Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 7, No. 1, March, 2026

Academic Performance Prediction Models Based on Multi-Head Attention LSTM Mechanisms

Shiyuan Zhou [1*]

[1] School of Information Engineering, Jiaxing Nanhu University, Jiaxing, Zhejiang, 314001, China.

Received 29 December 2025; Revised 19 February 2026; Accepted 22 February 2026; Published 01 March 2026

Abstract

This study focuses on accurately predicting student academic performance to support personalized teaching and optimize educational resource allocation. Addressing the limitations of traditional prediction methods in handling long-term dependencies within time-series data and extracting key features, this paper proposes an innovative model that integrates multi-head attention mechanisms with Long Short-Term Memory (LSTM) networks. The method utilizes LSTMs to efficiently process sequential data while employing multi-head attention mechanisms to concentrate on critical information synergistically. Experiments demonstrate that this model achieves significantly higher prediction accuracy than traditional methods on specific datasets, with an MAE of 4.12 and an R^2 of 0.94, fully showcasing its outstanding performance. This model pioneers new pathways for academic performance prediction in education, supporting scientific educational decision-making and driving high-quality educational development.

Keywords: Multi-Head Attention; Network of Long Short-Term Memory; Academic Performance Prediction; Big Data Mining.

1. Introduction

In today's era of rapidly advancing digital education, students' academic achievement is impacted by a complex interplay of multiple factors [1]. With the continuous advancement of educational informatization, schools have accumulated vast amounts of student learning data, including daily homework scores, exam results, classroom performance, and records of online learning activities. This data holds valuable insights into students' learning states, knowledge mastery levels, and performance trends [2]. Effectively mining and utilizing this data to provide personalized learning recommendations for students and assist teachers in optimizing instructional strategies has become a critical issue requiring urgent resolution in the education sector [3]. Traditional forecasting methods struggle in the long run due to the reliance inherent in time-series data, making it difficult to accurately capture the complex patterns underlying changes in student performance [4]. Accurate academic performance prediction not only enables students to proactively understand their learning status and potential risks across subjects but also empowers teachers to gain comprehensive insights into both the overall class learning dynamics and individual student progress. Furthermore, it assists school administrators in allocating educational resources more scientifically and rationally [5]. Consequently, there is an urgent need for a predictive model capable of efficiently processing time-series data and precisely extracting key features.

Currently, both domestic and international research in educational data mining and academic performance prediction has commenced early and yielded substantial results. On one hand, traditional statistical methods such as multiple linear regression were widely applied in early studies [6]. Liu et al. [7] established linear models for prediction by analyzing the correlation between factors like students' family backgrounds and study time investment with their grades. However,

* Corresponding author: shiyuan.zhou@jxnhu.edu.cn

https://doi.org/10.28991/HIJ-2026-07-01-016

such methods prove inadequate when handling complex nonlinear relationships and time-series data. Wang et al. [8] employed a grey prediction model to forecast student grades with small samples, achieving practical results. Abín et al. [9] combined principal component analysis and other methods to reduce dimensions and extract features from multidimensional factors influencing student performance before establishing a prediction model. On the other hand, machine learning methods have gradually gained prominence [10]. Alturki et al. [11] employed decision tree algorithms to construct models predicting course pass rates based on students' past grades and course selection patterns. Wang & Lewis [12] achieved notable results by classifying exam scores using support vector machines. In recent years, deep learning technologies have garnered significant attention. Li & Gao [13] utilized neural networks to predict final grades in specific subjects, automatically extracting latent features from data. Gatzka [14] introduced attention mechanisms to focus on critical aspects of student learning data, but the deep integration of multi-head attention with LSTMs remains exploratory with substantial room for advancement. Despite these efforts, academic performance prediction research faces multiple challenges and limitations [15]. First, student learning data in real educational settings often contains noise interference. Second, acquiring large-scale, high-quality labeled data is difficult. Third, models are prone to overfitting under small-sample training conditions, failing to generalize effectively to new data. Finally, models suffer from insufficient interpretability.

Despite the progress made in these studies, a significant gap remains in the literature: most existing models struggle to simultaneously account for long-term dependencies in time-series data and the dynamic weight allocation of different features. Single LSTM models often fail to distinguish the importance of specific input features, while traditional machine learning methods lack the depth to handle complex sequential data. To fill this gap, this study introduces a multi-head attention mechanism to enhance the LSTM network, thereby significantly improving the model's ability to capture key features while maintaining its capacity to process temporal information.

Addressing these research gaps, this paper proposes a multi-head attention LSTM-based academic performance prediction model, aiming to provide an innovative solution for educational achievement forecasting. This paper deeply integrates the multi-head attention mechanism with LSTM networks to construct a student academic performance prediction model. It designs a step-by-step process for building an academic performance prediction based on deep learning algorithms. Experimental validation demonstrates that the model achieves high prediction accuracy across multiple educational datasets, showing significant improvement over traditional methods and other existing models. This provides a more reliable basis for educational decision-making, advancing personalized teaching and the optimized allocation of educational resources in the context of educational informatization.

