers can use grades to understand their learning situations. Therefore, grade prediction plays a crucial role in teaching management and academic warning systems.

In recent years, there has been increasing attention from scholars in research on grade prediction for students. Several relevant studies have also emerged. For example, Ramesh et al. [5] collected information such as students' gender, place of residence, and parents' occupation to explore

The associate editor coordinating the review of this manuscript and approving it for publication was Yonghong Peng.

the influence of different factors on final exam scores. Baradwaj et al. [6] in order to predict students' performance at the end of the semester, collected information such as student test scores, classroom tests, seminars, and assignment scores. They used decision tree models to classify and predict students' performance. Some scholars also used past grades alone to predict students' future grades [7].

Although previous research has achieved good results, two problems remain. First, similar to the study by Ramesh, it is difficult to collect relevant personal information from students and involves issues of personal privacy and security, making it difficult for widespread application. Second, there is a lag in predicting grades using students' learning performance information during the learning process [8]. Although using historical grades can achieve early prediction, the lack of real-time factors hinders accurate predictions. Reducing the difficulty of data collection so that the model can be widely applied with good timeliness and predictive accuracy remains an urgent problem that needs to be solved.

To address these issues, this paper proposes a university student performance prediction model (MFAPM) based on multi-feature fusion and attention mechanisms. It focuses on analyzing students' historical grades from multiple dimensions and explores the relationship features between courses and courses, students and students, and courses and students through methods such as representation learning and collaborative filtering. The attention mechanism was used to uncover the relationships between different dimensional features. Finally, the predicted grades of the courses are outputted in the prediction layer. The proposed method was extensively tested on the collected datasets, and the experimental results demonstrated its effectiveness. The novel contributions of this study are summarized as follows.

- We collected and constructed a dataset consisting of real student historical performance and course triplets and validated the effectiveness of the model on this dataset.
- By integrating features from different dimensions, we developed a model based on multi-feature fusion and attention mechanism. This model can effectively extract feature information from historical performance data and predict student grades.
- Focusing on students' historical performance reduces the difficulty of data collection and avoids the lag in grade prediction. As a result, they have a wide range of applications.

The remainder of this paper is organized as follows. Section II discusses related research. Section III details the data involved in the experiments and the methods used to obtain them and provides a sound analysis of the data. Section IV presents the proposed performance prediction method. Section V presents and analyzes the specific results of the experiments. Section VI provides a discussion of the work of this study. Finally, Section VII concludes the paper.

II. RELATED WORKS

Student performance plays a crucial role in assessing their learning status. It not only helps students with academic planning and setting reasonable goals, but also enables them to intervene early for students who may face demic difficulties and provide personalized guidance and support. In most universities, student performance is also one of the key factors determining whether they can successfully graduate [9], [10]. Accurate prediction of student performance is of significant importance in improving teaching quality and instructional management efficiency for teachers. The task types for grade prediction can be divided into regression tasks, which aim to predict specific numerical grades, and classification tasks based on grade-level divisions.

According to the different data types used in the prediction of grades, the types can be divided into online, behavioral, and academic data. Online data refers to data generated by students on online learning platforms. Yang et al. [11] proposed a random forest method based on MOOC data improvement, defined a hybrid indicator to measure the importance of features, and established rules for feature selection. Behavioral data refers to behavioral information generated by students during the learning process, which is related or indirectly related to grades. Yao et al. [12] collected behavioral data recorded on students' campus cards and proved that factors such as diligence, regularity, and sleep patterns are closely related to their grades. They also proposed a multitask prediction framework based on ranking learning algorithms to predict students' academic performance. Lian et al. [13] studied the impact of students' borrowing behavior in libraries on their academic performance and proposed a supervised content-aware matrix factorization method for grade prediction. They combined it with a library recommendation system to improve the quality of the book recommendations. Xu et al. [14] extracted real Internet usage data from students, including Internet access time, traffic volume, and connection frequency, using decision trees, neural networks, and support vector machines–three common machine learning algorithms–for predicting academic performance. Academic data refers to information on students' performance in assignments, quizzes, exams, and GPA during their learning process. Gedeon et al. [15] used a feedforward neural network approach to predict students' final exam scores based on their performance in experiments, assignments, and quizzes during the teaching process. Al-Barrak et al. [16] utilized a decision tree algorithm to predict students' final GPA based on their previous course grades. Marbouti et al. [17] collected homework, quizzes, and midterm exam scores from the first five weeks of students and created a prediction model for the course using feature selection methods and the naive Bayes algorithm. Huang et al. [18] collected comprehensive GPA and grades from four related courses of students to predict the performance in ''Dynamics'' course using methods such as multilayer perceptron neural network, radial basis function neural network, and support vector machine. In addition to the mentioned types of methods above, some scholars use personal information such as gender, birthday, major field of study, residential city, status, and father's employment status etc., Amra et al. [19] collected these information about students then predicted their academic performance using naive Bayes algorithm. The relationship between courses has also been studied. Tsiakmaki et al. [20] used a transfer learning method to explore the relationships between different courses.

Although many methods have been proposed for the performance prediction of students in the above studies, these methods still have some limitations, most of which have difficulty obtaining students' behavioral data and personal information, which is not universally applicable; some studies use academic data, although they can avoid the difficulty of data collection, but there is a certain lag, and most of the methods only predict performance from a single point of view. However, different feature information can be extracted from different perspectives, and different features in the formation may not be used to improve the prediction accuracy.

III. DATASET

A. STUDENT ACHIEVEMENT DATA

Engineering plays an important role in social development and economic development. Currently, the number of undergraduate engineering students far exceeds that of other disciplines [21], making predictions about their academic performance more representative. This study collected real-grade data from students majoring in electronic information engineering at a public university. The data includes 574 students from grades 18 to 20, and after cleaning and organizing the data, it contains a total of 10 professional courses with 5740 grade records. Table 1 shows the courses and the corresponding grade levels included in the dataset.

TABLE 1. Data related to course and grade distribution.

Grade Level	Course
Freshman	advanced mathematics(AM), college physics(CP), circuits(CT)
sophomore	linear algebra(LA), analog electronic technology(AET), digital electronic technology(DET), electromagnetic field and electromagnetic waves(EFEW), signals and systems(SS), digital signal processing(DSP)
junior	high-frequency electronic circuits(HFEC)

High-frequency electronic circuits is a junior course with a high level of difficulty. It has more historically related courses, which can be analyzed better. Therefore, the researching takes high-frequency electronic circuits as the research object and uses their course grades as the prediction target, dividing them into four levels: excellent, good, fair, and poor according to the rules in Table 2.

TABLE 2. Rules for grade classification of course performance.

Classification rules	Grade
90 <= G <= 100	Excellent
80 <= G < 90	Good
60 = G < 80	Medium
0 <= G < 60	Poor

Figure 1 shows the distribution of scores for high-frequency electronic circuits. The majority of students' scores were concentrated between 90-100 points, followed by the range of 80-90 points and 60-80 points. The interval of 0-60 points had the fewest number of students. From the distribution, it can be observed that in real-life situations, score distributions are often uneven. Uneven score data may have an impact on the learning of the prediction models.

