An Innovative Hybrid Model of ARIMA and Random Forest for Predicting MOOC Learning Outcomes

Po-Yuan Su[1], Chung-Ho Su[2]*

[1] Department of Digital Content Design, Ling Tung University, Taiwan
[2]Department of Animation and Game Design, Shu-Te University, Taiwan
a0938825858@gmail.com, mic6033@stu.edu.tw

Abstract

With the rapid growth of digital education, MOOCs have emerged as an essential tool for promotiPng lifelong learning among the public. However, pervasive issues of high learner attrition and low completion rates necessitate the establishment of a predictive mechanism capable of forecasting learning risks while providing decision-making support in advance. This study proposes an innovative predictive framework that integrates time series analysis and machine learning to enhance the prediction accuracy and model stability of MOOC learning outcomes. The prediction model process is divided into five stages: data preprocessing, time series forecasting using the ARIMA model, feature engineering and dimensionality reduction through PCA, RF modeling, and model performance evaluation. Initially, the ARIMA model was used to capture the linear trend, followed by RF modeling to fit the nonlinear components of the residuals. The results indicate that the proposed ARIMA+PCA+RF hybrid framework demonstrates significantly superior prediction performance on both the training set (80%) and the testing set (20%). In the final evaluation stage, the hybrid model's performance indicators on the test data are as follows: F1-score = 0.8961, MSE = 1.912, RMSE = 1.3827, MAE = 1.0243, R^2 = 0.9024, clearly outperforming the average performance of traditional models (RA, SVM, ARIMA, RF). Particularly in terms of RMSE and the F1-score, the hybrid model consistently ranks highest among all comparative models, achieving an RMSE of 1.13 and an R^2 of 0.91 in the F1–F14 combination. This hybrid predictive framework enhances the interpretability and predictive accuracy of student performance models, enabling early identification of at-risk learners and facilitating timely intervention.

Keywords: Time series, ARIMA, Random Forest, MOOCs, Learning outcomes

1 Introduction

With the rise of digital education, MOOCs have emerged as a major innovation in global higher education since 2008 [1], facilitating lifelong learning through platforms such as Coursera, edX, and FutureLearn [2]. However, their average completion rate stands at only 5%–15% [3], indicating a significant gap in learning motivation and sustained engagement. Alghamdi et al. [4] emphasized that developing effective behavioral prediction and risk warning systems is crucial for improving course completion rates. Time series and machine learning methods are widely applied to learning prediction tasks. Deep architectures such as Long Short-Term Memory (LSTM) exhibit time series processing capabilities, achieving prediction accuracies of up to 90% [5]. The RF model has also been proven to enhance prediction accuracy in terms of dropout prediction [4]. Moreover, clickstream data exhibit long-term dependence characteristics, where overlooking temporal factors may reduce prediction accuracy [6]. Boroujeni and Dillenbourg [7] indicate that the behavioral logs of MOOC platforms contain numerous timestamps, making them suitable for time series modeling. Khalil and Ebner [8] further emphasized the importance of temporal information for personalized instruction and real-time intervention. Current research trends are moving toward cross-model integration to enhance the time series characteristics of learning behavior data and the ability to handle data heterogeneity.

Kloft et al. [9] utilized SVM to confirm the relationship between behavioral changes and learning outcomes, while Tapio and Tarepe [10] proposed a hybrid model that integrates ARIMA (AutoRegressive Integrated Moving Averag) and RF to accommodate linear and nonlinear characteristics, illustrating greater predictive potential. Although existing studies have employed time series and machine learning for MOOC prediction [11], they primarily utilize single models, failing to integrate data characteristics and modeling advantages. ARIMA excels at capturing trends and seasonal structures [12], but has limitations in processing nonlinear and higher-order interactions. Although RF possesses powerful nonlinear modeling and feature selection capabilities [13], it struggles to understand the cumulative and dependent nature of time series data, potentially leading to the loss of information within behavioral data. Furthermore, current prediction models tend to focus on static classifications (e.g., dropout prediction), failing to capture the evolution and temporal dynamics of learning behaviors [6]. In handling MOOC data, RF performs well with high-

*Corresponding Author: Chung-Ho Su; Email: mic6033@stu.edu.tw
DOI: https://doi.org/10.70003/160792642026092705008

dimensional, heterogeneous, and noisy data, possessing built-in variable importance assessment and robustness [13]. However, single models struggle to address data diversities and structural complexities, necessitating integrative strategies for reinforcement.

Since most current MOOC prediction research relies on single models that struggle to simultaneously capture both time dependencies and nonlinear structures in learning behaviors, this study proposes an innovative hybrid framework that integrates ARIMA and RF to unify time series trends with multidimensional behavioral characteristics, thereby improving the accuracy and stability of learning outcome predictions. The specific research objectives are as follows:

1. To apply the ARIMA model to extract linear trends and cyclical features from learning behavior data.
2. To utilize the RF model to construct nonlinear components within the ARIMA residuals and enhance the model's validity.
3. To integrate ARIMA and RF for constructing a hybrid predictive framework to validate performance advantages in prediction tasks.
4. To analyze the stability and applicability of the model across learning contexts using multiple evaluation indicators.

2 Related Work

2.1 Time Series

As a standard forecasting method, time series analysis focuses on modeling and predicting data generated over time and has been widely applied across fields, including economics, finance, healthcare, and education. Mao et al. [14] noted that educational data often exhibit characteristics of continuous records, such as learning records, course progression, and test results. Data processed by time series modeling can aid in capturing learning trajectories and predicting learning outcomes. As a classic linear time series model, the ARIMA model maintains a prominent role in educational forecasting studies due to its theoretical rigor and strong parameter interpretability [12].

Kontopoulou et al. [14] noted in their literature review on various time series forecasting techniques that ARIMA is suitable for data exhibiting trends and periodic characteristics, demonstrating relatively stable performance in short-term forecasting tasks. In recent years, hybrid models combining ARIMA with machine learning techniques have gained popularity for addressing the limitations of single models in processing educational sequential data. For instance, Tapio and Tarepe [10] proposed an ARIMA-RF hybrid model in their research on student enrollment predictions; ARIMA was utilized to capture the linear trend, followed by RF modeling of the residuals to account for nonlinear fluctuations. Experimental results demonstrate that in terms of forecasting errors (MAPE), this hybrid model outperforms standalone ARIMA and RF models, confirming the effectiveness of the hybrid approach in complex sequential data. In learning analytics, Liu et al. [5] employed the LSTM model to predict students' learning states on MOOC platforms, successfully identifying student learning progress and dropout risks. This finding indicates the high potential of time series models in educational forecasting.

2.2 ARIMA

Since the introduction of the ARIMA model by Box and Jenkins [12], this method has become one of the most representative statistical forecasting techniques in time series analysis. ARIMA integrates Autoregressive (AR), Integrated (I), and Moving Average (MA) components. It can effectively capture trends and periodic variations in stationary linear time series, and is widely employed for data forecasting in economics, energy, and social sciences [12]. In educational research, ARIMA has increasingly been used to model time series data related to student learning journeys, including test scores, assignment completion rates, and course participation behaviors. Kontopoulou et al. [15] further note that while machine learning models exhibit resilience and predictive accuracy on specific unstructured data, the ARIMA model also has certain limitations. Its predictive performance depends on the assumptions of stationarity and linearity in the sequence data; there is a noticeable decline in prediction accuracy when confronted with nonlinear data. Qin et al. [16] found that when taking higher education enrollment rates as an example, ARIMA performs poorly in situations with short historical data, indicating that data length is also a critical factor. Büyükşahin and Ertekin [17] proposed a hybrid architecture that combines ARIMA with Artificial Neural Networks (ANNs). Initially, ARIMA was employed to model the linear trend, and subsequently, ANN was utilized to handle the nonlinear components in the residuals, enhancing prediction accuracy. This strategy can also be extended to predicting learning outcomes in MOOCs. Combining ARIMA's strengths in trend forecasting with machine learning's capabilities in processing nonlinear data can establish a more resilient and effective predictive framework [10].