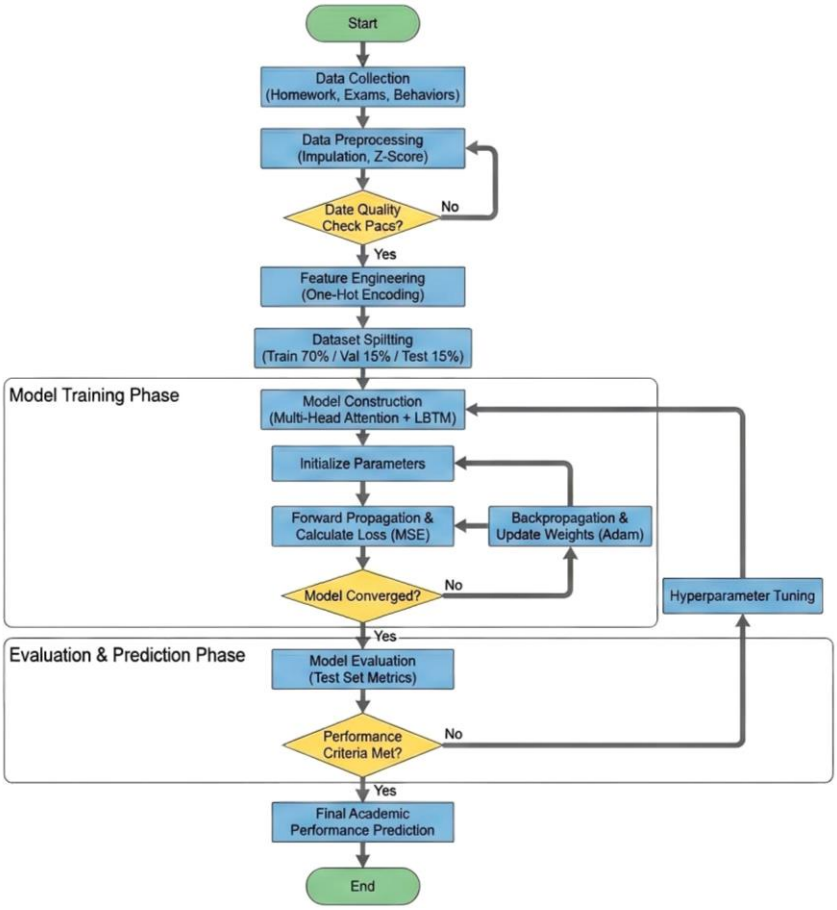

Figure 1. Basic Process of Academic Performance Prediction

2. Related Works

2.1. Overview of Academic Performance Prediction

Academic performance prediction is a critical task in educational data mining. It aims to establish models that forecast future academic outcomes based on historical student data—including homework grades, exam scores, classroom performance, and other multi-source information—using data analysis and mining techniques [16]. These data typically exhibit multi-source and heterogeneous characteristics, encompassing numerical grade data, textual comment data, and time-series learning process data, collectively reflecting students' learning status and knowledge mastery [17]. Figure 1 illustrates the fundamental workflow of academic performance prediction, where each step from data collection to the application of final prediction results is crucial.

Common prediction methods can be categorized into traditional statistical approaches [18] and emerging data mining techniques [19]. Traditional statistical methods, such as linear regression, predict outcomes by analyzing linear relationships between grades and factors like study time or family background. Nevertheless, these techniques have drawbacks when managing complex nonlinear relationships and time-series data [20]. With the advancement of data mining technologies, machine learning methods have gained prominence. For instance, decision tree algorithms construct tree-like models based on diverse features within grade data—such as past performance and course selections—to predict whether a student will pass a specific course [21]. Support vector machines achieve classification predictions for exam scores by mapping data to high-dimensional spaces [22]. In recent years, deep learning techniques—particularly the integration of Networks LSTM (Long Short-Term Memory) and multi-head attention mechanism—have brought breakthroughs to academic performance prediction [23]. Table 1 compares common data-driven algorithms for academic performance forecasting.

Table 1. Comparison of Common Academic Performance Prediction Methods

Method	Advantages	Disadvantages
Linear Regression	Simple and easy to understand, high computational efficiency	Difficult to handle complex nonlinear relationships, weak in identifying long-term relationships in time series information
Decision Tree	Strong interpretability capable of handling non-numeric data	Prone to overfitting and low efficiency in processing large-scale data
Support Vector Machine (SVM)	High prediction accuracy suitable for small sample data	Sensitive to parameter selection, high computational complexity, and difficult to handle time series data directly
Deep Learning	Efficiently manages time series data's long-term dependencies, automatically extracts data features.	Complex model structure and long training time require large-scale data support.

2.2. Long Short-Term Memory Networks

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network that was created especially for processing sequential data. They effectively deal with the gradient's disappearance and explosion issues encountered by traditional RNNs when handling long sequences, owing to their unique gating mechanism [24]. LSTMs regulate information flow through forget gates, input gates, and output gates, thereby capturing long-term dependencies in sequential data.

The basic unit, Figure 2, depicts the LSTM's structure, comprising a cell state and three gating structures. The amount of prior cell state information is determined by the forget gate, which is retained; the input gate regulates how much new input data enters the cell state; and how much of the output gate determines current cell state is output.

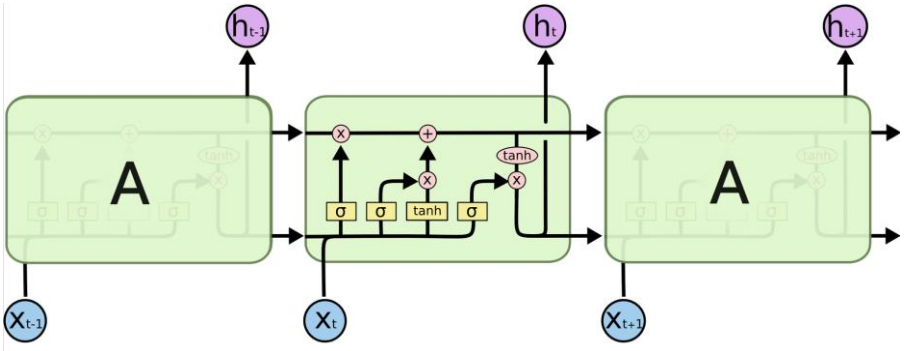

Figure 2. Basic LSTM Unit Structure

Based on its structure and principles, LSTM possesses advantages such as long-term memory capability, information filtering functions, and robustness against noise. In academic performance prediction, LSTM can take a student's historical grade sequence as input. By learning the temporal dependencies within the sequence, it predicts future grades, as seen in Figure 3.

Figure 3. Application of LSTM in Academic Performance Prediction

2.3. Multiple-Head Attention System

The multi-head attention mechanism captures correlations between elements at different positions within sequence data. By carrying out calculations in parallel across several attention heads, it broadens the model's informational understanding dimensions, enabling more comprehensive extraction of multi-granularity features from sequence data [25]. Figure 4 illustrates the structure regarding the mechanics of multi-head attention, demonstrating its process of concurrently computing multiple attention heads.

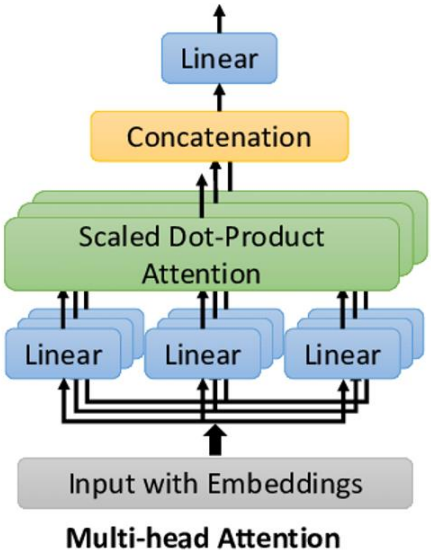

Figure 4. Multi-Head Attention Mechanism Architecture

The attention of several heads mechanism first linearly transforms the input vector into query (Q), key (K), and value (V) matrices. Each attention head independently calculates attention weights according to the following formula:

$$Attention(Q\ K\ V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Here, d_k denotes the key's dimensions vector. Each attention head's outputs are concatenated and then passed through a linear transformation to yield the final output. The whole process can be described as:

$$MultiHead(Q\ K\ V) = Concat(head_1\ head_2\ \cdots\ head_h)W^O \tag{2}$$

where $head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$ and h denote the quantity of attention heads, and W^O denotes the concatenated linear transformation matrix.

Based on its structure and principles, the mechanism of multi-head attention captures multi-granularity features, enhances expressive power, and improves training efficiency. In the field of natural language processing, it is widely applied to tasks like machine translation and text generation [26]. In machine translation, this mechanism enables models to simultaneously focus on different words within input sentences and their syntactic and semantic relationships, thereby generating more accurate and fluent target-language sentences [27]. In academic performance prediction, the mechanism for multi-head attention can capture multidimensional correlations between student grades and other relevant factors, thereby improving the accuracy of prediction models. As illustrated in Kumari [28], one attention head may focus on the student's exam trend, while another may concentrate on the impact of course difficulty on performance, as shown in Figure 5.