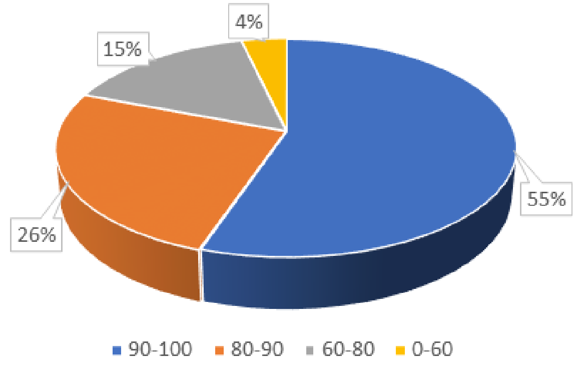


FIGURE 1. The distribution of grades in high-frequency electronic circuits.

To investigate the correlation between the course and other courses in terms of grades, we calculated Spearman's rank correlation coefficient (SRC) between the course and other courses. Figure 2 shows the relationship between the grade distribution and the magnitude of the Spearman's correlation coefficient (SRC) for HFEC, AET, AM, and EFEW. From Figure 2, it can be seen that there are different relationships between different disciplines, and there is a higher correlation between AET and HFEC than between AM and EFEW.

B. KNOWLEDGE POINTS DATASETS

Different courses have different correlations with each other, and these course correlations can provide some support for in the prediction of grades, but how to better mine these relationships is still an open question. This paper uses representation learning approach to mine course relations from the perspective of semantic space.

The data-collection process for representation learning is illustrated in Figure 3. The corpus information related to each

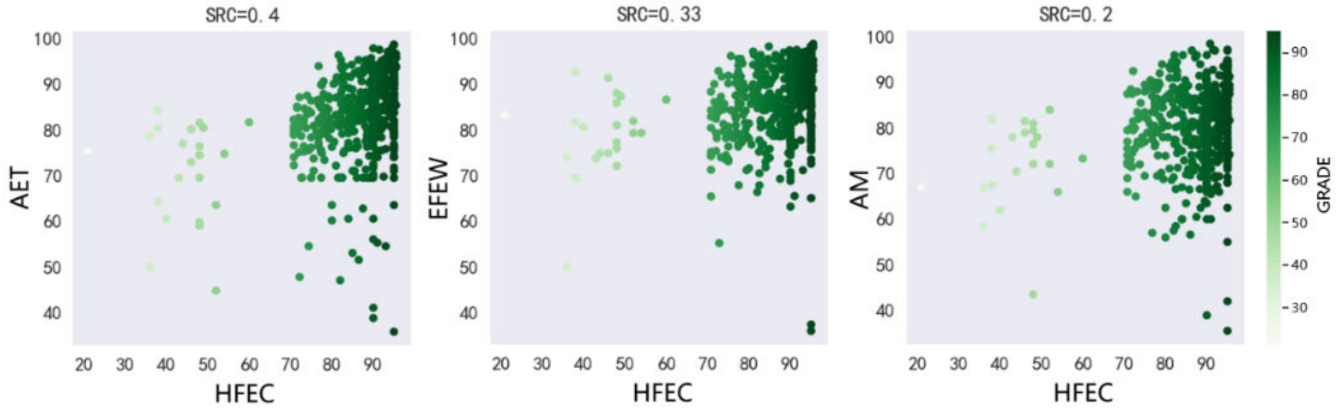


FIGURE 2. Correlation coefficient and distribution relationship between AET, EFEW, AM and HFEC.

course was collected from the textbooks and encyclopedia websites of the course, and the keywords were extracted as the knowledge points of the corresponding course using the TF-IDF keyword extraction algorithm. To exclude meaningless keywords, the extracted knowledge points are cleaned and filtered to avoid affecting the representation effect of the representation model. Finally, the obtained keywords and their corresponding courses comprise a triadic set.

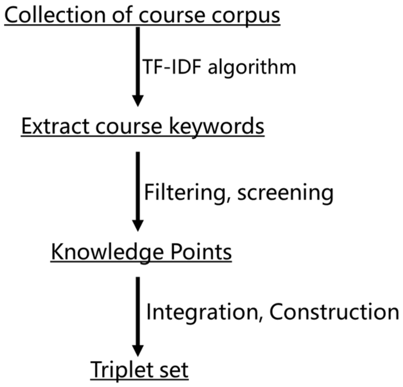


FIGURE 3. The construction process of a triplet dataset.

The composition of the constructed knowledge base is presented in Table 3. The course knowledge base involves 10 courses, with a cumulative total of 12,624 knowledge points across different courses. Using these knowledge points, a set of triplets is constructed, as shown in Figure 4. Different courses may be related to the same knowledge point, and different knowledge points may also be related to the same course. By training representation learning models on the constructed triplet data, embedding vector representations for different courses can be obtained.

IV. PROPOSED METHOD

This study focuses on students' past academic performance, expands feature information from multiple dimensions, and proposes MFAPM. The MFAPM model consists mainly of the following: three layers, including feature extraction layer, attention layer and prediction layer. The feature extraction

TABLE 3. Number of knowledge points included in each course in the triplet set.

Course	Number of knowledge points
advanced mathematics(AM)	1337
linear algebra(LA)	724
college physics(CP)	1016
circuits(CT)	1079
electromagnetic fields and electromagnetic waves(EFEW)	1541
analog electronic technology(AET)	1326
digital electronic technology(DET)	1160
signals and systems(SS)	1303
digital signal processing(DSP)	1333
high-frequency electronic circuits(HFEC)	1805

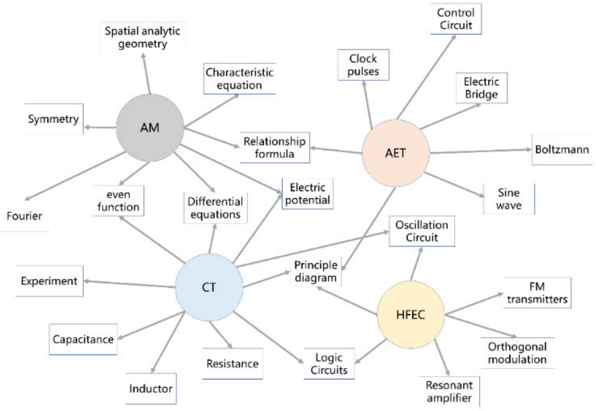


FIGURE 4. Representation of the relationship between knowledge points and courses.

layer is used to extract different dimensional features, the attention layer is used to mine the importance relationship between different features, and the prediction layer is used to predict the results. The framework diagram of the MFAPM is shown in Figure 5.

A. DEFINITIONS AND NOTATIONS

Given a dataset of student grades $\mathrm{G} = \{\mathrm{g}_1, \mathrm{g}_2, \ldots, \mathrm{g}_\mathrm{i}\}$, where gi represents the grade set of the ith student. Let $\mathrm{C} = \{\mathrm{c}_1, \mathrm{c}_2, \ldots, \mathrm{c}_\mathrm{k}\}$ represent the embedding vectors of various courses obtained through learning. The relationship between courses is represented by $\mathrm{Y} \in (\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_\mathrm{k})$. The prediction target of the model is based on the four levels divided according to the above rules, and can be represented as $\mathrm{T} \in (\mathrm{A, B, C, D})$, where A, B, C, and D represent excellent, good, average, and poor, respectively, and are encoded using one-hot encoding.

