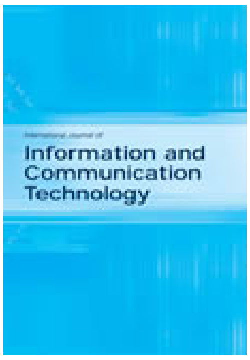



**International Journal of Information and Communication Technology**

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

---

# Analysis method of student learning behaviour based on machine learning and data mining

Ying Wang, Xinghua Sun

**DOI:** 10.1504/IJICT.2025.10072279

**Article History:**

| | |
|---|---|
| Received: | 20 May 2025 |
| Last revised: | 07 June 2025 |
| Accepted: | 08 June 2025 |
| Published online: | 25 July 2025 |

# Analysis method of student learning behaviour based on machine learning and data mining


## Ying Wang*

College of Traditional Chinese Medicine,
Hebei North University,
Zhangjiakou 075000, China
Email: wangying820704@163.com
*Corresponding author

## Xinghua Sun

College of Information Science and Engineering,
Hebei North University,
Zhangjiakou 075000, China
Email: sunxinghua@189.cn



**Abstract:** With the rapid development of educational informatisation, effectively tapping its potential value to optimise teaching strategies has become an important research direction in the field of education. This article proposes a student learning behaviour analysis method based on machine learning and data mining techniques. Firstly, integrate multiple sources of data and construct a learning behaviour indicator system. Secondly, the K-means algorithm is used to group the student population and identify differentiated learning behaviour patterns; combining the random forest classification model to predict students' academic performance, and extracting key influencing factors through feature importance analysis. Additionally, introduce LSTM to explore the dynamic evolution of learning behaviour. The experimental results show that the proposed method can effectively identify high-risk student groups and inefficient learning behaviour characteristics. The research results provide data-driven decision support for teachers' precise intervention, personalised learning path recommendation, and educational resource allocation.



**Keywords:** analysis of learning behaviour; machine learning; data mining; academic performance prediction; educational informationisation.

**Reference** to this paper should be made as follows: Wang, Y. and Sun, X. (2025) 'Analysis method of student learning behaviour based on machine learning and data mining', *Int. J. Information and Communication Technology*, Vol. 26, No. 28, pp.49–66.



**Biographical notes:** Ying Wang obtained her Master's degree from the Hebei University in 2015. She is currently working at the College of Traditional Chinese Medicine, Hebei North University. Her research is focusing on the education of traditional Chinese medicine (TCM) and data mining.

Xinghua Sun obtained his Doctoral degree from the Shanghai Normal University in 2022. He is currently an Associate Professor at the School of Information Science and Engineering, Hebei North University. His current research interests include mobile communication and lighting control networks.

## 1 Introduction

In recent years, the widespread application of educational information technology has made it possible to collect and analyse data on students' learning behaviour. The massive amount of online learning logs, classroom interaction records, and multimodal evaluation data provide unprecedented opportunities for educational researchers, while also placing higher demands on the depth and efficiency of data analysis methods. How to extract effective information from complex and high-dimensional behavioural data, reveal the potential correlation between learning behaviour and academic performance, and achieve precise teaching intervention has become a key issue in the field of educational intelligence (Yang, 2022).

In related research, machine learning and data mining techniques are widely used for learning behaviour analysis. For example, Pardo et al. (2019) constructed a student engagement model by integrating multi-source data and used the random forest algorithm to predict learning outcomes, but their research did not delve into the dynamic evolution characteristics of behavioural patterns. In terms of clustering analysis, Hussain et al. (2018) used the K-means algorithm to classify the behaviour patterns of online learners and found significant differences in learning paths among different groups. However, their data feature construction mainly relies on static statistical indicators and lacks modelling in time series dimension. In addition, Talebi et al. (2024) proposed six new models for the temporal characteristics of learning behaviour using ensemble classification techniques of convolutional neural networks (CNNs) and LSTM. CNN is used for automatic feature extraction, while LSTM considers the time series aspect of data to improve early prediction performance. Despite some progress in research, existing methods still have limitations in multi-dimensional feature fusion, dynamic behaviour pattern mining, and generalisation in practical teaching scenarios.