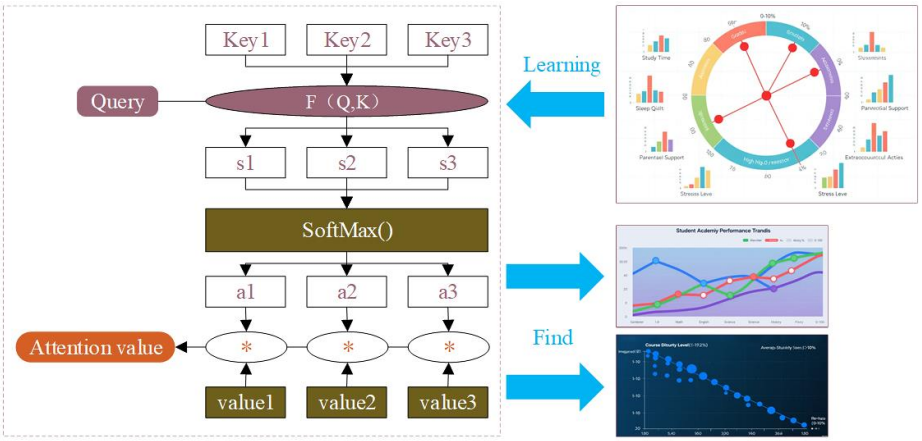

Figure 5. Application of Multi-Head Attention Mechanism in Academic Performance Prediction

3. Construction of an Academic Performance Prediction Model Based on Multi-Head Attention LSTM

3.1. Data Collection and Preprocessing

This study's data collection encompasses students' academic performance and related information. This study strictly adhered to ethical guidelines. Prior to data processing, all personally identifiable information (PII) was removed and replaced with anonymous IDs to protect student privacy. Furthermore, the data collection process was approved by the relevant school administration and complies with the ethical standards for educational data usage. The collected data includes daily homework scores, semester exam scores, classroom performance ratings, interaction records from online learning platforms, course selection history, and basic background information [29]. This multidimensional dataset comprehensively reflects students' learning behaviors and states, providing rich feature information for model training.

Data preprocessing primarily addressed issues such as missing values, significant differences in data dimensions, and feature encoding, as illustrated in Figure 6. Some feature data for certain students in the collected dataset might be missing. For numerical features, missing values were imputed using the feature's mean; for categorical features, the mode was used for imputation. In addition to handling missing values, statistical outlier detection techniques were applied to mitigate noise interference. Specifically, the Interquartile Range (IQR) rule was utilized to filter out extreme anomaly records that significantly deviated from the normal distribution range of grades, thereby enhancing data quality. To enhance training efficiency and accuracy, numerical features underwent standardization due to significant differences in data dimensions. The Z-Score normalization method [30] was employed, converting each feature value into the difference between the feature value and the mean divided by the standard deviation. This ensures the normalized data has a variance of one and a mean of zero. For feature encoding, categorical features were converted into numerical features for easier model processing. One-hot encoding was used, mapping each categorical value to a binary vector. In the literature, course selections included categories such as “Mathematics,” “English,” and “Physics.” After one-hot encoding, convert categorical features to one-hot encoding. Subsequently, feature importance analysis was performed using the Pearson Correlation Coefficient to quantify the linear correlation between each feature and academic performance, and features with coefficients exceeding a specific threshold were selected as model inputs. Each course corresponds to a binary vector. For example, Mathematics is represented as [1,0,0], English as [0,1,0], and Physics as [0,0,1]. Table 2 shows the data preprocessing methods.

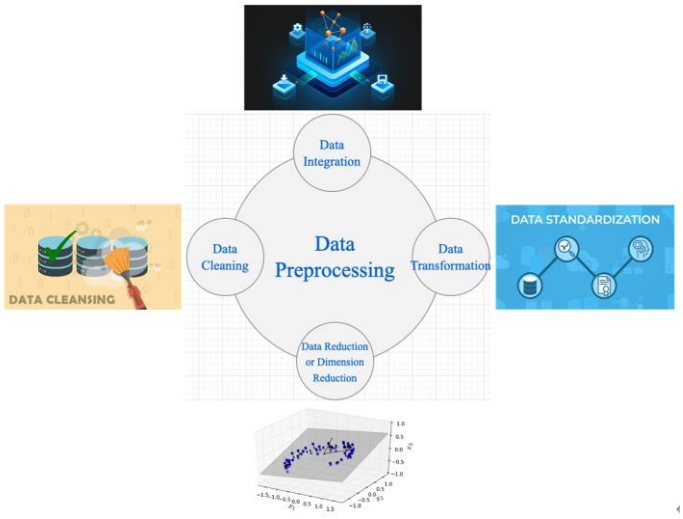

Figure 6. Schematic Diagram of Data Preprocessing Operations

Table 2. Specific Data Preprocessing Techniques

Operation	Problem Type	Technique and Method
Missing Value Processing	Numerical Features	Mean Imputation
Missing Value Processing	Categorical Features	Mode Imputation
Data Standardization	Large differences in feature scales	Z-Score Standardization
Feature Encoding	Categorical Features	One-Hot Encoding

3.2. Model Architecture Design

The overall architecture of the multi-head attention LSTM-based academic performance prediction model is shown in Figure 7. The model takes preprocessed student academic data sequences as input, including historical grades, learning behaviors, and other features; it outputs predicted academic performance for the next phase. The model primarily consists of an input layer, a multi-head attention layer, an LSTM layer, an output layer, and a fully connected layer.