B. FEATURE EXTRACTION LAYER

The feature extraction layer contains three modules: the ST-CRS module to extract the relationship features between students and courses, the CRS-CRS module to extract the relationship features between courses and courses, and the ST-ST module to extract the relationship features between students and students. Among them, the ST-CRS module extracts students' grades for each course from the dataset to rep the relationship between students and courses. After constructing a triplet set for courses, the CRS-CRS module learns the constructed knowledge base through the TransD [22] model to obtain the low-dimensional real-valued representation vector of each entity. The TransD model introduces a transformation between entity and relationship spaces to better handle complex relationships compared with knowledge representation learning models such as TransE [23] and TransR [24], which have better interpretability. The embedding matrix and scoring function of the TransD model are calculated as follows:

$$M_{rh} = r_p h_p^\top + I^{m \times n} \tag{1}$$

$$M_{rt} = r_p t_p^\top + I^{m \times n} \tag{2}$$

$$h_\perp = M_{rh} h, t_\perp = M_{rt} t \tag{3}$$

$$f_r(h, t) = - \parallel h_\perp + r - t_\perp \parallel_2^2 \tag{4}$$

where h_p, t_p and r_p are the embedding vectors of head entities, tail entities, and relations in entity space, $I^{m \times n}$ is the unit matrix; M_{rh} and M_{rt} are the mapping matrices; $h_\perp$ and $t_\perp$ are the embedding vectors of head entities and tail entities projected into the relation space through the projection matrix; and $f_r(h, t)$ is the scoring function of TransD.

Using the TransD representation learning model, a low-dimensional embedding vector of course entities can be obtained. The embedding vectors of the courses are fused with the corresponding course grades of the students, the Principal Component Analysis (PCA) algorithm [25] is used to reduce the dimensionality of the fused data, and the relationship between the course vectors is mined to provide correlation support for the prediction of students' grades. The processing flow is illustrated in Figure 6.

The collaborative filtering algorithm was introduced into the ST-ST module, which is widely used in recommender systems [26] and also has many applications in the field of grade prediction. Collaborative filtering predicts students' grades by analyzing the similarities between students and the correlation between subjects. Based on this idea, this study uses the collaborative filtering algorithm to mine the relationship between similar students and select the top k students' grades that are most similar to the target student as the basis for prediction, as shown in Algorithm 1.

Algorithm 1 The Pseudocode of Student Collaborative Filtering

Input: 'data_set' - Student grade dataset; 'target_student' - Target student; 'k' - Number of similar students to find
output: 'similar_students' - List of k students most similar to the target student.

```
1. Begin
2. Initialize an empty list 'similarities'.
3. For each 'student' in 'data_set':
       3.1 Calculate the 'similarity' between 'target_student'
           and 'student'.
       3.2 Append 'similarity' to the 'similarities' list.
4. Sort the 'similarities' list in descending order and get the
       top 'k' indices into 'indices'.
5. Initialize an empty list 'sim_stus'.
6. For each index 'i' in 'indices':
       6.1 Get the student ID from 'data_set' at index 'i' and
           append it to 'sim_stus'
7. Set 'similar_students' as the list 'sim_stus'.
8. Return 'similar_students'.
9. end
```

C. ATTENTION LAYER

To explore the relationship between different dimensions of feature, an attention mechanism is introduced to au to extract the importance of different features. Based on the previous content, the study extracted features from three dimensions: the relationship between courses (Y), the relationship between students (sim-stus), and the relationship between students and courses (g). These three-dimensional features are fused into a feature matrix M_{SD}, which is then input into the attention mechanism module. The specific calculation process is as follows.

The query matrix Q, key matrix K, and value matrix V were obtained by mapping the M_{SD} to different representation spaces using linear transformation matrices.

$$Q = M_{SD} * W_q \tag{5}$$

$$K = M_{SD} * W_k \tag{6}$$

$$V = M_{SD} * W_v \tag{7}$$

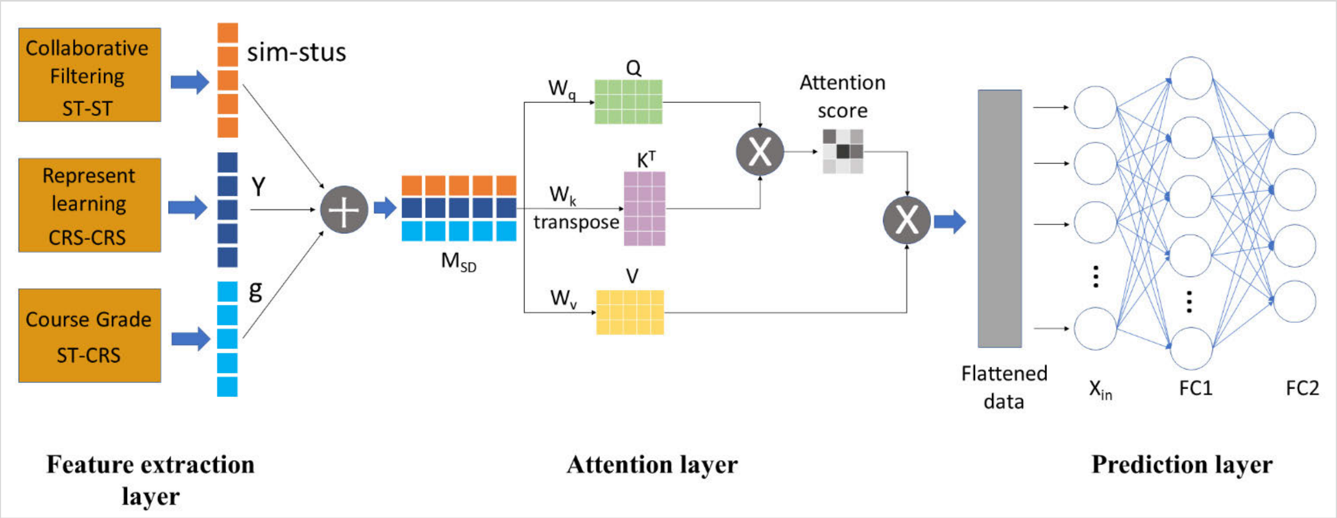
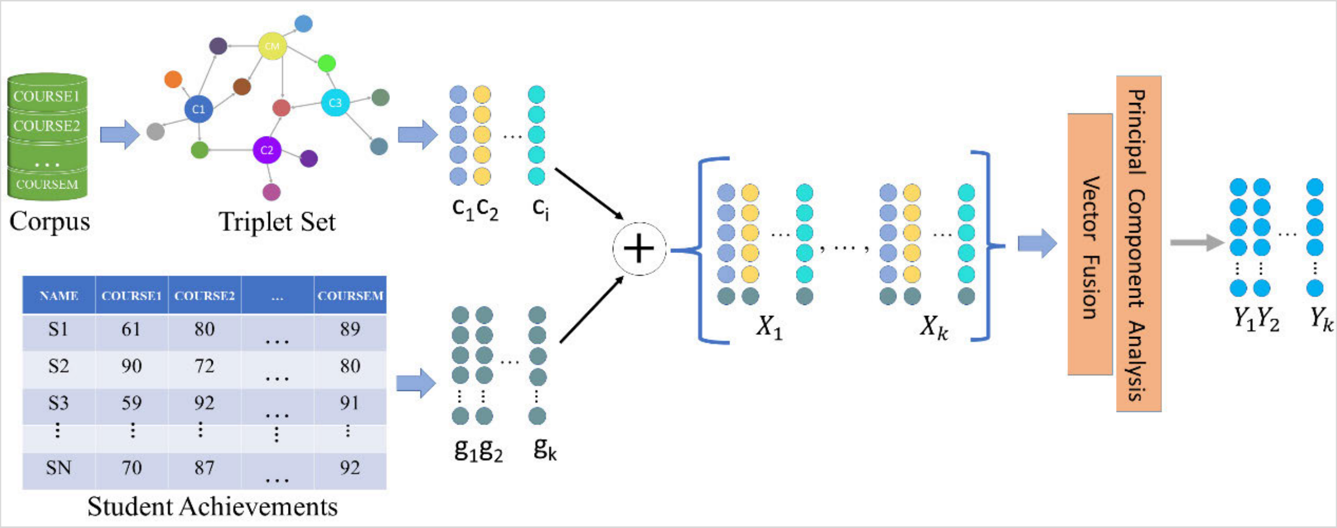