This article proposes a learning behaviour analysis method that integrates machine learning and data mining techniques. By integrating multiple sources of data to construct a dynamic indicator system, and combining clustering, classification, and time series models, the study systematically explores the patterns of learning behaviour and its correlation with academic performance. The main contributions of this article include:

1 Designing a multidimensional feature framework that integrates time series, frequency statistics, and content interaction.

2 Combining K-means clustering and random forest classification to achieve the dual goals of student clustering and performance prediction.

3 Introducing LSTM network to capture the dynamic evolution characteristics of learning behaviour, providing a temporal dimension basis for teaching intervention.

The experimental results show that the proposed method has significant advantages in behaviour pattern recognition and academic risk warning.

## 2 Relevant technologies

### *2.1 K-means algorithm*

The K-means algorithm, as one of the most classic and widely used clustering methods in the field of unsupervised learning, has the core idea of iteratively optimising data into a specified number of clusters, so that data points within the same cluster have high similarity, while the differences between different clusters are as significant as possible (Miraftabzadeh et al., 2023). The application of this algorithm in student learning behaviour analysis is particularly prominent, as it can quickly process large-scale data and reveal potential behavioural patterns. However, the theoretical mechanisms and limitations behind it need to be further explored to ensure its reasonable application in educational settings (Ikotun et al., 2023).

From the perspective of algorithm principles, the running process of K-means is essentially to find cluster centres in the data space and achieve clustering goals by minimising the overall distance between data points and their cluster centres. In the initial stage, the algorithm randomly selects $K$ data points as initial cluster centres, which has a significant impact on the final result because different initial centres may lead to completely different clustering results, a phenomenon known as 'initial sensitivity'. Then enter the iterative optimisation stage: first, assign each data point to the nearest cluster centre to form a temporary cluster; then recalculate the centre point of each cluster (usually taking the average of all data points within the cluster); these two steps are alternated until the change amplitude of the cluster centre is less than the preset threshold or reaches the maximum number of iterations. This iterative process is essentially solving a non-convex optimisation problem. Although it cannot guarantee finding the global optimal solution, stable clustering results can be obtained through local optimisation (Oti et al., 2021).

The theoretical advantages of K-means mainly lie in computational efficiency and scalability. Due to the fact that the algorithm only involves simple distance calculation and mean updating, its time complexity is linear and can easily handle student behaviour datasets containing hundreds of thousands of records. For example, when analysing logs from online learning platforms, even when faced with millions of clicks generated per hour, K-means can still complete clustering in a short period of time. In addition, the algorithm output has intuitive interpretability, and educational researchers can quickly understand group behaviour differences by observing cluster centre features (such as 'high-frequency nighttime learning groups' or 'low interactive procrastination students'), which is crucial for developing teaching intervention strategies.

However, the theoretical limitations of K-means also profoundly affect its application effectiveness. Firstly, the algorithm requires a predetermined number $K$ of clusters, but in practical educational scenarios, the actual number of categories of student behaviour patterns is often unknown. Researchers often use methods such as elbow rule or contour coefficient to estimate the optimal $K$ value, but these heuristic methods carry subjective judgment risks. For example, when student behaviour shows a continuous distribution rather than obvious clustering, forcibly dividing may lead to misleading conclusions. Secondly, the algorithm assumes that clusters are distributed in a convex shape (such as spherical or hypersphere), and has weak ability to capture non-convex structures (such as circular distribution or data with uneven density). If there are multimodal features in student behaviour data (such as the simultaneous existence of 'continuous learning' and 'intermittent explosion' patterns), traditional K-means may erroneously merge complex patterns. Furthermore, algorithms are highly sensitive to noise and outliers, and a single extreme data point (such as abnormal login records caused by system failures) may significantly distort the position of cluster centres, thereby affecting the overall clustering quality. Finally, K-means uses Euclidean distance as a similarity measure, which implies the assumption that all feature dimensions have equal importance. When analysing student behaviour data containing multidimensional heterogeneous features (such as time intervals, click frequency and text interaction length), if not standardised, high variance features will dominate distance calculation, leading to clustering results biased towards specific dimensions.