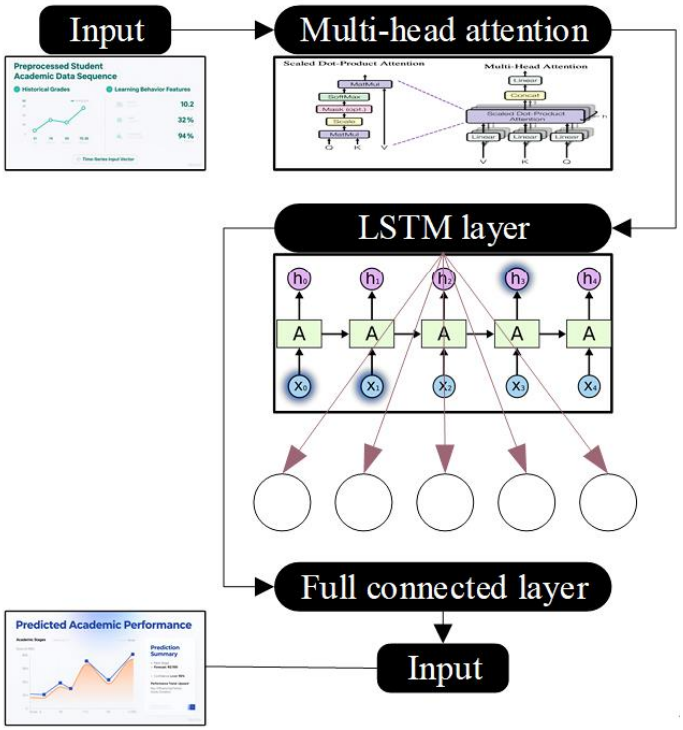

Figure 7. Overall Architecture Diagram

The multi-head attention LSTM-based academic performance prediction model includes an input layer, a multi-head layer of attention, an LSTM layer, a fully connected layer, and an output layer. The input layer is given preprocessed student academic data sequences, with its dimension determined by the number of collected features. The multi-head attention layer contains multiple attention heads, each independently calculating the correlation weights between elements at different positions within the input sequence. This approach enables the model to capture key information and feature relationships from multiple perspectives within the input sequence. Theoretically, this combination addresses the limitations of using a single model type. While LSTM resolves the vanishing gradient problem of traditional RNNs through gating mechanisms and excels at handling dependencies across the time dimension, the multi-head attention mechanism maps inputs into different subspaces via parallel computation to capture spatial correlations among features from multiple dimensions. The theoretical synergy lies in the fact that the LSTM provides temporal context, while the attention mechanism provides feature saliency weighting, constructing a spatiotemporal modeling framework capable of simultaneously understanding "when" changes occur and "which" factors drive them.

During the model design phase, preliminary experiments were conducted to compare the "Attention-LSTM" (attention before LSTM) architecture with "LSTM-Attention" (LSTM before attention) and pure self-attention architectures. The results indicated that applying the multi-head attention mechanism before the LSTM layer effectively extracted global correlations among features prior to temporal processing, yielding the optimal convergence speed and prediction accuracy on our dataset. The temporal processing is mostly handled by the LSTM layer. information of the

input sequence, capturing long-term trends and dynamic patterns in student performance. It regulates the flow level of knowledge via the forget gate, input gate, and output gate, effectively processing long sequence data. The LSTM layer's unit count requires adjustment based on data complexity and experimental outcomes. The fully connected layer further integrates and transforms the features output from the LSTM layer to extract higher-level feature representations. Neurons in this layer are connected to all neurons in the preceding layer, enabling nonlinear combinations of input features. The ReLU activation function is typically employed in fully connected layers, as it introduces nonlinearity and mitigates the vanishing gradient problem [31]. The quantity of neurons in the output layer is established by the prediction target. In this study, since the prediction targets the next-stage academic performance, the output layer contains a single neuron without an activation function, directly outputting the predicted score value. Table 3 displays the outcomes of the analysis of each layer.

Table 3. Analysis of Layers in Multi-Head Attention LSTM

Layer Name	Function	Parameter Analysis
Input Layer	Receives preprocessed student academic data sequence	Input Dimension
Multi-Head Attention Layer	Calculates association weights within input sequences	Number of Attention Heads
LSTM Layer	Captures long-term trends and dynamic patterns of student performance	Number of LSTM Units
Fully Connected Layer	Further integration and transformation of features	Activation Function
Output Layer	Predicts the next phase of academic performance	Output Dimension

3.3. Model Training Process

- ***Loss Function***

During model training, the loss function used to quantify the difference between expected and actual scores is the mean squared error (MSE). The MSE is calculated as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

Here, y_i symbolizes the real score, $\hat{y}_i$ represents the predicted score, and N denotes the number of training samples. By minimizing the mean squared error loss function, the model strives to make its predicted scores as close as possible to the actual scores.

- ***Optimizer***

To efficiently optimize model parameters and reduce the loss function value, the optimizer Adam (Adaptive Moment Estimation) is employed. Combining the advantages of momentum and adaptive learning rates, the Adam optimizer automatically adjusts the learning rate and accelerates the model's convergence process. Its parameter update here is the formula:

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{m_t}{\sqrt{v_t}+\varepsilon} \tag{4}$$

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot \nabla\theta_t \tag{5}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot (\nabla\theta_t)^2 \tag{6}$$

Among these, θ denotes model parameters; α represents the rate of learning; m_t and v_t denote first-order moment estimate and second-order moment estimation, respectively; β_1 and β_2 are the decay rates controlling first-order and second-order moment, respectively; ε is an extremely small constant used to prevent division by zero.

- ***Training Steps***

The training and construction process of the multi-head attention LSTM-based academic performance prediction model includes initializing model parameters, data partitioning, forward propagation, loss calculation, backward propagation, and iterative training. The workflow is seen in Figure 8, with specific steps as follows:

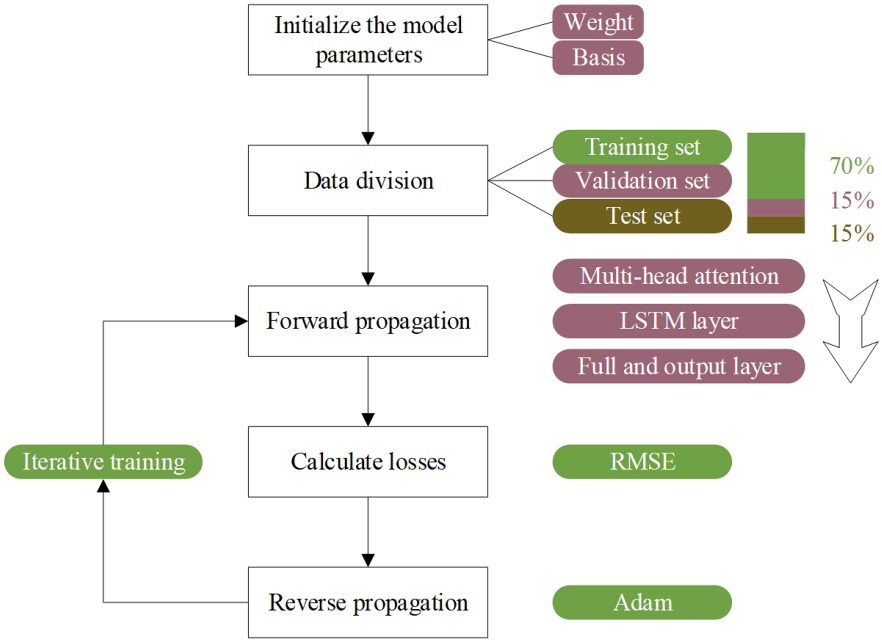

Figure 8. Multi-Head Attention LSTM Prediction Model Training Workflow Diagram

- ***Initialize model parameters.*** Before training begins, randomly initialize the weight matrices and bias terms within the model to establish an initial state for training.
- ***Data partitioning.*** Divide the collected dataset into sets for testing, validation, and training, with proportions of 70%, 15%, and 15%, respectively. The training set is used for model training; the validation set is used to evaluate model performance during training to adjust hyperparameters; the test set serves as the last assessment of the model's capacity for generalization and prediction performance.
- ***Forward propagation.*** The input training set data is fed into the model. After computations through the multi-head attention layer, LSTM layer, fully connected layer, and output layer, the predicted academic scores are obtained.
- ***Loss calculation.*** The current model's loss value is computed using the mean squared error loss function based on predicted and actual scores.
- ***Backpropagation.*** The calculated loss value is propagated backward through the model to compute gradients for each layer. Model parameters are adjusted according to these gradients and the Adam optimizer's update rules.
- ***Iterative training.*** Repeat the above steps until the preset maximum number of training epochs is reached or the model's performance on the validation set ceases to improve. After each training epoch, record the model's loss values on both the instruction and validation sets to monitor the training process and convergence.

