Research on Student Academic Performance Prediction and Improvement Strategies Based on LSTM

Jianwei Huang[1] and Yanyu Huang[2*]

[1] Fujian Chuanzheng Communications College, Fuzhou, China
[2] Quanzhou Normal University, Quanzhou, China

*15559110766@163.com

Abstract. This study proposes an academic early warning and intervention framework based on LSTM and behavioral features. The framework first leverages study duration, attendance rates, sleep hours, and historical grades to construct a multi-dimensional dataset and trains an LSTM prediction model, which achieves an RMSE of 3.1747 on the test set, demonstrating high predictive precision. Furthermore, the study transforms the trained model into an interactive simulation environment to conduct feature intervention simulations targeting three student personas: "High Effort, Low Efficiency," "Insufficient Investment," and "Weak Foundation." Experiments demonstrate that tailored improvement strategies can significantly enhance predicted performance, yielding a maximum score increase of 8.9 points, thereby achieving a transition from static grade prediction to dynamic personalized strategy generation.

Keywords: Student, Long Short-Term Memory networks, Academic performance prediction, Improvement strategies.

1 Introduction

Against the backdrop of the deep integration of information technology and education, precise academic performance prediction serves as a critical foundation for building active early warning systems and realizing personalized learning support. It facilitates a paradigm shift in education from reactive remediation to proactive intervention, thereby optimizing resource allocation [1].

However, current academic guidance faces dual challenges at both practical and research levels. On one hand, traditional methods rely heavily on teacher experience, characterized by limitations such as response lag and restricted coverage [2]. On the other hand, prediction research based on machine learning (e.g., Decision Trees, SVM) typically halts at outputting scores, failing to translate results into actionable and interpretable behavioral recommendations, thus creating a disconnect between prediction and intervention [3,4].

To address this, this study proposes a "Prediction-Simulation" integrated framework, aiming to transform static prediction models into dynamic strategy inference platforms. First, an LSTM prediction model is constructed to capture the complex relationships

I. A. Khan et al. (eds.), *Proceedings of the 2026 5th International Conference on Educational Innovation and Multimedia Technology (EIMT 2026)*, Atlantis Highlights in Social Sciences, Education and Humanities 51,
https://doi.org/10.2991/978-94-6239-691-3_21

between behaviors and grades. Subsequently, leveraging the model as a virtual experimental environment, we adjust key behavioral features (e.g., study duration, attendance) for typical student personas to quantitatively simulate the academic improvement effects of different intervention strategies, thereby bridging the gap between prediction and action.

2 Dataset Introduction

2.1 Data Structure

The study utilizes a dataset sourced from the Kaggle public repository, which captures critical lifestyle and academic behavioral characteristics influencing student performance. The dataset comprises five core variables: the target variable is the final grade (exam_score); the predictive features include study duration (hours_studied) and attendance rate (attendance_percent), which characterize study engagement; historical grades (previous_scores), reflecting the foundation of prior knowledge; and sleep duration (sleep_hours), representing physiological state. Collectively, these variables constitute a feature space for investigating the relationships among behavioral patterns, foundational knowledge, and academic achievement.

2.2 Correlation Analysis

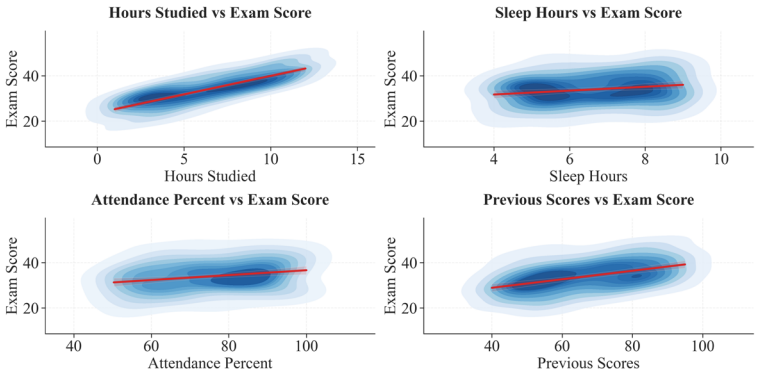

Fig. 1. Feature correlation analysis.

Figure 1 illustrates the two-dimensional Kernel Density Estimation (KDE) and linear regression trends between each feature and academic performance. Study duration and previous scores exhibit a pronounced strong positive correlation, with probability density cores concentrated along the diagonal and significant regression slopes, indicating that these are the dominant determinants of academic performance. In contrast, while attendance rate correlates positively with scores, its distribution appears more dispersed, reflecting substantial individual heterogeneity. The association between sleep duration and performance is the weakest, characterized by a flat regression trend and

diffuse distribution; this suggests that the direct impact of merely increasing sleep is limited, as it functions primarily as a foundational safeguard. The Pearson correlation coefficient between characteristics and academic performance is presented in Table 1.