Thus, the role of ARIMA in hybrid prediction models remains a significant area of research potential, particularly in the application of large and continuous learning data such as those in MOOCs.

2.3 Random Forest (RF)

RF is an ensemble learning model proposed by Breiman [18]. Predictions can be made through multiple randomly generated decision trees, while Bagging techniques are incorporated to enhance stability and reduce the risk of overfitting. Moreover, RF can effectively handle nonlinear relationships and high-dimensional heterogeneous data. It also possesses a built-in variable importance assessment mechanism, offering solid interpretability and practical application value [13]. In financial time series analyses, Tyralis et al. [19] validated that through variable selection and feature expansion, RF can effectively construct short-term prediction models, demonstrating substantial flexibility. In the educational domain, RF has been used extensively for learning behavior analysis and learning outcome prediction. Alamri et al. [20] modeled first-week behavioral data on an MOOC platform to predict whether students would drop out by using only two features, achieving a prediction accuracy of 94%. Another empirical study indicates that RF performs exceptionally well in binary classification tasks for learning outcomes, achieving an accuracy rate of 95.7%, significantly outperforming traditional logistic regression. Tapio and Tarepe [10] proposed a hybrid predictive framework that combines ARIMA for handling linear trend management with RF for modeling the nonlinear structures of residuals. The results show that the hybrid model can significantly enhance overall prediction accuracy.

2.4 MOOCs

With the rapid growth of MOOCs, leveraging the continuous behavioral data generated by learners on these platforms for effective learning outcome predictions has become a key focus in educational technology research. Interaction records of learners on MOOC platforms, such as video viewings, assignment submissions, test scores, and login frequencies, typically exhibit time dependencies, making them suitable for dynamic prediction through time series analysis and machine learning techniques [5]. Liu et al. [5] constructed a prediction model using LSTM networks to estimate students' learning states and potential performances in courses. It achieved a prediction accuracy of 90%, verifying LSTM's effectiveness in capturing time and long-term dependencies in learning behaviors. Jo, Maki, and Tomar (2018) analyzed clickstream log data to confirm that sequence models can identify learning patterns and early learning risks, facilitating the formulation of individualized intervention strategies. Tapio and Tarepe [10] combined ARIMA and RF models for student enrollment prediction and found that the hybrid model can address linear trends and nonlinear fluctuations simultaneously, with overall prediction error (MAPE) outperforming the individual applications of both models.

3 Proposed Method

The process architecture of this study is illustrated in Figure 1. It can be divided into four main stages: data selection, time series modeling with the ARIMA model, dimensionality reduction using Principal Component Analysis (PCA), and prediction with the RF model. First, MOOC learner data is processed, and the time series features are constructed. Then, the ARIMA model is employed to capture learning trends and predict unit scores. PCA is used to simplify static and interactive features. Next, the outputs from the ARIMA model are integrated with other features for RF training to predict learning outcomes (e.g., scores and completion rates). Finally, model comparisons and stability assessments are conducted using multiple evaluation indicators to verify the predictive performance and practical value of the ARIMA+PCA+RF hybrid model.

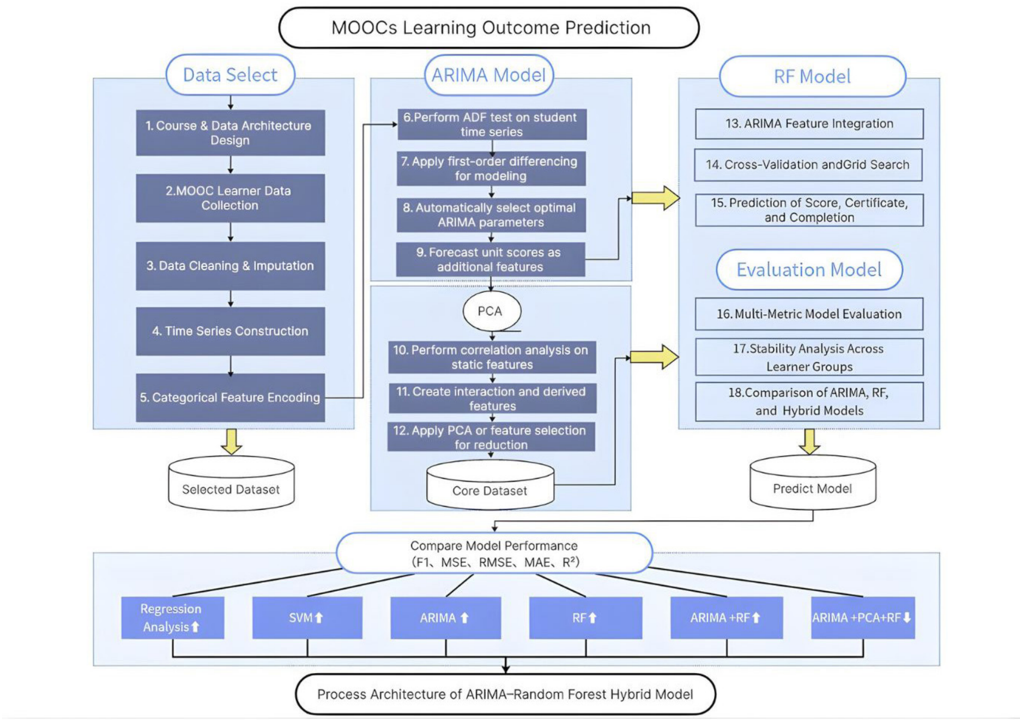


Figure 1. Process architecture of ARIMA–RF hybrid model

3.1 Data Collection and Preprocessing

The data for this study is sourced from the Ewant platform, specifically designed and tracked for specific MOOC courses. The data collection period spans from September 1, 2023, to January 31, 2024, yielding complete learning records for 140 learners, covering 21 columns that include behavior, scores, and background information. According to the research scope, the data was first categorized as time series features (e.g., unit_score) and static learning features (e.g., video_watch_time, assignment_submits, login_frequency, student_sex, student_year). The course was designed in multiple units. Each student submitted assignments sequentially by unit, resulting in time-series unit score data that serves as the core input for the ARIMA prediction model. During data preprocessing, missing and outlier values were checked and processed according to the nature of each column using techniques such as mean, mode, and logical imputation (e.g., forward filling). Subsequently, time series for each student's unit scores (unit_score) were developed based on the unit number (unit_number), facilitating the execution of the Augmented Dickey-Fuller (ADF) test to check for series stationarity. If the series is non-stationary, first-order differencing is applied to render it suitable for ARIMA modeling. This step facilitates the extraction of trends and seasonality components regarding changes in individual students' learning outcomes. For static feature processing, all categorical variables (e.g., gender, device type, background, year in school) were transformed into computable features and encoded using one-hot encoding. This served to retain categorical information and prevent the model from inferring spurious ordinal relationships. After preprocessing, the data was input to the ARIMA model for individualized sequence prediction, and to the RF model as static learning behavior inputs, enabling subsequent hybrid modeling analyses.