3.4. Model Evaluation Metrics

To comprehensively evaluate the performance of the constructed multi-head attention LSTM-based academic achievement prediction model, evaluation metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2) were employed. Detailed comparative analysis is presented in Table 4.

Table 4. Comparative Analysis of Multi-Model Evaluation Metrics

Metric	Principle	Purpose
MAE	Calculates the mean absolute discrepancy between the expected and actual scores	Reflects the average deviation between predicted and actual values
RMSE	Square root of the mean squared error	Reflects model performance for large prediction errors
R^2	1-(Residual Sum of Squares)/(Total Sum of Squares)	Represents the proportion of data variation explained by the model

4. Steps for Building an Academic Performance Prediction Model Based on Multi-Head Attention LSTM

The steps for constructing an academic performance prediction model based on multi-head attention LSTM primarily include data preprocessing, model architecture design, model training configuration, model training and validation, model testing and evaluation, model optimization and adjustment, and model deployment and application. The method for building the academic performance prediction model is seen in Figure 9, with specific steps as follows:

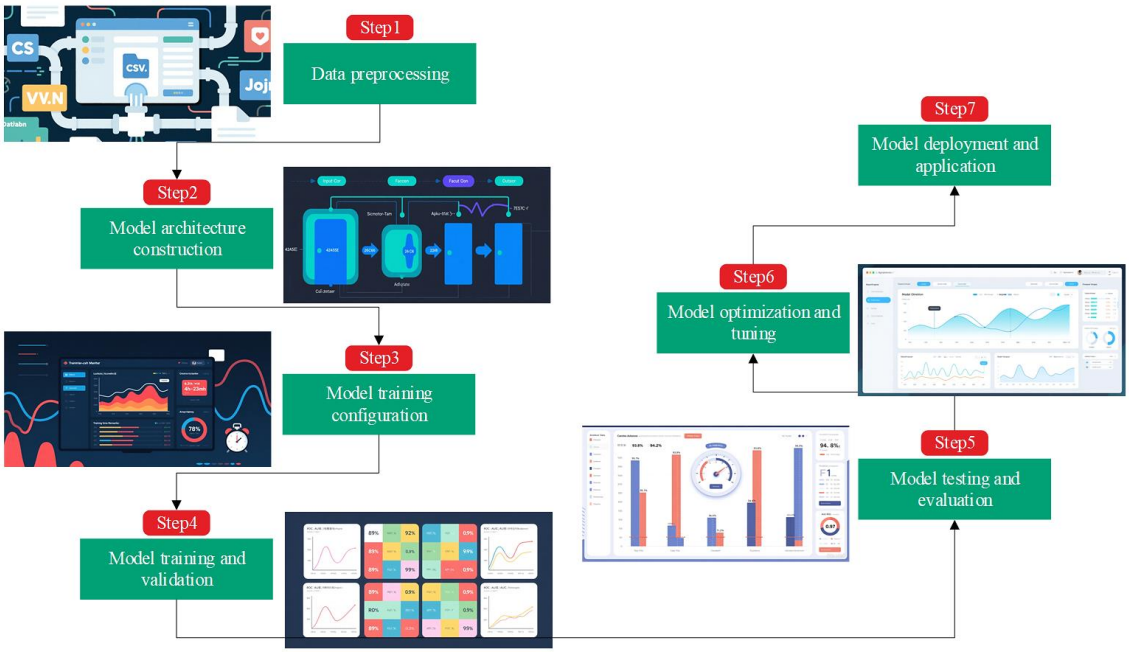

Figure 9. Schematic Diagram of Academic Performance Prediction Model Construction Method

Step 1: Data Preprocessing. Collect student academic performance data, check for missing values, identify and remove duplicate records and obviously anomalous values, standardize numerical features, convert categorical features to one-hot encoding, and select features highly correlated with academic performance as model input features.

Step 2: Model Architecture Setup. Define the input layer dimensions, add multi-head attention layers, set appropriate attention heads and dimensions per head, determine the number of LSTM units, and select suitable activation functions.

Step 3: Model Training Configuration. Choose squared mean error as the loss function and use the Adam optimizer; determine training parameters such as training epochs and batch size.

Step 4: Model Training and Validation. During training, sequentially feed input data from the training set into the model, performing forward propagation and backpropagation to adjust parameters. Validate the model using the validation set to compute loss values and evaluation metrics.

Step 5: Model Testing and Evaluation. Apply the test set using the trained model to obtain predicted scores. Calculate the model's evaluation metrics on the test set.

Step 6: Model Optimization and Adjustment. Based on the model's performance on the validation and test sets, adjust the model's hyperparameters. Simultaneously modify factors such as the number of LSTM layers and the position of multi-head attention layers to observe their impact on model performance, selecting the optimal model architecture.

Step 7: Model Deployment and Application. Package the trained model to enable seamless integration into educational management systems or teaching support platforms. Collect usage feedback from teachers and students, and refine the model based on this feedback to enhance its practicality and reliability.

5. Experimental Design and Results Analysis

5.1. Experimental Design

This experiment utilizes academic performance data from 1,000 secondary school students over the past three years. The dataset covers six consecutive academic semesters. To ensure consistency in the time-series analysis, each student's data was constructed as a sequence containing 6 time steps, ensuring that all subjects were observed over identical time spans. The dataset encompasses multidimensional information, including daily homework scores, semester exam grades,

classroom performance evaluations, online learning platform interaction records, and course selection history. The dataset was partitioned into a test set (15%), validation set (15%), and training set (70%).

To validate the effectiveness of the proposed multi-head attention LSTM-based academic performance prediction model, common prediction models were employed for comparison. To determine the optimal model configuration, a Grid Search strategy was employed to systematically tune key hyperparameters, including the number of attention heads, LSTM units, and the learning rate. Based on performance on the validation set, the parameter combination that minimized the loss function was selected. Specific details and Table 5 display the parameter settings.

Table 5. Common Comparison Prediction Models

Algorithm	Principle
Linear Regression (LR)	Regularization parameter α=0.1 fits the intercept term
Support Vector Machine (SVM)	Kernel C=100.0 gamma=0.01 for the Radial Basis Function
Deep Neural Network (DNN)	Two hidden layers with 64 and 32 neurons respectively, ReLU activation function, learning rate=0.01, batch size=32, 100 training epochs
Long Short-Term Memory (LSTM)	64 memory units in the LSTM layer, tanh activation function, Adam optimizer, batch size=32, 100 training epochs

5.2. Analysis of Experimental Results

To validate the effectiveness and efficiency of the multi-head attention LSTM-based academic performance prediction method, this section compares it with linear regression, support vector machines, DNN, and LSTM, yielding the results shown below.