FIGURE 5. MFAPM model framework.



NAME	COURSE1	COURSE2	...	COURSEM
S1	61	80	...	89
S2	90	72	...	80
S3	59	92	...	91
⋮	⋮	⋮	⋮	⋮
SN	70	87	...	92

FIGURE 6. The process of exploring the relationship between courses.

where W_q, W_k and W_v are the projection matrices in the different spaces.

The dot product attention is used to calculate the similarity matrix S (Attention Score) between the query matrix and key matrix, and then normalize S by column to obtain the weight matrix A. Weight matrix A is multiplied with the value matrix V to obtain the feature matrix O weighted by the attention mechanism.

$$S = Q * K^T \tag{8}$$

$$A_{i,j} = softmax(S_{i,j}) = \frac{exp\left(S_{i,j}\right)}{\sum_k^n exp\left(S_{i,k}\right)} \tag{9}$$

$$O = A * V \tag{10}$$

where K^T is the transpose matrix of the key matrix K, n is the number of rows of S, $S_{i,j}$ is the element of the jth row and column i of S, and $A_{i,j}$ is the element of the ith row and column j of A.

D. PREDICTION LAYER

After the output features of the attention layer are obtained, the prediction layer maps them nonlinearly. As shown in Figure 1, this part contains two fully connected layers, FC1 and FC2, and the ReLU is used as the activation function between the two fully connected layers. A back-propagation algorithm was used to train the network model. The classification loss was calculated using cross entropy. The model was optimized using iterative and stochastic gradient descent algorithms to converge the loss function. To prevent the model from overfitting during training, an L2 regularization term is introduced to constrain the parameters. Thus, the final loss function of the model is expressed as follows:

$$L = -\frac{1}{N}\sum\nolimits_{i=1}^{N} y_i log y_i' + \lambda \left|\left|\theta\right|\right|^2 y_i \in T \tag{11}$$

where N is the number of samples in the training set, y_i is the real sample label of the ith student, y_i' is the probability distribution of the predicted ith student, λ is the regular term coefficient, and θ is the set of parameters of the model.

V. EXPERIMENT AND RESULTS

A. EVALUATION INDICATORS

In learning models, the Mean Reciprocal Rank (MRR), Hit@10, and Hit@3 were used to evaluate the effectiveness of embedding corresponding vectors. Among them, Hit@n refers to the proportion of correctly predicted entities among the top n predicted entities. The closer the values of MRR and Hit@n are to 1, the better is the performance of the representation model. The formula for the MMR is as follows:

$$\mathrm{MMR} = \frac{1}{|P|}\sum\nolimits_{i=1}^{|T|} \frac{1}{rank_i} \tag{12}$$

where $|P|$ is the number of triples, and $rank_i$ shows the ranking position of the entity that the model correctly predicts on the ith test triple among all predicted entities, that is, the ranking of the correct entity.

To evaluate the predictive performance of the MFAPM, four evaluation metrics were used: *Accurracy*, *Precision*, *Recall*, and $F1$. The calculation formulas for each evaluation method are as follows:

$$Accurracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{16}$$

where True Positives (TP) represents the true positive examples predicted as positive. False Positives (FP) represent false positive examples that are predicted as positive but actually negative. False Negatives (FN) represent false negative examples that are predicted as negative but actually positive. True Negatives (TN) represent true negative examples that are predicted as negative.

B. EXPERIMENTAL SETUP

The MFAPM model was built based on Python 3.7.12 and PyTorch 1.8.1+cu111 framework, with the optimizer using adaptive moment estimation optimization. To evaluate the model, a 5-fold cross-validation method was used to divide the dataset. The batch size was set to 20, initial learning rate was set to 0.001, and number of iterations was set to 5000. Additionally, the FC1 and FC2 fully connected layers in the model had neuron quantities of 128 and 4, respectively.

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) REPRESENTING LEARNING EXPERIMENT

The TransD representation model was trained based on the constructed triad set, and all data were used for training and testing. In training the model, the dimension of the embedding vector was set to 200, maximum distance γ was set to 10, and number of iterations was set to 2000. Table TABLE 4 shows the evaluation results of the TransD representation model under the constructed dataset for MMR, Hit@10, and Hit@3, where raw data means that negative samples are added to the evaluation data, and filter data means that no negative samples are added to the evaluation data.

TABLE 4. Evaluation results of TransD model on constructing triplet sets.

Data type	MMR	Hit@10	Hit@3
Raw data	0.627827	0.986917	0.741877
Filter data	0.999408	0.999535	0.999349

Based on the evaluation results in Table TABLE 4, it can be observed that the TransD representation model performs well in satisfying the representation of the triad of knowledge points and course composition. By calculating Spearman's coefficients for different course grades, the relationship between courses can be determined from the perspective of the grades. To verify the validity of the course embedding vector, the Manhattan distance of different course entities in the representation space is calculated and normalized to present the correlation between courses. The specific formulae are as follows:

$$MD_{ij} = \sum\nolimits_{k=1}^{N} |x_k - y_k| \tag{17}$$

$$SIMC = 1 - \left(\frac{MD - min\,(MD)}{max\,(MD) - min\,(MD)}\right) \tag{18}$$

where MD_{ij} denotes the Manhattan distance between course embedding vectors C_i and C_j, MD denotes the Manhattan distance between different courses, $MD_{ij} \in MD$, $C_i = \{x_1, x_2, \ldots, x_N\}$, $C_j = \{y_1, y_2, \ldots, y_N\}$, N is the dimension of the embedding vector, $min\,(MD)$ and $max\,(MD)$ denote the maximum and minimum values of the Manhattan distance, respectively.