3.2 Time Series Prediction of the ARIMA Model

Following data preprocessing, a time series was established for the unit score (unit_score) of each learner. The ARIMA model was employed for individualized predictions. Data were structured into sequences according to unit order and tested for stationarity via the ADF test. If a sequence is non-stationary, first-order differencing is applied to remove trends and variability, ensuring suitability for ARIMA modeling. The ARIMA model requires specification of the autoregressive order (p), differencing order (d), and moving average order (q). This study employs the Python auto_arima function to automatically select the optimal parameter combinations based on AIC and BIC, tailoring prediction models for each student. This method can effectively handle educational time series data with small samples and substantial individual variances while mitigating subjective biases. Upon completing model construction, a one-step forecast was performed to estimate the next unit score. These forecasts, endowed with time dependencies, can supplement static features and furnish the following RF model with more comprehensive input information, thereby enhancing both the overall predictive capacity and individualized identification accuracy of the hybrid model.

The algorithmic steps for the ARIMA model in this stage are as follows:

1. Define the Search Space: Set upper limits for the autoregressive order (p), differencing order (d), and moving average order (q), For instance in Equation (1):

$$p, d, q \in [0, 5] \tag{1}$$

Adjustments can be made according to the actual data structure as required.

2. Differencing order (d) in the ADF Test: Detect the sequence stationarity automatically. If it is non-stationary, differencing is repeatedly applied until the sequence passes the ADF test and the d value is determined.

3. Model Fitting & Evaluation: Perform the ARIMA model fitting for each (p,d,q) combination, and calculate both AIC and BIC indicators show in Equation (2):

$$AIC = 2k - 2\ln(L); BIC = \ln(n)k - 2\ln(L) \tag{2}$$

where k is the number of model parameters, L is the maximum likelihood estimate, and n is the sample size.

4. Select the Optimal Model Combination: Use the minimum AIC or BIC as the basis for model selection to identify the most suitable ARIMA combination for each student.

5. Develop Individual ARIMA Models (Per-Learner Model): Fit a special ARIMA model for each student to reflect individual learning performance changes.

6. Conduct One-Step Forecasting: Use the trained model to predict the next unit score (at the next time point yt+1).

3.3 Feature Engineering and Dimensionality Reduction

Following the completion of time series predictions, this study systematically modeled the static features of learners on MOOC platforms to enhance the predictive performance of the hybrid model. The initial features encompass learning behavior records (e.g., video_watch_time, quiz_score, and login_frequency) and demographic variables (e.g., student_year, student_sex, and student_background), all of which are non-time-dependent static data. First, Pearson's correlation and point-biserial correlation analyses were employed to select columns significantly associated with learning outcomes (e.g., final_learning_score). Redundant and irrelevant variables were eliminated to reduce the noise risk related to high-dimensional data. Next, derived variables grounded in theoretical and empirical foundations were developed, such as the "active learning index" and "interaction density index," to enhance the model's ability to discern behavioral heterogeneity. Finally, dimensionality reduction was conducted using PCA and embedded feature selection. PCA can retain principal components that account for over 85% of the cumulative variance, while the embedded method can remove low-contribution variables based on feature importance from the RF model. This dual approach can effectively simplify the feature structure, enhancing the model's predictive performance and generalizability.

Feature Engineering and Dimensionality Reduction Calculation Formulas:

1. Data Standardization: Z-score standardization is applied to all continuous numerical columns. Their means and standard deviations are adjusted to 0 and 1, respectively, thereby mitigating the influence of scale differences on the results in Equation (3).

$$Z = \frac{x_i - \mu}{\sigma} \tag{3}$$

2. Establishment of the Covariance Matrix

Based on the standardized data X, establish an $n \times n$ covariance matrix in Equation (4):

$$C = \frac{1}{n-1} X^T X \tag{4}$$

3. Eigen Decomposition: Perform eigenvalue and eigenvector decomposition on covariance matrix C in Equation (5):

$$C_U = \lambda_U \tag{5}$$

where λ is the eigenvalue and v is the corresponding eigenvector.

4. Sorting and Principal Component Selection

Sort the eigenvalues by size and select the corresponding eigenvectors (principal components), with the selection criteria being 85% of the cumulative variance explained in Equation (6):

$$\text{Cumulative Variance Explained} = \sum_{i=1}^{k} \frac{\lambda_i}{\sum_{j=i}^{n} \lambda_j} \geq 85 \tag{6}$$

5. Projection

Project the original data onto the newly defined space composed of the selected principal components in Equation (7):

$$Z = XW \tag{7}$$

where W is the transformation matrix comprising the first k principal components.

6. Application to Subsequent Models: Input the transformed reduced-dimensional features (Z) into the RF model as part of the static features to enhance both model performance and generalizability.

3.4 Random Forest Modeling and Prediction

In this stage, the predictions for the next unit scores of individual students from the prior ARIMA model were integrated with the static variables processed through feature engineering. Then, a hybrid input dataset that incorporates both time series dynamics and static learning behaviors was formed as the basis for predictions by the RF model. This integration can enhance the model's grasp of learning behavior heterogeneity and compensate for the limitations of single models in fitting and explanatory power. The RF model is an ensemble learning method that conducts predictions through the construction of multiple decision trees. It has resistance to overfitting, as well as stable generalizability. In this study, 5-fold cross-validation and grid search were used to fine-tune hyperparameters (e.g., n_estimators and max_depth), aiming to ensure model stability across different sample splits. The prediction objectives encompass three categories: final_learning_score (for continuous tasks), certificate_eligible, and course_completion (for binary classification tasks), corresponding to RF regression and classifier modeling, respectively. To address classification imbalance problems, a class weight adjustment strategy was used. The hybrid model constructed in this stage integrates the advantages of statistics and machine learning, thus laying the groundwork for subsequent model performance comparisons.

RF Modeling Formulas:

1. Bagging Sample Generation Formula

From the original training set $D = \{(x_i, y_i)\}_{i=1}^{N}$, B subsample datasets $D(b)$ are generated through random sampling with replacement in Equation (8):

$$D^{(b)} \sim \text{BootstrapSample}(D), b = 1, 2, \ldots, B \tag{8}$$

2. Random Feature Subset Selection

At each node split in every tree, m features (from the total features M) are selected randomly as candidates for the split in Equation (9):

$$\text{Select } m << M \text{ features randomly at each node} \tag{9}$$

3. Classification Prediction Output (Majority Voting)

For classification tasks, each tree Tb provides a predicted category y^b for the input x, and the final prediction is determined by majority voting in Equation (10):

$$\hat{y} = \text{mode}\left(\{T_1(x), T_2(x), \ldots, T_B(x)\}\right) \tag{10}$$

4. Regression Prediction Output (Averaging)

For regression tasks, the predicted value is the average of all trees' predictions in Equation (11):

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{11}$$

5. Feature Importance Calculation (Gini Importance)

If the Gini index is used as a measure of node purity, the importance of feature f is defined as the total contribution of that feature to the reduction of node purity across all trees in Equation (12):

$$\text{Imp}(f) = \sum_{b=1}^{B} \sum_{t \in T_b} 1_{\text{split on} f} \cdot \Delta i(t) \tag{12}$$

where $\Delta i(t)$ is the reduction in node purity after a split at node t.