Each model's performance comparison is shown in Table 6. As shown in Table 6, the proposed multi-head attention LSTM-based academic performance prediction model outperforms other comparison models across all three evaluation metrics: MAE, RMSE, and R^2. Specifically, compared to a standard neural network, MAE decreased by 1.55 points, RMSE decreased by 2.11 points, and R^2 increased by 0.07; compared to LSTM, MAE decreased by 0.86 points, RMSE decreased by 1.11 points, and R^2 increased by 0.03. This improvement in performance can be attributed to the introduction of the multi-head attention mechanism. Unlike standard LSTMs that treat all input time steps equally, the proposed model dynamically adjusts weights based on the prediction target, suppressing the interference of noise data mentioned earlier. The significant reduction in MAE and RMSE indicates that this mechanism effectively mitigates the negative impact of outliers, rendering the model more robust when facing complex and variable student behavioral data. This demonstrates that the proposed model possesses significant advantages in prediction accuracy and fitting performance. Compared to the grey prediction model used by Wang et al. [8] and the decision tree algorithms employed by Alturki et al. [11] mentioned in the introduction, the model proposed in this study not only handles nonlinear relationships but also successfully captures dynamic changes in time series, which traditional statistical methods fail to achieve. Furthermore, in contrast to Gatzka [14], who introduced a single attention mechanism, the multi-head design in this study further enriches the dimensions of feature extraction, proving the necessity of deep feature fusion in educational data mining.

Table 6. Performance Comparison of Models

Model	MAE	RMSE	R^2
Linear Regression	8.24	10.56	0.72
Support Vector Machine	6.85	8.92	0.81
Deep Neural Network	5.67	7.34	0.87
LSTM	4.98	6.12	0.91
Multi-Head Attention LSTM	4.12	5.23	0.94

Figure 10 illustrates the loss variation during the training process of the academic performance prediction model based on multi-head attention LSTM. Training is represented on the horizontal axis, iterations, while the vertical axis denotes loss values. The figure reveals that both training and validation set losses gradually decrease as iterations increase, indicating continuous learning and optimization with progressively improving performance. Concurrently, the validation set loss consistently remains lower than the training set loss throughout training, demonstrating strong generalization capabilities and the absence of overfitting.

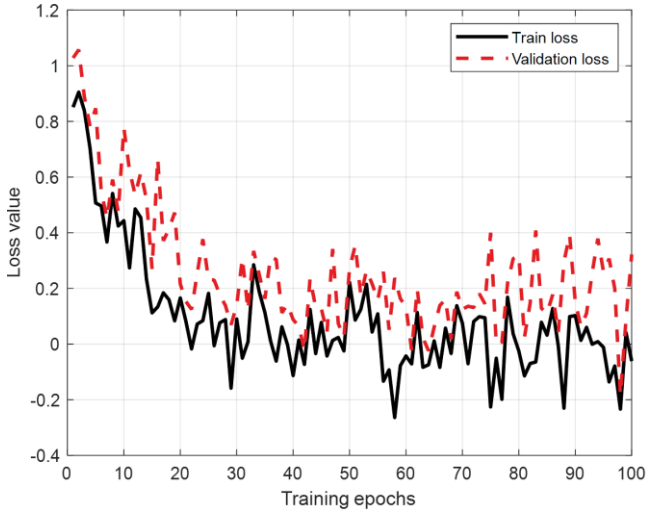

Figure 10. Training Loss Curve

Figure 11 illustrates the contrast between the actual values and the anticipated outcomes of the academic performance prediction model based on multi-head attention LSTM on the test set. The horizontal axis represents the test sample ID, while the vertical axis denotes the score value. The figure reveals that the predicted values of the proposed model closely correspond to the real values, exhibiting consistent overall trends. This further validates the model's predictive accuracy and reliability.

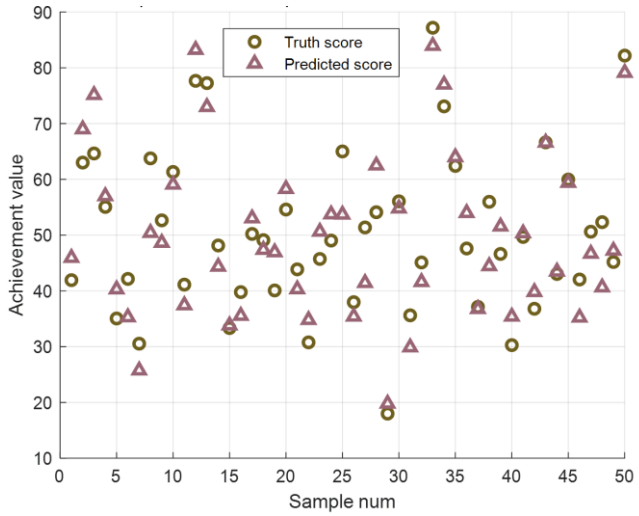

Figure 11. Comparison of Predicted and Actual Values

Table 7 presents a comparison of computational complexity across models. As shown, the linear regression model exhibits the lowest Space complexity (O(1)) and time complexity (O(n)), but its predictive performance is relatively poor. The support vector machine (SVM) has the highest $O(n^3)$ for time complexity and $O(n^2)$ for space complexity, resulting in prolonged training times on large datasets and limiting its practical applicability. The time complexity of the standard neural network, LSTM, and multi-head attention LSTM is $O(n^2)$, with an O(n) space complexity. Among these, the multi-head attention LSTM demonstrates significant advantages in prediction accuracy and model fit. Although its time complexity is identical to that of the standard neural network and LSTM, the introduction of the attention mechanism with several heads slightly increases training time. Nevertheless, the overall training duration remains within an acceptable range.

Table 7. Comparison of Computational Complexity Among Models

Model	Time Complexity	Space Complexity
Linear Regression	O(n)	O(1)
Support Vector Machine	$O(n^3)$	$O(n^2)$
Deep Neural Network	$O(n^2)$	O(n)
LSTM	$O(n^2)$	O(n)
Multi-Head Attention LSTM	$O(n^2)$	O(n)

Figure 12 illustrates the distribution of prediction errors for the multi-head attention LSTM-based academic performance prediction model on the test dataset. The axis that runs horizontally is the magnitude of prediction errors,

while the vertical axis indicates the frequency of error occurrence. The figure reveals that prediction errors are predominantly concentrated between -5 and 5, exhibiting an approximately normal distribution shape. This indicates that the model's prediction errors are relatively small and demonstrate a certain degree of stability.