Figure 7 shows the correlation of course grades based on Spearman's coefficient (SRC) and the correlation of course embedding vectors based on the Manhattan distance (SIMC), both normalized for comparison. The embedding vectors are more objective than the grades; therefore, there may be some differences between the two correlations. However, in terms of relative trends, the correlations between courses were generally consistent. This further proves the effectiveness of using representation learning to explore relationships between courses.

2) COMPARISON EXPERIMENT

In the comparative experiment, the MFAPM model was compared with four machine learning classification methods:

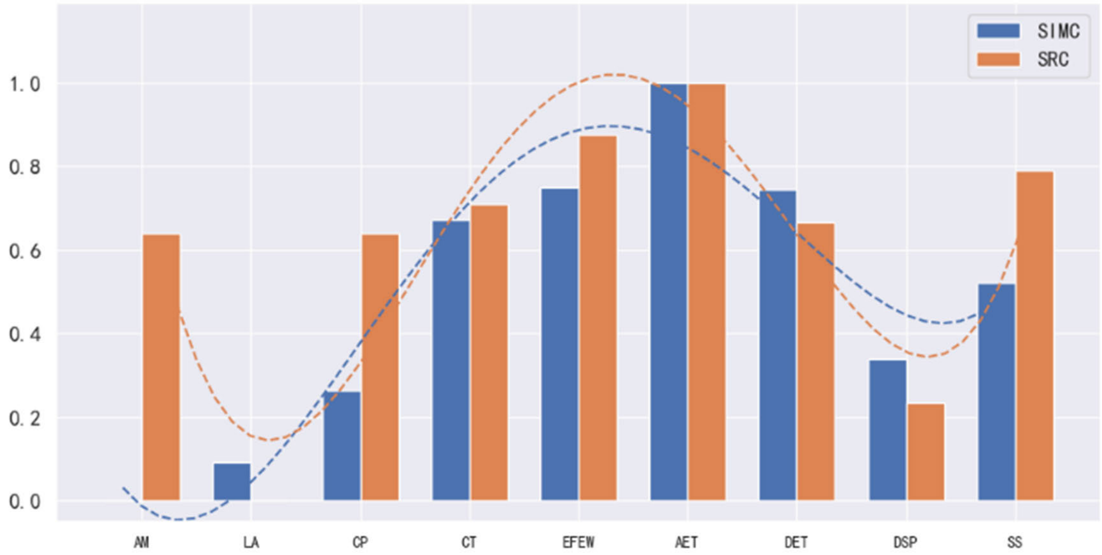


FIGURE 7. Comparison of correlations between courses of different dimensions.

Multilayer Perceptron (MLP), Logistic Regression (LR), Support Vector Machine (SVM), and Classification and Regression Trees (CART) as baseline models. The baseline model uses student-grade data as the dataset, where the kernel function of the SVM is chosen as the Linear Kernel. The multi-layer perceptron contains three layers, with the number of neurons in each layer being 27, 128, and 4, respectively, and the ReLU function is used as the activation function for the hidden layers.

The loss curve of the MFAPM is shown in Figure 8. From the graph, it can be observed that the MFAPM model converged when the number of iterations reached 1000, and the loss value reached its lowest point at 5000 iterations. Table 5 presents the comparative experimental results for the various models. Compared with MLP, LR, SVM, and CART four prediction methods, the proposed MFAPM model achieved better predictive results for all four evaluation metrics for student performance prediction tasks. Accuracy, precision, and Recall reached 72.5%, 82.2%, 73.5%, and F1 score reaches 73.6%, respectively. Compared to the other four comparison models, the MFAPM model performs exceptionally well on these four evaluation metrics, with improvements of 9%, 23.5%,16.6% and 21 .1% respectively, compared to the best baseline model.

TABLE 5. Comparison of experimental results using different methods.

Method	Accurracy	Precision	Recall	F1
MLP	60	41.8	42.0	40.4
LR	63.5	53.2	56.2	51.8
SVM	62.6	58.7	51.1	48.9
CART	60.9	50.1	56.9	52.5
MFAPM	72.5	82.2	73.5	73.6

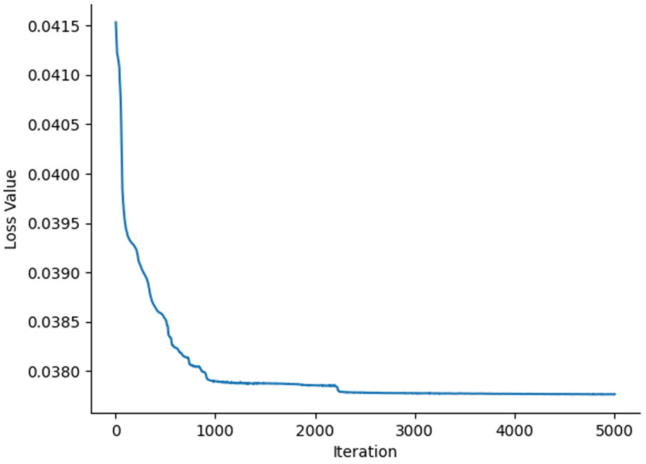


FIGURE 8. The training loss curve of MFAPM model.

3) EXPLORING THE IMPACT OF EACH FEATURE EXTRACTION MODULE

Three feature extraction modules are included in the MFAPM model, and to investigate the effects of different feature extraction modules extracted by the feature layer on the prediction accuracy of the model, ST-CRS is used as the benchmark, and the combination of ST-CRS + CRS-CRS, ST-CRS + ST-ST and ST-CRS + CRS-CRS + ST-ST (MFAPM) feature modules are tested on the prediction accuracy.

The training curves for the different module selection methods are shown in Figure 9. From the figure, it can be observed that using ST-CRS alone yielded the worst performance. However, when ST-CRS was combined with CRS-CRS or ST-ST, there was a significant improvement in the curve. Additionally, the performance of ST-CRS+CRS-CRS