3.5 Model Evaluation and Stability Test

In this stage, a multi-layered evaluation was conducted on the performance and stability of the hybrid model (ARIMA–RF) in predicting learning outcomes of MOOCs. Appropriate indicators were selected according to three types of prediction tasks. For continuous targets (e.g., final_learning_score), MSE, RMSE, MAE, and R^2 were utilized. For binary classification tasks (e.g., certificate_eligible and course_completion), the F1-score was primarily used to comprehensively evaluate the model's precision and recall performance. To further examine the model's generalizability across learner populations, this study conducted a cluster analysis based on student backgrounds (grade, gender, device) and behavioral characteristics (login frequency) to observe whether prediction performance is biased by population differences, aiming to ensure fairness and stability in the model's educational applications. The hybrid model proposed in this study was compared laterally with two benchmark models (pure ARIMA and pure RF). Analysis results reveal that the hybrid model consistently outperforms other frameworks in terms of accuracy, stability, and overall predictive performance. This confirms the practical application potential and promotional value of the strategy that integrates static and dynamic features.

To ensure theoretical rigor and methodological alignment in the evaluation results, five common performance indicators (see Table 1) were selected from the existing literature: F1-score, MSE, RMSE, MAE, and coefficient of determination (R^2). The F1-score applies to binary classification issues. It can balance precision and recall, with values closer to 1 indicating better classification performance [21]. MSE, RMSE, and MAE are measures of regression error. They can be used to calculate the deviation between predicted and actual values, with smaller error values indicating higher predictive accuracy [22-24]. Additionally, R^2 serves as an indicator of explanatory power, with values closer to 1 indicating the model's higher capacity to account for data variability [25].

Table 1. Indicators description

Indicators	Equation	Range	Reference
F1-score	$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	(0-1)	[21]
MSE	$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$	(≥0)	[22]
RMSE	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$	(≥0)	[23]
MAE	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	(≥0)	[24]
R	$R^2 = 1 - \frac{\Sigma\left(y_i - \hat{y}_i\right)^2}{\Sigma y_i - \hat{y}^2}$	Near 1	[25]

4 Results

4.1 Descriptive Statistics

Table 2 shows that this study analyzed the background data of 140 learners. Regarding gender distribution, there are 87 female (62.14%) and 53 male (37.86%) participants, indicating a female majority in the study sample. In terms of professional backgrounds, most learners are from the design field (45.71%), followed by management (23.57%), information technology (20.71%), and other fields (10.00%). In terms of age distribution, most learners are within the range of 18–20 years old (45.00%), followed by 20–23 years (28.57%) and 15–18 years (22.14%), with fewer participants over 23 years old (4.29%). Table 3 shows that MOOCs dataset Schema description.

Table 2. Descriptive statistics of participants (N = 140)

Variables	Frequency	Percentage (%)
Gender		
Female	87	62.14
Male	53	37.86
Profession		
Others	14	10.00
Information	29	20.71
Management	33	23.57
Design	64	45.71
Age		
15–18	31	22.14
18–20	63	45.00
20–23	40	28.57
Over 23	6	4.29

4.2 Prediction Effectiveness of the ARIMA Model

To enhance the time sequence and individual precision of learning outcome predictions, time series models for the unit score (unit_score) of each student were established; an adaptive ARIMA model was employed to conduct analysis and forecasting. After sorting each student's scores for the last five units, in the initial step, a stationarity check was performed using the ADF test. If it is non-stationary (p-value > 0.05), first-order differencing is applied to convert the data to a stationary series, while a stationary series is modeled directly. Statistics indicate that most models are configured as (1, 0, 1) or (2, 1, 1). The ADF p-values of the data average between 0.03 and 0.07, demonstrating the stability required for time series modeling. To evaluate prediction errors, RMSE was used as the indicator. The overall RMSE was 1.24312, exhibiting stable prediction performance. A key contribution of this stage is providing the individualized future performance trends (ARIMA_score) for each student to serve as a time series feature that supplements the limitations of static behavioral data, thereby enhancing the model's predictive capability regarding future learning behavior tendencies (Table 4).

4.3 Dimensionality Reduction of PCA Eigenvalues

To enhance the prediction performance of the model and reduce feature redundancy, the original 21 static and behavioral features were screened to reduce the dimensionality. Initially, columns with prediction potentials were retained through variable inspection and preprocessing, followed by dimensionality reduction using PCA. PCA selected variables based on the Kaiser criterion (retaining principal components with eigenvalues > 1) and cumulative explained variance. Results indicate that the first five principal components account for approximately 83% of the variability. By screening representative original variables using principal component loading (|loading| > 0.60), the following ten core features were retained: video_watch_time (F1), quiz_score (F2), login_frequency (F3), time_spent_on_platform (F4), assignment_submits (F5), forum_posts (F6), quiz_attempts (F7), video_pause_count (F10), days_since_last_login (F8), and student_year (F9). These features encompass three significant aspects: learning engagement, interaction, and performance, contributing to improving the model's explanatory power and performance. Table 5 presents the Pearson correlation coefficients among the ten core features selected by PCA, exploring the linear relationships between these features. Overall results indicate significant positive correlations among most variables, particularly the highest correlation between video_watch_time (F1) and quiz_score (F2) (r = .780, ***p < .001), highlighting the close relationship between learning engagement and outcomes. Moreover, login_frequency (F3) and time_spent_on_platform (F4) also exhibit a strong correlation (r = .767, ***p < .001), reflecting that students who log in more frequently tend to spend more time learning. In contrast, the correlations of days_since_last_login (F8) and student_year (F9) with other variables are relatively low, suggesting their informational independence and providing a broader feature diversity for the model. Moreover, video_pause_count (F10) shows a moderate significant positive correlation with most variables (e.g., F1–F6), possibly indicating that students engage in active thinking and repetitive learning behaviors while watching videos.

Table 3. Dataset schema of MOOCs learning records

Column Name	Description	Data type	Feature role
student_id	Unique identifier for each student	Categorical	Identifier
unit_number	Sequential number of the learning unit	Ordinal Integer	Sequence Index
unit_score	Unit performance score (0–100)	Continuous	Time-series Input
video_watch_time	Total watch time of videos per unit	Continuous	Behavioral Feature
forum_posts	Number of forum posts per unit	Count Integer	Behavioral Feature
assignment_submits	Number of assignment submissions per unit	Count Integer	Behavioral Feature
login_frequency	Login frequency per unit	Count Integer	Behavioral Feature
video_completion_rate	Percentage of video fully watched	Percentage	Behavioral Feature
time_spent_on_platform	Total time spent on the platform	Continuous	Behavioral Feature
quiz_score	Average quiz score per unit	Continuous	Behavioral Feature
video_pause_count	Number of video pauses	Count Integer	Behavioral Feature
last_login_time	Time interval since last login	Time-based Count	Behavioral Feature
days_since_last_login	Number of days since last login	Count Integer	Behavioral Feature
mobile_or_desktop	Type of device used	Categorical	Background Feature
student_year	Year in school (e.g., 15–22)	Ordinal Integer	Background Feature
student_sex	Gender (1: female, 2: male)	Categorical	Background Feature
student_background	Academic background category	Categorical	Background Feature
certificate_eligible	Whether the student qualified for certificate	Boolean	Prediction Target
course_completion	Course completion status	Boolean	Prediction Target
final_learning_score	Final learning score after all units	Continuous	Prediction Target