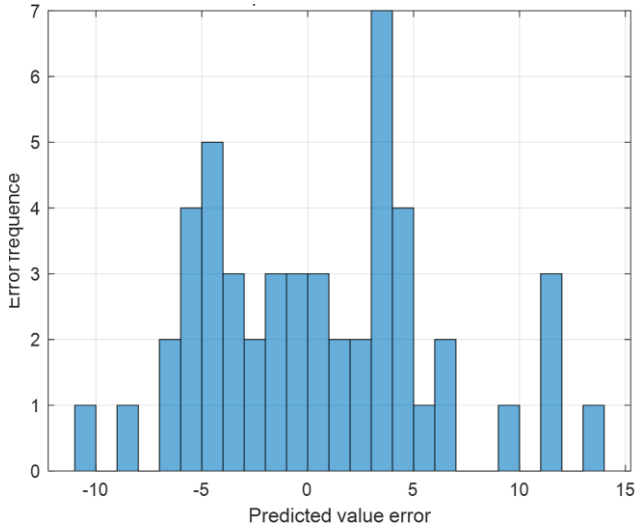

Figure 12. Prediction Error Distribution Map

Table 8 demonstrates that the multi-head attention LSTM-based academic performance prediction model exhibits outstanding performance across different subjects. The forecast mistakes for mathematics, Chinese, English, physics, and chemistry are 3.56, 4.23, 3.45, 4.78, and 4.56, respectively, all lower than those of other comparison models. This indicates the proposed model possesses strong versatility and adaptability, enabling accurate prediction of academic performance across diverse subjects.

Table 8. Performance Comparison of Models Across Subject Predictions

Subject	Linear Regression	Support Vector Machine	Deep Neural Network	LSTM	Multi-Head Attention LSTM
Math	7.85	6.23	4.87	4.12	3.56
Chinese	8.42	7.15	5.67	4.89	4.23
English	7.21	5.89	4.56	3.98	3.45
Physics	9.12	7.65	6.23	5.34	4.78
Chemistry	8.76	7.34	5.89	5.12	4.56

Figure 13 illustrates the attention weight distribution within the multi-head attention mechanism of the proposed academic performance prediction model. The horizontal axis is a representation of input features, while the vertical axis denotes attention weight values. The figure reveals that the model assigns higher attention weights to features such as historical grades and homework completion rates, indicating their significant influence on academic performance prediction. Analyzing the attention weight distribution can provide more targeted guidance and recommendations for both teachers and students.

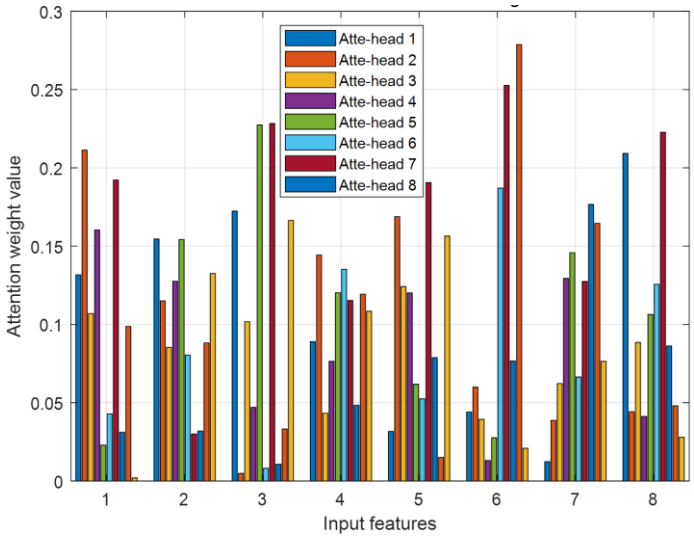

Figure 13. Attention Weight Distribution Map

6. Conclusion

This paper successfully constructs and validates an academic performance prediction model based on multi-head attention LSTM, addressing the deficiencies of existing educational data mining methods in handling long-sequence dependencies and identifying feature importance. The study first preprocessed multi-source heterogeneous data through Z-score standardization and one-hot encoding, followed by the design of a deep learning architecture incorporating multi-head attention and LSTM layers.

Experimental results demonstrate that the proposed model outperforms traditional models across all evaluation metrics. Specifically, compared to the standard LSTM model, the MAE decreased by 0.86, the RMSE decreased by 1.11, and the R^2 increased to 0.94. This significant improvement confirms the effectiveness of the multi-head attention mechanism in capturing critical learning behavior features, such as homework completion rates and historical grade trends, while also validating the robustness of LSTM in processing time-series data. Compared to shallow models like linear regression and support vector machines, this model exhibits stronger nonlinear fitting capabilities and generalization potential.

This study not only provides educational administrators with a precise academic early warning tool to support scientific decision-making and resource allocation but also offers data support for personalized teaching interventions. Future work will focus on integrating more modal data, such as images and speech, to further enhance the model's interpretability and exploring the practical deployment and application of this model in large-scale online education platforms.

7. Declarations

7.1. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.2. Funding and Acknowledgments

This work was supported by the 2023 Annual Jiaxing Applied Basic Research Project (Grant No. 2023AY11026), the Research and Creation Project of Zhejiang Provincial Department of Culture Radio Television and Tourism for the Year 2024–2025 (Grant No. 2024KYY029), the Jiaxing Nanhu University 2024 Education and Teaching Reform Research Project (Grant No. 22042024107), the 2025 Annual Higher Education Research Project of Zhejiang Higher Education Association(Grant No. KT2025186).

7.3. Institutional Review Board Statement

Not applicable.

7.4. Informed Consent Statement

Not applicable.

7.5. Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

[1] Singh, A. K., & Kumar, A. (2025). Multi-objective: hybrid particle swarm optimization with firefly algorithm for feature selection with Leaky ReLU. Discover Artificial Intelligence, 5(1), 192. doi:10.1007/s44163-025-00428-0.

[2] Chen, M., Li, Y., Li, N., Zhong, Y., & Guo, G. (2024). Dangerous attack paths analysis for power networks based on adaptive limited depth search and improved Z-score pruning. Ain Shams Engineering Journal, 15(12), 103135. doi:10.1016/j.asej.2024.103135.

[3] Prasath, N. A. (2025). A Multi-Source Deep Learning Model Utilizing Campus Data to Enhance Early Academic Performance Prediction. Mathematical Modelling of Engineering Problems, 12(4), 1311–1320. doi:10.18280/mmep.120422.

[4] Jia, C., He, H., Zhou, J., Li, K., Li, J., & Wei, Z. (2024). A performance degradation prediction model for PEMFC based on bi-directional long short-term memory and multi-head self-attention mechanism. International Journal of Hydrogen Energy, 60, 133–146. doi:10.1016/j.ijhydene.2024.02.181.

[5] Li, F., Ma, G., Ju, C., Chen, S., & Huang, W. (2024). Data-driven forecasting framework for daily reservoir inflow time series considering the flood peaks based on multi-head attention mechanism. Journal of Hydrology, 645(Part B), 19. doi:10.1016/j.jhydrol.2024.132197.