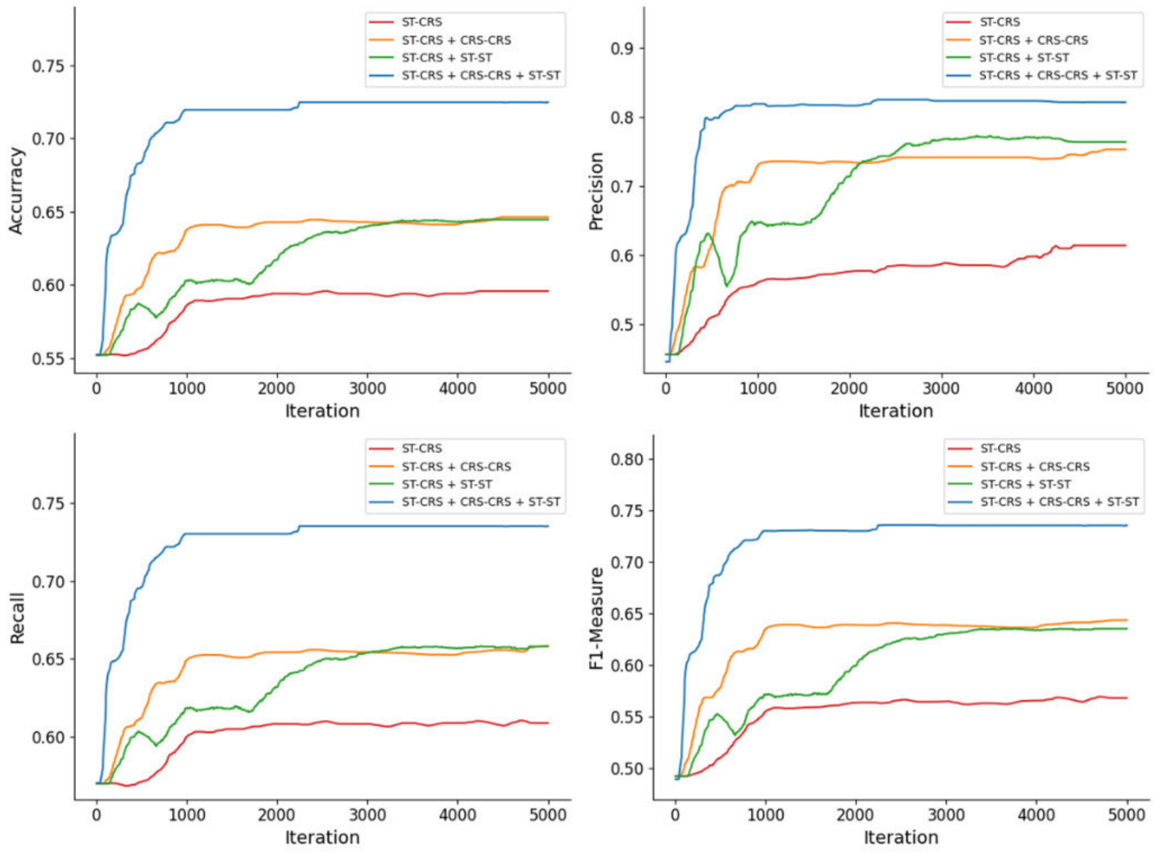


FIGURE 9. Compare the improvement effects of different modules.

TABLE 6. Experimental results of feature module selection %.

Feature Module Combination	Accuracy	Precision	Recall	F1
ST-CRS	59.4	61.4	60.9	56.8
ST-CRS + CRS-CRS	64.6	75.3	65.8	64.4
ST-CRS + ST-ST	64.5	76.4	65.8	63.5
ST-CRS + CRS-CRS + ST-ST	72.5	82.2	73.5	73.6

was similar to that of ST-CRS+ST-ST. The best results were obtained when all three feature modules were integrated into the gather.

From Table 6, it can be observed that although the accuracy of using the ST-CRS module alone is slightly lower than that of the comparative method in Table 5, the precision, recall, and F1 score are higher than those of the other baseline models. At the same time, according to the experimental results, adding CRS-CRS and ST-ST modules on top of the ST-CRS module can effectively improve the predictive ability of the model. The accuracy, precision, recall, and F1 score increased by 13.1%, 20.8%, 12.6%, and 16.8%, respectively. The experimental results demonstrate the effectiveness of each feature extraction module. Combining ST-CRS with other feature modules yields a better predictive performance than using only ST-CRS alone. All three feature extraction modules enhanced the predictive ability of the model.

4) EXPLORING THE IMPACT OF DATA BALANCE ON MODEL

Analyzing the test results in Table 6, it can be found that the accuracy of the MFAPM model is higher than the rest of the three indicators. To explore the reasons for this phenomenon, this study conducted a comparative analysis of the model's prediction results, randomly selected 20% of the data to test the trained model, and analyzed the prediction results in each category using a confusion matrix.

The confusion matrix for the prediction results is shown in Figure 10. By analyzing the confusion matrix, it can be observed that the model achieved the highest prediction accuracy for the classification of excellent grades. However, there are some differences in the prediction results for the classification of cations as good, moderate, and poor. It is worth paying special attention to the fact that the model is more inclined to predict the actual grades as excellent as the wrong prediction results. By comparing the results of the

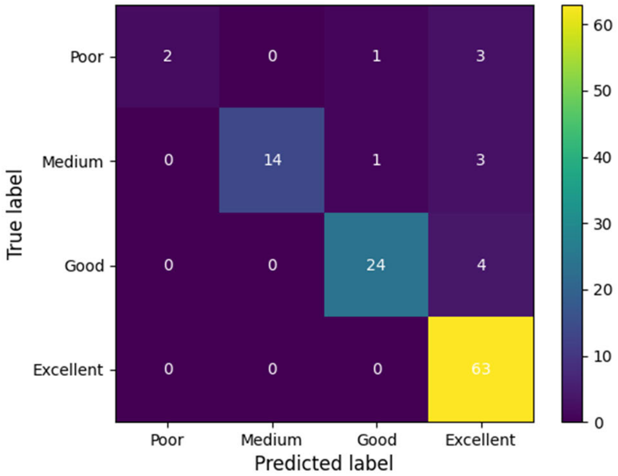


FIGURE 10. Confusion matrix of prediction results for raw data.

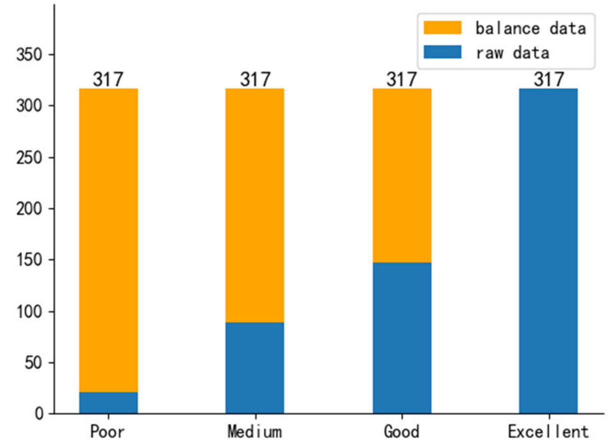


FIGURE 11. Comparison of data volume before and after expansion.

confusion matrices, we can conclude that the model is more powerful in predicting the majority of the sample categories than the minority of the sample categories in an unbalanced dataset. This phenomenon demonstrates the importance of data balancing in improving the model's predictive power across categories.

In order to explore the effect of balanced and unbalanced data on the model, the SMOTE (Synthetic Minority Oversampling Technique) algorithm is introduced in this section to expand the number of samples. SMOTE is a data-generation algorithm for solving the problem of category imbalance, which generates a small number of class samples in the training set, thus balancing the sample distribution. The distribution of the number of samples before and after expansion by setting the number of nearest neighbor samples to two is shown in Figure 11. After introducing the SMOTE algorithm, the model was trained according to the same parameters and evaluation metrics as those in the comparison experiments.