Table 4. Summary of ARIMA forecasting results for student unit score sequences

student_id	original_series	ADF_p_value	is_stationary	differencing_order	ARIMA(p,d,q)	ARIMA_score	Error
S001	[70, 71, 72, 73, 74]	0.03	TRUE	0	(1, 0, 1)	75	1.00537
S002	[71, 72, 73, 74, 75]	0.05	FALSE	1	(2, 1, 1)	76.5	1.00003
S003	[72, 73, 74, 75, 76]	0.07	TRUE	0	(1, 0, 1)	78	1.00912
S004	[73, 74, 75, 76, 77]	0.03	FALSE	1	(2, 1, 1)	78	1.00774
S005	[74, 75, 76, 77, 78]	0.05	TRUE	0	(1, 0, 1)	79.5	1.00757
S025	[74, 75, 76, 77, 78]	0.03	TRUE	0	(1, 0, 1)	79	1.00338
S026	[70, 71, 72, 73, 74]	0.05	FALSE	1	(2, 1, 1)	75.5	0.00757
S027	[71, 72, 73, 74, 75]	0.07	TRUE	0	(1, 0, 1)	77	2.00986
S028	[72, 73, 74, 75, 76]	0.03	FALSE	1	(2, 1, 1)	77	0.00876
S029	[73, 74, 75, 76, 77]	0.05	TRUE	0	(1, 0, 1)	78.5	1.00338
S030	[74, 75, 76, 77, 78]	0.07	FALSE	1	(2, 1, 1)	80	0.00757
							RMSE = 1.24312

Table 5. Pearson correlation matrix of core features

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
F1-video_watch_time	1									
F2-quiz_score	0.780***	1								
F3-login_frequency	0.773***	0.731***	1							
F4-time_spent_on_platform	0.769***	0.775***	0.767***	1						
F5-assignment_submits	0.763***	0.751***	0.688***	0.705***	1					
F6-forum_posts	0.779***	0.772***	0.754***	0.695***	0.757***	1				
F7-quiz_attempts	0.741***	0.712***	0.689***	0.706***	0.680***	0.726***	1			
F8-days_since_last_login	0.022	-0.042	-0.010	0.007	0.065	-0.053	0.058	1		
F9-student_year	0.042	-0.033	0.020	0.004	-0.011	0.027	0.014	0.035	1	
F10-video_pause_count	0.761***	0.741***	0.645***	0.724***	0.713***	0.692***	0.698***	0.033	0.041	1

Note: ***p < .001; **p < .01; *p < .05

4.4 Random Forest Prediction Effectiveness

To enhance the accuracy and explanatory power of learning outcome predictions in MOOCs, the dynamic features generated from PCA and time series predictions (ARIMA) were integrated to establish an RF model with nonlinear processing capabilities and generalization effects. The input features originate from two primary sources: (1) ten core learning behaviors and background variables selected after PCA dimensionality reduction (e.g., video_watch_time, login_frequency, quiz_score); (2) next_unit_score (F11) predicted by the ARIMA model for individual learning journeys. This dynamic feature offers high time series sensitivity and effectively compensates for the limitations in prediction posed by static features. Depending on the nature of the prediction tasks, the model employs RF Regressor and RF Classifier to predict three learning outcome indicators: final_learning_score (F12), certificate_eligible (F13), and course_completion (F14). Hyperparameter tuning was conducted on training data through 5-fold cross-validation and Grid Search (Probst et al., 2019 [26]; Biau & Scornet, 2016 [13]), and model generalization was assessed using Out-of-Bag (OOB) errors (Breiman, 2001 [18]; Janitza & Hornung, 2018 [27]).

Table 6 summarizes the input features from individual student data in the RF model (F1–F10), next_unit_score (F11) predicted by the ARIMA model, final_learning_score (F12) predicted by the model, as well as certificate_eligible (F13) and course_completion (F14) estimated based on the prediction results. The final column (error) indicates the differences between predicted values and actual scores, which can be used to evaluate model accuracy. The RMSE value shown at the bottom of Table 6 is 1.17831. This result indicates that the overall model demonstrates stability and good performance in predicting learning outcomes.

To enhance the interpretation and selection basis for input variables within the RF model, this study constructed a Pearson's correlation coefficient heatmap (as shown in Figure 2) to visually examine the linear relationships among the core features. In Figure 2, shades of red represent positive correlations, while shades of blue signify negative correlations. The color intensity reflects the degree of correlation among variables. The color bar on the right side indicates the range of correlation coefficients (from -1 to +1), centered at 0. Results demonstrate that most feature variables exhibit significant positive correlations, particularly showing a high linear relationship among video_watch_time (F1), quiz_score (F2), time_spent_on_platform (F4), and login_frequency (F3). This result indicates that learning behaviors are synchronous and interactive. Conversely, days_since_last_login (F8) and student_year (F9) exhibit low correlations, indicating their relative independence of feature information, which aids in enhancing the model's generalizability. This heatmap can not only serve as a validation tool after data preprocessing and feature engineering but also assist in confirming that multicollinearity among the variables selected by PCA has been effectively reduced. Thus, it enhances the stability and explanatory power of subsequent RF modeling. The correlation heatmap reinforces the rationality of the hybrid model's feature structure and lays a visual foundation for model refinement.

Table 6. Summary of input features and prediction results for the RF model

Student_ID	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	Errors
S001	4	18	6	20	4	0	0	2	3	3	60	63	Yes	Yes	1.8352
S002	10	30	2	9	1	3	1	6	1	4	76	75	Yes	Yes	1.7121
S003	7	53	1	16	4	2	0	7	1	4	65	68	Yes	Yes	2.2311
S004	6	43	10	16	1	2	5	6	1	1	74	82	Yes	Yes	1.1735
S005	2	29	10	6	0	0	4	6	0	4	95	94	Yes	Yes	1.1103
S025	1	14	3	23	3	2	6	4	3	3	61	60	No	No	1.0469
S026	9	29	1	28	0	5	2	7	2	3	71	79	Yes	Yes	1.8953
S027	6	37	7	27	4	2	2	6	2	2	96	100	Yes	Yes	1.5096
S028	7	46	4	19	2	4	2	6	5	4	67	65	Yes	Yes	0.6622
S029	0	79	1	28	4	6	1	2	2	2	84	87	Yes	Yes	1.1348
S030	10	20	5	3	4	1	2	6	2	2	55	53	No	No	1.1172
															RMSE = 1.17831