[6] Ahmad, K., Iqbal, W., El-Hassan, A., Qadir, J., Benhaddou, D., Ayyash, M., & Al-Fuqaha, A. (2023). Data-driven artificial intelligence in education: A comprehensive review. IEEE Transactions on Learning Technologies, 17, 12-31. doi:10.1109/TLT.2023.3314610

[7] Liu, Q., Wang, J., Dai, H., Ning, L., & Nie, P. (2025). Bridge Structural Damage Identification Based on Parallel Multi-head Self-attention Mechanism and Bidirectional Long and Short-term Memory Network. Arabian Journal for Science and Engineering, 50(3), 1803–1815. doi:10.1007/s13369-024-09035-0.

[8] Wang, J., Zhang, D., Huang, Q., & Cui, Z. (2025). Multiple-step accurate prediction of wave energy: A hybrid model based on quadratic decomposition, SSA and LSTM. International Journal of Green Energy, 22(1), 100–123. doi:10.1080/15435075.2024.2406849.

[9] Abín, A., Núñez, J. C., Rodríguez, C., Cueli, M., García, T., & Rosário, P. (2020). Predicting Mathematics Achievement in Secondary Education: The Role of Cognitive, Motivational, and Emotional Variables. Frontiers in Psychology, 11. doi:10.3389/fpsyg.2020.00876.

[10] Xu, F., & Qu, S. (2021). Data mining of students' consumption behaviour pattern based on self-attention graph neural network. Applied Sciences (Switzerland), 11(22), 10784. doi:10.3390/app112210784.

[11] Alturki, S., Alturki, N., & Stuckenschmidt, H. (2021). Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions. Journal of Information Technology Education: Innovations in Practice, 20, 121–137. doi:10.28945/4835.

[12] Wang, Y., & Lewis, S. E. (2022). Towards a theoretically sound measure of chemistry students' motivation; Investigating rank-sort survey methodology to reduce response style bias. Chemistry Education Research and Practice, 23(1), 240–256. doi:10.1039/d1rp00206f.

[13] Li, G., & Gao, W. (2022). Achievement Prediction of English Majors Based on Analytic Hierarchy Process and Genetic Algorithm. Mobile Information Systems, 2022, 6542300 1–6542300 10. doi:10.1155/2022/6542300.

[14] Gatzka, T. (2021). Aspects of openness as predictors of academic achievement. Personality and Individual Differences, 170. doi:10.1016/j.paid.2020.110422.

[15] Hamdani, M. (2021). Predicting Academic Achievement of Students: A Kolb's Learning Styles Approach. Journal of Educational, Health and Community Psychology, 10(4), 623–641.

[16] Véronneau, M. H., Vitaro, F., Poulin, F., Ha, T., & Kornienko, O. (2025). Academic Achievement, Externalizing Problems, and Close Friends in Middle School: Testing a Developmental Cascade Model Leading to Educational Attainment in the Late Twenties. Journal of Youth and Adolescence, 54(6), 1489–1505. doi:10.1007/s10964-025-02143-6.

[17] Zhang, X., & Yue, J. (2024). Predictive Modeling of Student Performance through Classification with Gaussian Process Models. International Journal of Advanced Computer Science and Applications, 15(6), 1214–1227. doi:10.14569/IJACSA.2024.01506123.

[18] Meruelo, A. D., Castro, N., Nguyen-Louie, T., & Tapert, S. F. (2020). Substance use initiation and the prediction of subsequent academic achievement. Brain Imaging and Behavior, 14(6), 2679–2691. doi:10.1007/s11682-019-00219-z.

[19] Brzozka, B. (2025). Machine Learning Algorithms in Predicting College Students' Grades: A Review. Journal of Applied Automation Technologies, 3, 1–12. doi:10.64972/jaat.2025v3.1.

[20] Nuankaew, P., & Nuankaew, W. S. (2022). Student Performance Prediction Model for Predicting Academic Achievement of High School Students. European Journal of Educational Research, 11(2), 949–963. doi:10.12973/EU-JER.11.2.949.

[21] Eltayar, A., Aref, S. R., Khalifa, H. M., & Hammad, A. S. (2023). Prediction of Graduate Learners' Academic Achievement in an Online Learning Environment Using a Blended Trauma Course. Advances in Medical Education and Practice, 14(000), 137–144. doi:10.2147/AMEP.S401695.

[22] Yildiz, M., & Börekci, C. (2020). Predicting Academic Achievement with Machine Learning Algorithms. Journal of Educational Technology and Online Learning, 3(3), 372–392. doi:10.31681/jetol.773206.

[23] Willems, J., van Daal, T., van Petegem, P., Coertjens, L., & Donche, V. (2021). Predicting freshmen's academic adjustment and subsequent achievement: Differences between academic and professional higher education contexts. Frontline Learning Research, 9(2), 28–49. doi:10.14786/flr.v9i2.647.

[24] Bird, K. A., Castleman, B. L., & Song, Y. (2025). Are algorithms biased in education? Exploring racial bias in predicting community college student success. Journal of Policy Analysis and Management, 44(2), 379–402. doi:10.1002/pam.22569.

[25] Zhang, Z., Deng, Q., He, W., & Yang, C. (2024). A New Method Based on Belief Rule Base with Balanced Accuracy and Interpretability for Student Achievement Prediction. Mathematics, 12(20), 3283. doi:10.3390/math12203283.

[26] Wallace, E. R., Ola, C., Leroux, B. G., Speltz, M. L., & Collett, B. R. (2021). Prediction of school age IQ, academic achievement, and motor skills in children with positional plagiocephaly. Paediatrics and Child Health (Canada), 26(3), E132–E137. doi:10.1093/pch/pxaa012.

[27] Pérez-González, J. C., Filella, G., Soldevila, A., Faiad, Y., & Sanchez-Ruiz, M. J. (2022). Integrating self-regulated learning and individual differences in the prediction of university academic achievement across a three-year-long degree. Metacognition and Learning, 17(3), 1141–1165. doi:10.1007/s11409-022-09315-w.

[28] Kumari, D. N. (2023). Contributory role of self-esteem, academic achievement and vocational aspirations in prediction of career attitude. International Journal of Arts, Humanities and Social Studies, 5(1), 107–110. doi:10.33545/26648652.2023.v5.i1b.54.

[29] Kawakami, K., Procopio, F., Rimfeld, K., Malanchini, M., von Stumm, S., Asbury, K., & Plomin, R. (2024). Exploring the genetic prediction of academic underachievement and overachievement. NPJ Science of Learning, 9(1), 39. doi:10.1038/s41539-024-00251-9.

[30] Corredor Valderrama, M. C., de Ruiz Sandoval, D. S., & Gómez, S. M. (2025). Prediction of Academic Achievement of University Students: Learning Styles and Strategies. Electronic Journal of Research in Educational Psychology, 23(1), 51–72. doi:10.25115/ZM2Q9N35.

[31] Iines R, P., Sami J, M., Vesa M, N., & Hannu K, S. (2023). ADHD symptoms and maladaptive achievement strategies: the reciprocal prediction of academic performance beyond the transition to middle school. Emotional and Behavioural Difficulties, 28(1), 3–17. doi:10.1080/13632752.2023.2189404.