Figure 12 shows the loss curves of training the MFAPM model on the equilibrium and original data. From Figure 12,

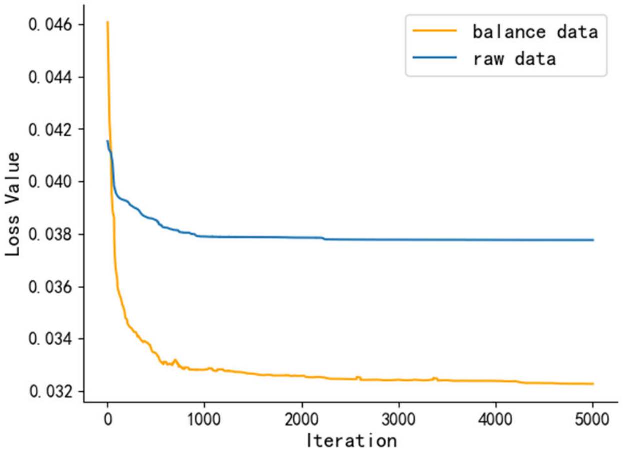


FIGURE 12. Comparison of balanced and raw data training curves.

we can observe that the balanced and original data present obvious differences during the training process. Although the balanced data have a higher initial loss value, they have significantly fewer iterations than the raw data in terms of the number of iterations required to reach convergence. In addition, the model trained with balanced data could fit the data better, exhibiting lower training losses.

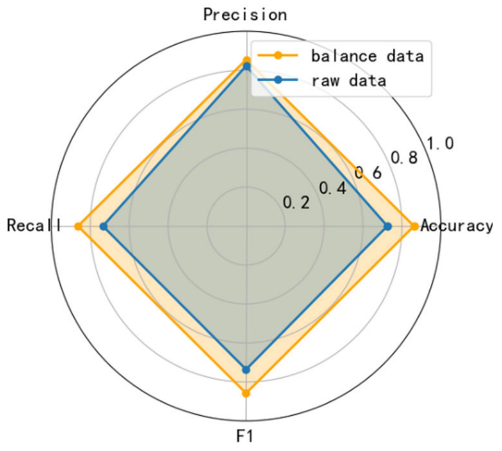


FIGURE 13. Comparison of balanced data and raw data performance.

As shown in Figure 13, the model trained with balanced data performed well in terms of the accuracy, precision, recall, and F1 values, reaching 86.2%, 85.4%, 86.3%, and 85.8%, respectively. This is an improvement of 13.7%, 3.2%, 12.8%, and 12.2%, respectively, compared with the original data, showing the significant ability of balanced data to enhance the predictive effectiveness of the model. In addition, the models trained with balanced data present a more balanced predictive ability, avoiding situations where the accuracy rate is higher than other metrics when using raw data.

Comparing Figures 10 and 14, it can be observed that the models trained with balanced data did not show a greater

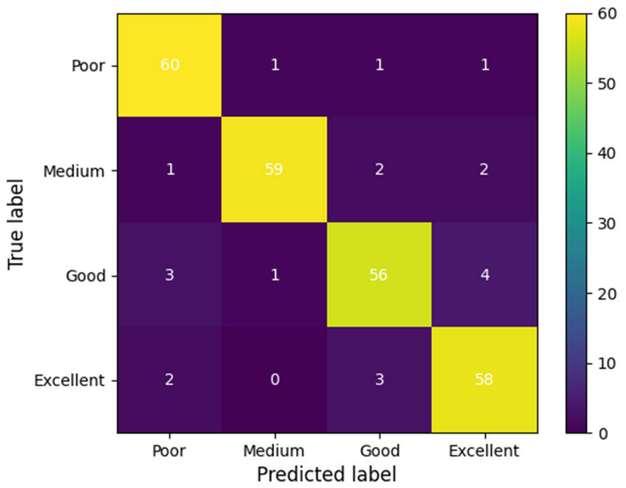


FIGURE 14. Confusion matrix for balanced data prediction results.

tendency to predict a specific category and succeeded in determining the false-positive rate for a few categories. This suggests that balanced data have a positive impact on improving model performance and prediction accuracy.

VI. DISCUSSION

This dissertation aims to explore how to reduce the difficulty of data collection and avoid prediction lag and to investigate the construction of grade prediction models with higher prediction accuracy based on historical grade construction. To this end, a student performance prediction model (MFAPM) based on multi-feature fusion and attention mechanism is proposed, by mining the relationship between courses, the relationship between students, and the relationship between students and courses, and a large number of experiments are carried out on real datasets, and the experimental results show the validity of the MFAPM model proposed in the dissertation, and in comparison with the benchmark model, the MFAPM exhibits better prediction performance. In addition, from the results of the feature-module selection experiments, it was found that the selected relational features contributed to the predictive ability of the model. It is worth noting that in Table 6, the accuracy rates of the model are all better than those of the other three metrics, which may be due to unbalanced educational data [27]. To explore the effect of unbalanced data on the model, the study also introduces the smote algorithm to expand the original data and train the MFAPM model; the experimental results show that training with balanced data can make the model's predictive ability more balanced and have a certain improvement in the model's prediction accuracy. Through extensive experimental validation, this study demonstrated the effectiveness of the model in solving the grade prediction problem. This research provides a useful tool for the field of education, aiming to improve assessment accuracy and improve the field of education.

Although the MFAPM model proposed in this thesis achieved the best prediction results among the comparison models, this study is not without limitations: (1) although the model can predict students' grades before the course starts, which solves the problem of lagging in some courses to a certain extent, it still needs to extract the feature information from the grades of some pre-requisite courses; (2) the factors considered in this study may still not be comprehensive enough, there may be other important character characteristics and factors that have not been considered to have an impact on students' performance. The paper constructed a performance prediction model from multiple dimensions, but in addition to students' performance characteristics, students' demographic characteristics [28], [29], and learning behavior characteristics [30] may be one of the factors affecting the performance of students. In summary, this study has made progress in addressing the problem of practical course evaluation; however, further research and improvements are needed to meet the challenges of different contexts and factors. Therefore, further optimization of multi-feature fusion and attention mechanisms to adapt to different disciplines and teaching environments will be considered in future work. Simultaneously, factors such as behavioral characteristics are combined to improve prediction accuracy, and the influence of factors such as student background on prediction is extended in depth.

VII. CONCLUSION

The relevance of this study is to help student management and teaching in higher education by predicting course grades to warn students academically and adjust teaching strategies. This paper proposes a grade prediction model for university students based on multi-feature fusion and attention mechanism by expanding from multiple dimensions based on the past grades of students, mining relational features using methods such as representation learning and collaborative filtering, and introducing the importance of the attention mechanism to automatically extract features. The study conducted sufficient experiments on real datasets, and the experimental results show that the MFAPM model proposed in this study achieves an accuracy of 72.5%, which has better prediction results than other benchmark models.

REFERENCES

[1] S. Hussain, "Survey on current trends and techniques of data mining research," *Int. J. Res. Advent Technol.*, vol. 7, no. 4, pp. 133–137, Apr. 2019.

[2] R. Paul, S. Gaftandzhieva, S. Kausar, S. Hussain, R. Doneva, and A. K. Baruah, "Exploring student academic performance using data mining tools," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 8, pp. 195–209, Apr. 2020.

[3] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[4] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 3, May 2020, Art. no. e1355.

[5] V. Ramesh, P. Parkavi, and K. Ramar, "Predicting student performance: A statistical and data mining approach," *Int. J. Comput. Appl.*, vol. 63, no. 8, pp. 35–39, Feb. 2013.

[6] B. Kumar Baradwaj and S. Pal, "Mining educational data to analyze students' performance," 2012, *arXiv:1201.3417*.