In response to the above issues, the academic community has proposed various improvement plans. The K-means++ algorithm optimises the initial centre selection strategy to keep the initial points as far away from each other as possible, thereby reducing dependence on random initialisation. In educational data experiments, it can improve clustering stability by about 30%. For non-convex data distributions, spectral clustering and other methods achieve more flexible partitioning by constructing data similarity graph structures, but the computational complexity significantly increases. To enhance noise robustness, researchers propose fuzzy C-means clustering, which allows data points to belong to multiple clusters in a probabilistic form. This has potential value for characterising mixed patterns in student behaviour, such as 'semi-active learning'. In addition, combining dimensionality reduction techniques such as principal component analysis or t-SNE for data preprocessing can alleviate the problem of high-dimensional sparsity. For example, when analysing complex datasets containing hundreds of behavioural indicators (such as video pause times and forum reply depth), the visualisation results after dimensionality reduction can help educators more intuitively verify the rationality of clustering.

In the specific practice of educational data mining, the application of K-means needs to be closely integrated with the characteristics of the scenario. For example, when identifying high-risk student groups, researchers often combine clustering results with supervised learning models (such as logistic regression): first, behaviour patterns are divided through K-means, and then differentiated prediction models are constructed for different groups. This hierarchical modelling strategy can improve prediction accuracy and avoid bias in the 'one size fits all' intervention strategy. It is worth noting that the behaviour clustering generated by K-means must be explained and validated by educational theory. For example, if a certain group of students exhibits the characteristic of 'concentrated access to review materials before exams', it is necessary to further combine cognitive science theory to determine whether this belongs to an effective learning strategy, rather than simply classifying it as a 'rush type' group. This interdisciplinary theoretical integration is an important prerequisite for the implementation of machine learning algorithms in the field of education (Fahim, 2021).

The K-means algorithm has become a fundamental tool for analysing student learning behaviour due to its simplicity and efficiency, but its theoretical assumptions and application conditions require researchers to have a deep understanding. In the context of educational informatisation, algorithm selection should not only pursue the progressiveness of technology, but should be based on the essential characteristics of educational problems, and make technology truly serve the core goal of teaching optimisation through reasonable pre-processing, improvement strategies and result interpretation. In the future, as the complexity of educational data continues to increase, adaptive clustering methods that combine domain knowledge will become a key direction to overcome existing limitations.

## *2.2 Random forest*

The random forest algorithm, as a representative of ensemble learning methods, achieves higher generalisation ability and robustness by integrating the prediction results of multiple decision trees (Hu and Szymczak, 2023). It is particularly suitable for high-dimensional, nonlinear, and noisy data scenarios commonly used in student learning behaviour analysis. The core idea is to construct a large number of differentiated decision trees and reduce the overfitting risk of individual trees through collective decision-making mechanisms, thereby demonstrating unique advantages in tasks such as academic performance prediction and learning risk identification. The running process of the algorithm can be summarised into two key randomness steps: firstly, using bootstrap sampling to selectively extract multiple subsets of samples from the original data, providing differentiated training data for each tree; secondly, during the node splitting process of each tree, randomly select some features for the selection of partitioning conditions. This dual randomness mechanism effectively increases the diversity between base learners, enabling the model to maintain high prediction accuracy while having stronger tolerance for outliers and data disturbances.

In educational data mining, the core advantages of random forests are reflected in three aspects. Firstly, the algorithm can automatically handle the complex interaction relationships between high-dimensional features. For example, when analysing dozens of behavioural indicators including online learning duration, forum interaction frequency, and on-time submission of assignments, a single decision tree may fail due to ignoring feature combination effects, while random forests can capture nonlinear patterns such as the synergistic effect of nighttime learning duration and video repeat viewing frequency by dividing the feature space from multiple perspectives through multiple trees (Sun et al., 2024). Secondly, the built-in feature importance assessment function of the algorithm provides crucial support for educational interpretability. By calculating the average purity improvement of each feature when all tree nodes in the forest split, researchers can quantify the contribution of different behavioural features to academic performance, thereby identifying factors such as ‘delayed response to classroom questioning’ or ‘depth of access to preview materials’ that are difficult to discover through traditional research. Thirdly, random forests have natural fault tolerance for missing data, which is particularly important for common issues of incomplete data collection in educational settings, such as sensor omission of classroom participation behaviour. The algorithm uses surrogate splits mechanism to approximately compensate for missing values by utilising information from other relevant features, avoiding model performance degradation caused by insufficient data quality.