Table 1. correlation analysis

Hours_studied	Sleep_hours	Attendance_percent	previous
0.78	0.19	0.23	0.43

3 Academic Performance Prediction and Improvement Strategies

3.1 Construction of the LSTM Prediction Model

The core of the LSTM lies in the design of its memory cell, which regulates information flow and storage through three logical control units: the forget gate [5,6], the input gate, and the output gate, as illustrated in Figure 2. The computational flow of the model at time step t is as follows:

Forget Gate: Decides how much information to discard from the previous cell state C_{t-1} , as shown in Equation (1).

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{1}$$

Input Gate and Candidate State: Determines how much new input information to update into the cell state, as shown in Equations (2) and (3).

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{3}$$

Cell State Update: Combines the retained historical information with the new input information, as shown in Equation (4).

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{4}$$

Output Gate and Hidden State: Determines the final output value based on the current state, as shown in Equation (5) and (6).

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{5}$$

$$h_t = o_t \odot \tanh(C_t) \tag{6}$$

where, σ represents the Sigmoid activation function, and $\odot$ denotes the Hadamard Product (element-wise multiplication of matrices or vectors); W and b respond to the weight matrices and bias terms, respectively.

The model output is a single scalar representing the predicted final exam score. The cleaned dataset of 200 samples was randomly partitioned into a training set and a test set at a ratio of 8:2. The training set is utilized for the iterative optimization of model parameters, while the test set is reserved exclusively for the independent evaluation of final performance and is excluded from the training process.

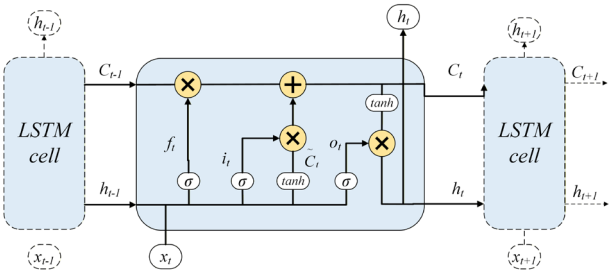

Fig. 2. LSTM unit structure

3.2 Academic Performance Prediction and Improvement Strategies

Building upon the constructed high-precision LSTM model, this study transforms the static prediction model into a dynamic simulation environment. By implementing targeted fine-tuning of student learning behavioral features at the input layer, this strategy precisely identifies the most effective behavioral adjustment scheme for the current state.

As illustrated in Table 1, based on data characteristics, the study categorizes the target student population into three typical personas: the "High Effort, Low Efficiency" type (characterized by high study engagement combined with insufficient sleep or attendance), the "Insufficient Investment" type (defined by a severe shortage of study duration), and the "Weak Foundation" type (marked by low historical grades). The data presented in Table 2 correspond to real-world samples, specifically selected from the dataset as representative cases for these three typical personas. For these distinct categories, the system executes differentiated simulations to generate personalized behavioural recommendations that offer the highest cost-effectiveness ratio.

Table 2. Typical student personas

Type	Hours Studied (h)	Sleep Hours (h)	Attendance Rate (%)	Previous Scores	Exam Score
High Effort, Low Efficiency	11.5	4.3	74.4	77	39.2
Insufficient Investment	2.1	8.3	50.3	75	26.5
Weak Foundation	7.1	5.8	63.6	46	27.1

4 Results and Discussion

4.1 LSTM Prediction Results

To evaluate the model's robustness and address concerns regarding dataset size, we conducted cross-validation with varying training-test ratios (Table 3). The model achieved optimal performance at an 8:2 split (R^2=0.8801, RMSE=3.1747). While accuracy naturally declined with smaller training sets, the degradation was gradual, confirming that the LSTM model remained stable and did not suffer from severe overfitting. Visual analysis (Figure 3) further corroborates this, showing that the predicted values closely track the actual scores and that the kernel density estimation (KDE) curves overlap.

As shown in Table 4, the LSTM model achieved an RMSE of 3.1747, comparable to that of the Random Forest (RMSE = 3.1023). While the current dataset relies on static behavioural summaries, student learning is inherently a sequential process. We prioritised the LSTM architecture to establish a scalable framework that seamlessly integrates future time-series data streams. Unlike traditional models (e.g., Random Forest), this approach allows for the direct incorporation of sequential data as it becomes available, making LSTM a more forward-looking choice for dynamic educational monitoring.

Table 3. Cross-validation

Training set: Test set	RMSE	R^2
8:2	3.1747	0.8801
7:3	3.2015	0.8716
6:4	3.5261	0.8026

Table 4. Model Comparison

Model	RMSE	R^2
LSTM	3.1747	0.8801
Random Forest	3.1023	0.8762
Linear regression	3.6014	0.8121

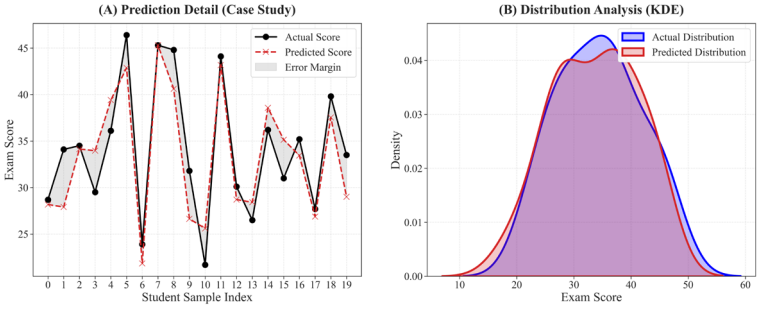

Fig. 3. LSTM grade prediction results.