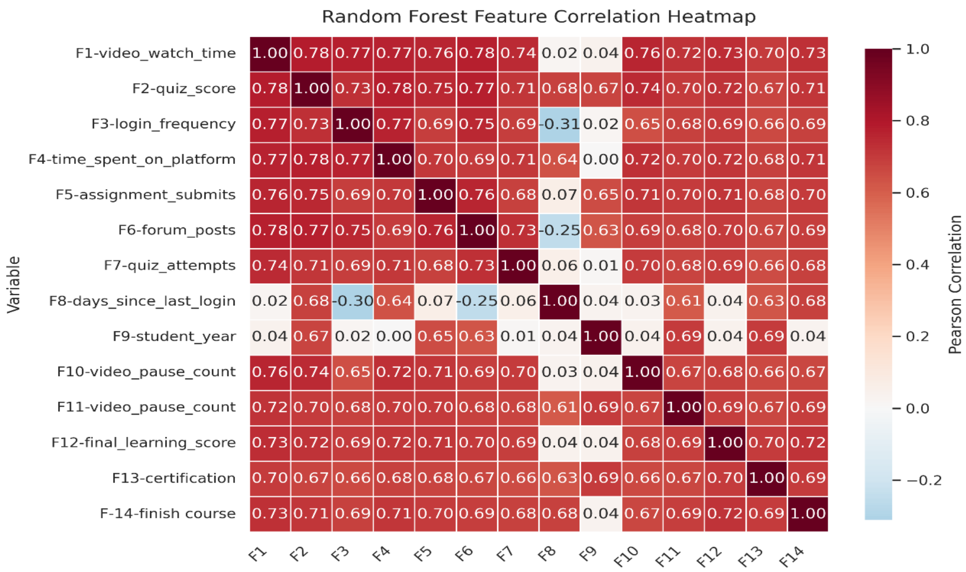


Figure 2. RF Feature correlation heatmap

4.5 Model Evaluation and Comparison

To comprehensively assess the performance of various prediction models, this study compared six models using five common evaluation indicators (F1-score, MSE, RMSE, MAE, R^2) across the training set (80%) and the test set (20%). Table 7 shows that the hybrid model (ARIMA+PCA+RF), which integrates PCA, time series modeling (ARIMA), and RF, exhibits the best overall performance. It achieves an R^2 of 0.9653 and an RMSE of 1.1792 in the training set, indicating a high degree of accuracy. To evaluate whether the model exhibits stability and generalizability, this study designed an independent validation dataset (20%) for testing. Results indicate that the ARIMA+PCA+RF model maintains low error (RMSE = 1.1626, MAE = 0.78) and high explanatory power (R^2 = 0.9823) in the test set, with minimal difference in performance compared to the training stage, suggesting no overfitting. In contrast, single modeling strategies such as RA and SVM exhibit significant variability in performance across the training and validation data, indicating inadequate comprehension of the complexities inherent in MOOC learning behaviors. Table 7 presents the prediction performance of six models on the training set (80%) and test set (20%). Evaluation indicators include the commonly used F1-score for classification tasks, as well as MSE, RMSE, MAE, and R^2 for regression models. The results demonstrate that the ARIMA+PCA+RF model outperforms all other models across all indicators. It achieves R^2 = 0.9653 in the training set and R^2 = 0.9823 in the test set, with RMSE values of 1.1792 and 1.1626, respectively, indicating excellent generalizability and stability.

Figure 3 and Figure 4 analyze the impact of various feature input combinations and training-test ratios on the prediction performance of models (measured by RMSE). Results indicate that the hybrid model (ARIMA+PCA+RF), integrating PCA, time series modeling (ARIMA), and RF, consistently exhibits the lowest RMSE across various configurations, demonstrating high accuracy and robustness in predicting learning behaviors. In Figure 3 (training 60%/testing 40%), even with a relatively small training sample, the ARIMA+PCA+RF model maintains the lowest RMSE across all feature combinations. It achieves 1.392995 with the complete feature set (F1–F14), significantly outperforming traditional methods such as RA (2.38289) and SVM (2.43776). This result indicates that the model exhibits good learning performance even in conditions of limited data. Figure 4, utilizing a configuration of training 80%/testing 20%, shows an overall decrease in RMSE. The ARIMA+PCA+RF model achieves an RMSE of 1.138195 with the complete feature set. This result indicates its continued best performance and confirms that more training data aids in model learning stability. Furthermore, when feature combinations expand from F1–F2 to F1–F14, all models exhibit a general decreasing trend in RMSE. The hybrid model demonstrates the greatest reduction, illustrating its adaptability to high-dimensional inputs and comprehensive advantages. The results support the practicality of the proposed hybrid model in addressing MOOC time series and behavioral prediction issues. It can accommodate nonlinear fitting and capture temporal dynamics, providing stable and precise prediction outcomes.

The ARIMA+PCA+RF hybrid model developed in this study demonstrates the best prediction performance. It achieves the lowest RMSE and highest R^2 across both training and test data, indicating that its overall stability and generalizability are superior to those of traditional methods. This framework successfully integrates the feature dimensionality reduction of PCA, the learning trend-capturing capabilities of ARIMA, and the nonlinear fitting abilities of RF. It can effectively merge static and dynamic information, demonstrating robust prediction potential. Moreover, this model possesses the capability to identify high-risk students early. Therefore, it can provide practical value for remedial teaching and resource allocation, and can serve as a basis for educational decision-making and personalized learning support.

Table 7. Performance comparison of models on training and test sets

Method	Trainnig (80% 140 records)					Trainnig (20% 140 records)				
	F1	MSE	RMSE	MAE	R^2	F1	MSE	RMSE	MAE	R^2
RA	0.65	1.85236	1.3601	1.06	0.8768	0.68	1.83234	1.3538	1.04	0.9125
SVM	0.75	2.36593	1.5388	1.21	0.6574	0.78	2.26693	1.5056	1.17	0.7231
ARIMA	0.78	1.65736	1.2874	1.03	0.5575	0.81	1.54873	1.2444	0.99	0.7177
RF	0.82	1.52598	1.2358	0.94	0.7814	0.85	1.48726	1.2137	0.88	0.8599
ARIMA+RF	0.89	1.41833	1.1912	0.88	0.8634	0.91	1.46543	1.1805	0.82	0.9334
ARIMA+PCA+RF	0.94	1.39276	1.1792	0.84	0.9653	0.96	1.35612	1.1626	0.78	0.9823