[7] M. N. R. Ayán and M. T. C. García, "Prediction of university students' academic achievement by linear and logistic models," *Spanish J. Psychol.*, vol. 11, no. 1, pp. 275–288, May 2008.

[8] Y. Ma, C. Cui, J. Yu, J. Guo, G. Yang, and Y. Yin, "Multi-task MIML learning for pre-course student performance prediction," *Frontiers Comput. Sci.*, vol. 14, no. 5, pp. 1–10, Oct. 2020.

[9] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," 2016, *arXiv:1606.06364*.

[10] M. Alban and D. Mauricio, "Predicting university dropout trough data mining: A systematic literature," *Indian J. Sci. Technol.*, vol. 12, no. 4, pp. 1–12, Jan. 2019.

[11] Y. Yang, P. Fu, X. Yang, H. Hong, and D. Zhou, "MOOC learner's final grade prediction based on an improved random forests method," *Comput., Mater. Continua*, vol. 65, no. 3, pp. 2413–2423, 2020.

[12] H. Yao, D. Lian, Y. Cao, Y. Wu, and T. Zhou, "Predicting academic performance for college students: A campus behavior perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 3, pp. 1–21, May 2019.

[13] D. Lian, Y. Ye, W. Zhu, Q. Liu, X. Xie, and H. Xiong, "Mutual reinforcement of academic performance prediction and library book recommendation," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 1023–1028.

[14] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with Internet usage behaviors using machine learning algorithms," *Comput. Hum. Behav.*, vol. 98, pp. 166–173, Sep. 2019.

[15] T. D. Gedeon and H. S. Turner, "Explaining student grades predicted by a neural network," in *Proc. Int. Conf. Neural Netw.*, Nagoya, Japan, Oct. 1993, pp. 609–612.

[16] M. A. Al-Barrak and M. Al-Razgan, "Predicting students final GPA using decision trees: A case study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7, pp. 528–533, 2016.

[17] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Comput. Educ.*, vol. 103, pp. 1–15, Dec. 2016.

[18] S. Huang and N. Fang, "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques," in *Proc. Frontiers Educ. Conf.*, Oct. 2012, pp. 1–2.

[19] I. A. A. Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," in *Proc. 8th Int. Conf. Inf. Technol. (ICIT)*, May 2017, pp. 909–913.

[20] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Transfer learning from deep neural networks for predicting student performance," *Appl. Sci.*, vol. 10, no. 6, p. 2145, Mar. 2020.

[21] J. Cao. (Dec. 29, 2022). *Number of Regular Students for Normal Courses in HEIs by Discipline*. Accessed: Jul. 19, 2023. [Online]. Available: http://www.moe.gov.cn/jyb_sjzl/moe_560/2021/quan-guo/202301/t20230103_1037969.html

[22] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2015, pp. 687–696.

[23] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.

[24] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.

[25] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010.

[26] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, Oct. 2009, Art. no. 421425.

[27] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, and N. A. M. Ghani, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, pp. 95608–95621, 2021.

[28] F. J. Kaunang and R. Rotikan, "Students' academic performance prediction using data mining," in *Proc. 3rd Int. Conf. Informat. Comput. (ICIC)*, Oct. 2018, pp. 1–5, doi: 10.1109/IAC.2018.8780547.

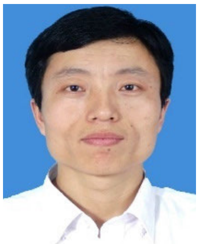
[29] B. K. Yousafzai, S. A. Khan, T. Rahman, I. Khan, I. Ullah, A. Ur Rehman, M. Baz, H. Hamam, and O. Cheikhrouhou, "Student-Performulator: Student academic performance using hybrid deep neural network," *Sustainability*, vol. 13, no. 17, p. 9775, Aug. 2021.

[30] S. Gaftandzhieva, A. Talukder, N. Gohain, S. Hussain, P. Theodorou, Y. K. Salal, and R. Doneva, "Exploring online activities to predict the final grade of student," *Mathematics*, vol. 10, no. 20, p. 3758, Oct. 2022.

DAOZONG SUN received the M.S. and Ph.D. degrees in agricultural electrification and automation from South China Agricultural University, in 2006 and 2013, respectively. He is currently an Associate Professor with the School of Electronic Engineering and the School of Artificial Intelligence, South China Agricultural University. His research interests include digital twins, deep learning, and educational data mining.

RONGXIN LUO received the B.S. degree in information and computer science from the Zhongkai University of Agriculture and Engineering, in 2020. He is currently pursuing the master's degree in artificial intelligence with South China Agricultural University. His current research interests include artificial intelligence and educational data mining.

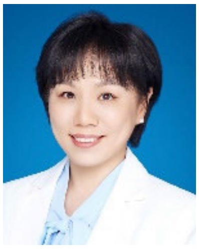
QI GUO received the M.S. degree in electronics and telecommunications engineering from Peking University, in 2011, and the Ph.D. degree in information and telecommunications engineering from Sun Yat-sen University, in 2018. She is currently a Lecturer with the Electronic Engineering College, South China Agricultural University. Her research interests include data mining, computational electromagnetism, and agricultural remote sensing.



JIAXING XIE received the M.S. degree in computer application technology from the Harbin University of Science and Technology, in 2005, and the Ph.D. degree in agricultural electrification and automation from South China Agricultural University, in 2016. Currently, he is a Lecturer with the School of Electronic Engineering and the School of Artificial Intelligence, South China Agricultural University. His research interests include artificial intelligence and IoT electronic technology.

HONGSHAN LIU received the B.S. degree in precision instruments from the National University of Defense Technology, in 1991, the M.S. degree in aerodynamics from the Nanjing University of Aeronautics and Astronautics, in 2001, and the Ph.D. degree in agricultural electrification and automation from South China Agricultural University, in 2016. He is currently an Associate Professor with the Electronic Engineering College, South China Agricultural University. His research interests include teaching and research of microelectronics technology and testing technology.

SHILEI LYU received the B.S. degree in automation from Northeastern University, in 2006, and the Ph.D. degree in radio physics from Sun Yat-sen University, in 2013. He is currently an Associate Professor with the School of Electronic Engineering and the School of Artificial Intelligence, South China Agricultural University. His research interests include artificial intelligence and RFID system applications.



XIUYUN XUE received the M.S. and Ph.D. degrees in agricultural electrification and automation from South China Agricultural University, in 2013 and 2021, respectively. She is currently a Senior Experimenter with the School of Electronic Engineering, College of Artificial Intelligence, South China Agricultural University. Her research interests include intelligent detection and control and intelligent application technology.

ZHEN LI received the M.S. degree in natural science from Oklahoma State University, in 2009, and the Ph.D. degree in agricultural electrification and automation from South China Agricultural University, in 2009. He is currently the Associate Dean and an Associate Professor with the Electronic Engineering College, South China Agricultural University. His research interests include intelligent programming and computer graphics technology.

SHURAN SONG received the M.S. degree in mechanized agriculture and the Ph.D. degree in agricultural electrification and automation from South China Agricultural University, in 2003 and 2012, respectively. She is currently a Professor with the Electronic Engineering College, South China Agricultural University. Her research interests include intelligent information processing and automatic control technology.