4.2 Personalized Improvement Strategies

Leveraging the high-precision LSTM simulation environment, this study conducted targeted intervention experiments on the three student categories defined in Section 3.2. As presented in Table 5, the results quantify the score improvement effects of distinct strategies and reveal corresponding pedagogical patterns.

Table 5. Personalized improvement strategy

Type	Baseline Score	Predicted Score	Improvement
High Effort, Low Efficiency	39.2	42.1	2.9
Insufficient Investment	26.5	34.9	8.4
Weak Foundation	27.1	36.0	8.9

Figure 4 shows the comparison of results before and after optimization through the improvement strategy. For "High Effort, Low Efficiency" students, reducing study time in the simulation and equivalently converting it to sleep resulted in score increases rather than decreases. This demonstrates that under conditions of fatigue, adequate rest is more effective than the mere prolongation of study hours, thereby validating the feasibility of "efficiency enhancement through burden reduction." For "Insufficient Investment" students, a substantial increase in study duration directly yielded a significant score improvement of 8.4 points, indicating that augmenting absolute study time remains the most direct means of performance enhancement for this cohort. For "Weak Foundation" students, a systemic intervention—synchronously upgrading historical grades while ensuring adequate sleep and attendance—achieved the maximum improvement magnitude (8.9 points). This suggests that realizing substantial breakthroughs requires comprehensive remediation to compensate for historical academic deficits.

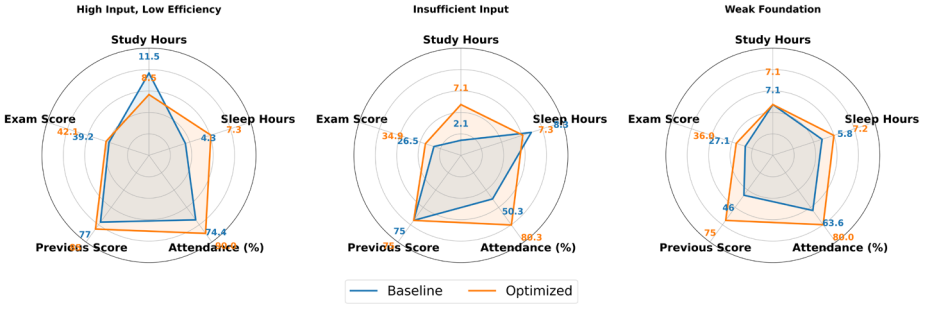

Fig. 4. Comparison of grades before and after

5 Conclusion

Addressing the practical dilemma of "easy prediction but difficult intervention" in the education sector, this study proposes a data-driven pathway for personalized student assistance. The primary conclusions are as follows: (1) The LSTM model demonstrates

exceptional modeling capabilities for educational behavioral data, achieving an R^2 of 0.8801 and an RMSE of 3.1747 on the test set, indicating high predictive precision. The proposed framework transforms the predictive model into an intervention deduction platform, enabling the planning of efficient improvement paths for students through simulation with low trial-and-error costs.

References

1. Liu, C., Wang, H., Du, Y., Yuan, Z.: A Predictive Model for Student Achievement Using Spiking Neural Networks Based on Educational Data. Applied Sciences. 12, 3841 (2022). https://doi.org/10.3390/app12083841.
2. 2Bujang, S.D.A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., Ghani, N.A.Md.: Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. IEEE Access. 9, 95608–95621 (2021). https://doi.org/10.1109/ACCESS.2021.3093563.
3. 3Doz, D., Cotič, M., Felda, D.: Random Forest Regression in Predicting Students' Achievements and Fuzzy Grades. Mathematics. 11, 4129 (2023). https://doi.org/10.3390/math11194129.
4. Kukkar, A., Mohana, R., Sharma, A., Nayyar, A.: A novel methodology using RNN + LSTM + ML for predicting student's academic performance. Educ Inf Technol. 29, 14365–14401 (2024). https://doi.org/10.1007/s10639-023-12394-0.
5. Abirami T, R. Vadivel: Student semester marks prediction using linear regression algorithms in machine learning. World J. Adv. Res. Rev. 18, 469–475 (2023). https://doi.org/10.30574/wjarr.2023.18.1.0591.
6. Qiu, W., Khong, A.W.H., Supraja, S., Tang, W.: A Dual-Mode Grade Prediction Architecture for Identifying At-Risk Students. IEEE Trans. Learning Technol. 17, 803–814 (2024). https://doi.org/10.1109/TLT.2023.3333029.