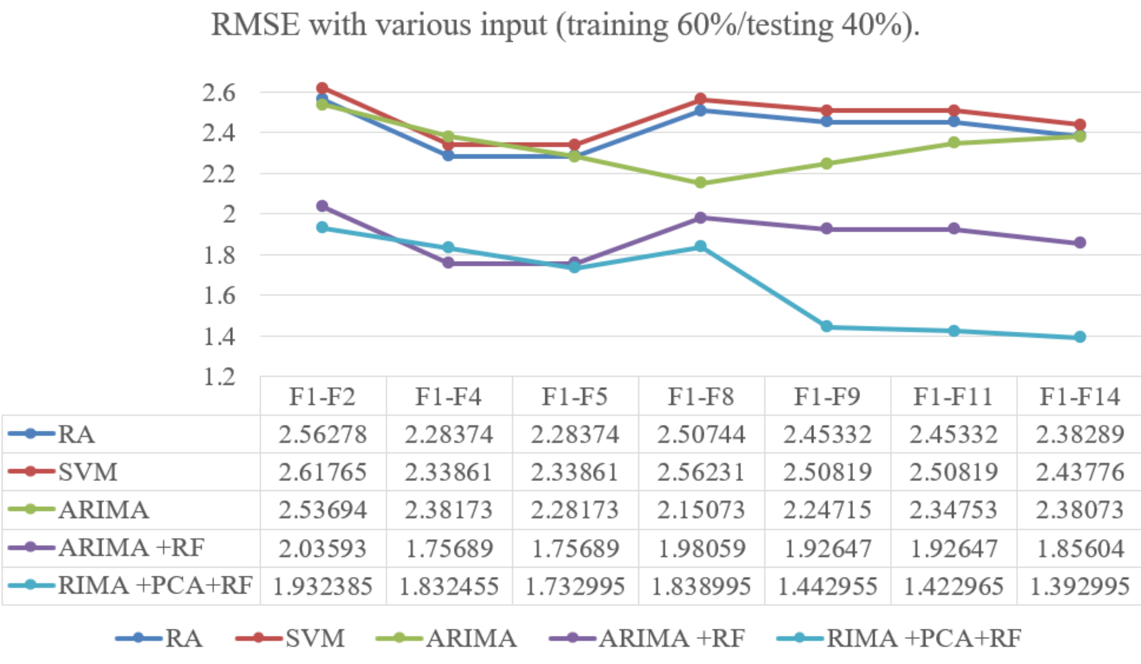
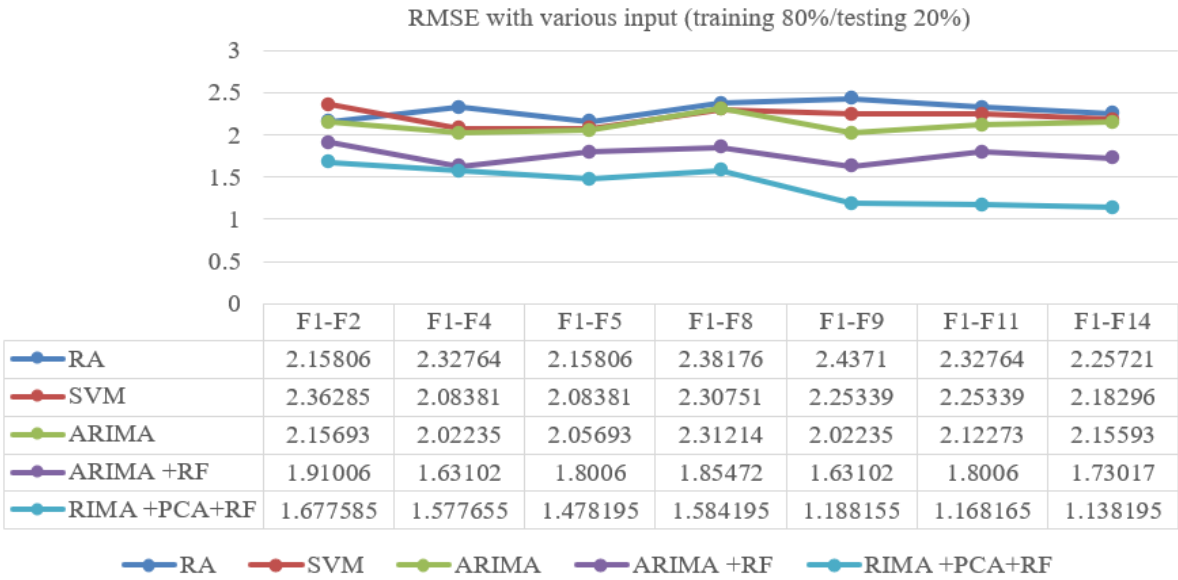


Figure 3. Variation of RMSE across different input feature sets (training: 60%, testing: 40%)



	F1-F2	F1-F4	F1-F5	F1-F8	F1-F9	F1-F11	F1-F14
RA	2.15806	2.32764	2.15806	2.38176	2.4371	2.32764	2.25721
SVM	2.36285	2.08381	2.08381	2.30751	2.25339	2.25339	2.18296
ARIMA	2.15693	2.02235	2.05693	2.31214	2.02235	2.12273	2.15593
ARIMA +RF	1.91006	1.63102	1.8006	1.85472	1.63102	1.8006	1.73017
RIMA +PCA+RF	1.677585	1.577655	1.478195	1.584195	1.188155	1.168165	1.138195

Figure 4. RMSE trend across different input features (80% Training / 20% Testing Split)

5 Conclusion

5.1 Discussion and Recommendations

This study proposed a hybrid prediction framework that integrates ARIMA, PCA, and RF to enhance the accuracy and stability of learning outcome predictions in MOOCs. ARIMA captures temporal trends in learning behaviors, PCA simplifies high-dimensional feature spaces, and RF captures nonlinear rules; this framework effectively addresses problems of variable redundancy, time dependency, and heterogeneity in MOOC data. Empirical results show that ARIMA+PCA+RF outperforms traditional models in both continuous and binary prediction tasks. It consistently leads in indicators such as the F1-score, MSE, RMSE, and R^2, with good generalizability. These findings are consistent with past research that confirms the advantages of hybrid models [28]. Moreover, the findings confirm that PCA can improve redundancy and stability [29], RF exhibits robust classification and nonlinear processing capabilities [18], while ARIMA can address the deficiencies of static features for dynamic behaviors [30]. The proposed hybrid model not only embodies methodological innovation but also holds potential application value in contexts such as individualized early warnings, remedial teaching, adaptive curricula, and educational decision-making supports. Regarding research recommendations, first, MOOC platforms can use this hybrid model as an early warning mechanism to provide timely reminders and remedial resources for potentially high-risk students. Second, we suggest incorporating students' time series scores to construct ARIMA models and capture individualized learning trends. Additionally, selecting representative behavioral features (e.g., video_watch_time and login_frequency) through PCA can reduce redundancy and enhance the model's efficiency and explanatory power. Finally, we recommend that platforms provide dynamic feedback based on prediction results, such as video re-watching, assignment reminders, and peer interactions, to promote completion rates and learning motivation.

5.2 Implications

The prediction performance of the integrated model proposed in this study holds significant implications for both academia and practice.

1. Academic Implications

(1) Cross-Domain Integration Innovation: Through integrating time series, dimensionality reduction, and machine learning methods, this study proposes a prediction model with generalizability and empirical foundations to serve as a new paradigm for learning analysis and educational data research.

(2) Strengthening Time Series Analysis of Learning: This study pioneers the application of the ARIMA model to individual score sequences, enhances the understanding of dynamic behavioral changes, and expands the application potential of time-dependent data in learning analysis.

(3) Optimizing Feature Engineering Practices: By effectively screening representative key variables through PCA, this study provides operational guidance for modeling and simplifying high-dimensional educational data.

2. Practical Implications

(1) Enhancing Prediction and Early Warning Capabilities of MOOCs Platforms: The model proposed in this study can assist platforms as early as possible in identifying high-risk students, initiating dynamic intervention mechanisms, and optimizing supports throughout their learning process.

(2) Supporting Individualized Teaching Interventions: Prediction results can serve as precise teaching bases for learning reminders and resource recommendations,

facilitating the implementation of learner-centered strategies.

(3) Providing a Basis for Teaching Decision-Making: Model outputs can serve as quantitative references for teachers and administrative units of academic affairs to adjust courses and resource allocations, enhancing teaching quality.

5.3 Future Research

Although the ARIMA+PCA+RF model proposed in this study performs well in predicting learning outcomes in MOOCs, some limitations remain. First, the sample source is limited to a single course (140 records), which may restrict the model's generalizability across diverse courses or platforms. Second, the ARIMA model demands high integrity for time series data; missing value imputation may impact prediction accuracy. Future research should consider using deep learning architectures such as LSTM to strengthen time series modeling and incorporate more learning behavior data (e.g., clickstream data and interaction pathways) to expand feature dimensions.

References

[1] T. R. Liyanagunawardena, A. A. Adams, S. A. Williams, MOOCs: A systematic study of the published literature 2008–2012, *International Review of Research in Open and Distributed Learning*, Vol. 14, No. 3, pp. 202–227, July, 2013. https://doi.org/10.19173/irrodl.v14i3.1455

[2] O. Zawacki-Richter, A. Bozkurt, U. Alturki, A. Aldraiweesh, What research says about MOOCs – An explorative content analysis, *International Review of Research in Open and Distributed Learning*, Vol. 19, No. 1, pp. 242–259, February, 2018. https://doi.org/10.19173/irrodl.v19i1.3356

[3] K. Jordan, Massive open online course completion rates revisited: Assessment, length and attrition, *International Review of Research in Open and Distributed Learning*, Vol. 16, No. 3, pp. 341–358, June, 2015. https://doi.org/10.19173/irrodl.v16i3.2112

[4] S. Alghamdi, B. Soh, A. Li, A comprehensive review of dropout prediction methods based on multivariate analysed features of MOOC platforms, *Multimodal Technologies and Interaction*, Vol. 9, No. 1, Article No. 3, January, 2025. https://doi.org/10.3390/mti9010003

[5] F. Xiong, K. Zou, Z. Liu, H. Wang, Predicting learning status in MOOCs using LSTM, *Proceedings of the ACM Turing Celebration Conference - China (ACM TURC)*, Chengdu, China, 2019, pp. 1–5. https://doi.org/10.1145/3321408.3322855

[6] Y. Jo, K. Maki, G. Tomar, Time series analysis of clickstream logs from online courses, *arxiv preprint*, arXiv:1809.04177, September, 2018. https://arxiv.org/abs/1809.04177

[7] M. S. Boroujeni, P. Dillenbourg, Discovery and temporal analysis of MOOC study patterns, *Journal of Learning Analytics*, Vol. 6, No. 1, pp. 16–33, April, 2019. https://doi.org/10.18608/jla.2019.61.2

[8] M. Khalil, M. Ebner, Learning analytics: Principles and constraints, *Proc. World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Montreal, Canada, 2015, pp. 1789–1799.

[9] M. Kloft, F. Stiehler, Z. Zheng, N. Pinkwart, Predicting MOOC dropout over weeks using machine learning methods, *Proc. EMNLP Workshop on Analysis of Large Scale Social Interaction in MOOCs*, Doha, Qatar, 2014, pp. 60–65. https://doi.org/10.3115/v1/W14-4111

[10] R. Tapio, D. Tarepe, Comparative analysis of Random Forest and hybrid ARIMA-Random Forest models for student enrollment forecasting in higher education, *Journal of Advances in Mathematics and Computer Science*, Vol. 40, No. 3, pp. 124–136, March, 2025. https://doi.org/10.9734/jamcs/2025/v40i31982

[11] Z. Ren, H. Rangwala, A. Johri, Predicting performance on MOOC assessments using multi-regression models, *arxiv preprint*, arXiv:1605.02269, May, 2016. https://arxiv.org/abs/1605.02269

[12] G. E. P. Box, G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, San Francisco, USA: Holden-Day, 1970.

[13] G. Biau, E. Scornet, A random forest guided tour, *TEST*, Vol. 25, No. 2, pp. 197–227, June, 2016. https://doi.org/10.1007/s11749-016-0481-7

[14] S. Mao, C. Zhang, Y. Song, J. Wang, X. J. Zeng, Z. Xu, Q. Wen, Time series analysis for education: Methods, applications, and future directions, *arxiv preprint*, arXiv:2408.13960, August, 2024. https://arxiv.org/abs/2408.13960

[15] V. I. Kontopoulou, A. D. Panagopoulos, I. Kakkos, G. K. Matsopoulos, A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks, *Future Internet*, Vol. 15, No. 8, Article No. 255, August, 2023. https://doi.org/10.3390/fi15080255

[16] L. Qin, K. Shanks, G. A. Phillips, D. Bernard, The impact of lengths of time series on the accuracy of the ARIMA forecasting, *International Research in Higher Education*, Vol. 4, No. 3, pp. 58–68, September, 2019. https://doi.org/10.5430/irhe.v4n3p58

[17] Ü. Ç. Büyükşahin, Ş. Ertekin, Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition, *Neurocomputing*, Vol. 361, pp. 151–163, October, 2019.

[18] L. Breiman, Random forests, *Machine Learning*, Vol. 45, No. 1, pp. 5–32, October, 2001. https://doi.org/10.1023/A:1010933404324

[19] H. Tyralis, G. Papacharalampous, Variable selection in time series forecasting using random forests, *Algorithms*, Vol. 10, No. 4, Article No. 114, December, 2017. https://doi.org/10.3390/a10040114

[20] A. Alamri, M. Alshehri, A. Cristea, F. D. Pereira, E. Oliveira, L. Shi, C. Stewart, Predicting MOOCs dropout using only two easily obtainable features from the first week's activities, *Proceedings of the 15th International Conference on Intelligent Tutoring Systems (ITS 2019)*, Kingston, Jamaica, 2019, pp. 163–173. https://doi.org/10.1007/978-3-030-22244-4_20

[21] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLoS ONE*, Vol. 10, No. 3, Article No. e0118432, March, 2015. https://doi.org/10.1371/journal.pone.0118432

[22] R. J. Hyndman, G. Athanasopoulos, *Forecasting: principles and practice*, 2nd ed., Melbourne, Australia: OTexts, 2018.

[23] T. Chai, R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, *Geoscientific Model Development*, Vol. 7, No. 3, pp. 1247–1250, June, 2014. https://doi.org/10.5194/gmd-7-1247-2014

[24] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, *Climate Research*, Vol. 30, No. 1, pp. 79–82, December, 2005. https://www.jstor.org/stable/24869236

[25] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, New York: Springer, 2013.

[26] P. Probst, M. N. Wright, A. L. Boulesteix, Hyperparameters and tuning strategies for Random Forest, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 9, No. 3, Article No. e1301, May/June, 2019. https://doi.org/10.1002/widm.1301

[27] S. Janitza, R. Hornung, On the overestimation of Random Forest's out-of-bag error, *PLoS ONE*, Vol. 13, No. 8, Article No. e0201904, August, 2018. https://doi.org/10.1371/journal.pone.0201904

[28] W. Xing, X. Chen, J. Stein, M. Marcinkowski, Temporal prediction of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization, *Computers in Human Behavior*, Vol. 58, pp. 119–129, May, 2016. https://doi.org/10.1016/j.chb.2015.12.007

[29] I. T. Jolliffe, J. Cadima, Principal component analysis: A review and recent developments, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 374, No. 2065, Art. No. 20150202, April, 2016. https://doi.org/10.1098/rsta.2015.0202

[30] K. R. Koedinger, S. D'Mello, E. A. McLaughlin, Z. A. Pardos, C. P. Rosé, Data mining and education, *Wiley Interdisciplinary Reviews: Cognitive Science*, Vol. 6, No. 4, pp. 333–353, July/August, 2015. https://doi.org/10.1002/wcs.1350

Biographies

Po-Yuan Su holds a Ph.D. in Design from National Yunlin University of Science and Technology. His research focuses on digital game and multimedia design evaluation, applying user-centered methods, analytics, and creative media to enhance learning experiences, interactivity, and design quality.

Chung-Ho Su is a Professor of Animation and Game Design at Shu-Te University, Taiwan. He holds dual Ph.D.s in Information Management and Design from National Yunlin University of Science and Technology. His research spans digital media and game design, design thinking, soft computing, and design education evaluation.