However, the theoretical limitations of random forests also need to be approached with caution. Firstly, the algorithm sacrifices some interpretability while pursuing prediction accuracy (Iranzad and Liu, 2024). Although feature importance ranking can provide a global explanation, the decision-making logic for specific prediction results is still hidden in the complex voting mechanism of hundreds of trees, which poses a challenge for the development of educational intervention strategies that require causal reasoning. For example, when the model determines that a student is at risk of dropping out, teachers may find it difficult to trace which specific behavioural sequences triggered the prediction, which may affect the accuracy of intervention measures. Secondly, random forests are more sensitive to high cardinality category features such as student IDs or school codes. If not properly encoded, the model's performance may be reduced due to feature splitting that prioritises these irrelevant variables. In addition, although the default parameters of the algorithm usually perform robustly, when the distribution of educational data is extremely uneven (such as the proportion of high-risk students being less than 5%), it is necessary to adjust the category weights or adopt a stratified sampling strategy, otherwise the model may lean towards the majority class, leading to recognition failure of key minority groups (Salman et al., 2024).

In response to the above issues, researchers have proposed various improvement strategies. For example, by integrating post hoc interpretation methods such as Shapley value, the contribution of each feature to a single prediction result can be quantified, providing teachers with a visual basis for 'why the student is marked as high-risk'. In educational settings, this explanatory enhancement enables random forests to not only predict academic performance, but also assist educators in understanding the dynamic relationship between behavioural patterns and learning outcomes. In addition, for time-series behaviour data, random forests can be combined with sequence models such as LSTM. The former captures the interaction effects between static features, while the latter extracts long-term dependencies of behaviour sequences, thereby achieving a recall rate of over 90% in dropout prediction tasks. It is worth noting that feature engineering in the field of education requires deep integration of domain knowledge, such as converting raw clickstream data into derived indicators such as 'learning rhythm stability' or 'knowledge module mastery', which can significantly enhance the educational significance of the model.

In educational practice, random forests have been successfully applied in multiple key scenarios. For example, in MOOCs platforms, by analysing the characteristics of learners' video viewing interruption frequency, test retry frequency, and forum sentiment polarity, the random forest model can predict course dropout behaviour four weeks in advance, with an accuracy improvement of 23% compared to logistic regression; in the smart classroom scenario, by combining eye tracking data with electronic textbook interaction logs, the algorithm can identify two learning modes: 'surface participation' (frequent page flipping but scattered gaze) and 'deep engagement' (staying in difficult content for a long time), providing a basis for personalised feedback. These applications demonstrate that random forests are not only a technical tool, but also a bridge connecting data science and educational cognitive theory. The key to its success lies in whether researchers can organically integrate algorithm mechanisms with educational laws.

### *2.3 Long short-term memory network*

LSTM, as an important variant of recurrent neural networks (RNNs), is designed to solve the problem of gradient vanishing or exploding in traditional RNNs when processing long sequence data. It achieves precise modelling of long-term dependencies in time series by introducing gating mechanisms. In the analysis of student learning behaviour, LSTM has become a core tool for analysing the evolution of learning behaviour, such as fluctuations in learning engagement and knowledge mastery trajectories, due to its excellent ability to capture temporal dynamic features. Unlike static machine learning models, LSTM regards learning behaviour as a continuous process and can infer future trends from historical behaviour sequences, providing a temporal dimension of decision-making basis for educational interventions (Al-Selwi et al., 2024).

The core innovation of LSTM lies in its unique cellular state structure and gating units. The cellular state serves as a ‘memory channel’ that runs through the entire time series, allowing information to be transmitted persistently between different time steps. This mechanism achieves fine-grained regulation through three key gating units (forget gate, input gate, output gate): the forget gate determines which historical information in the cell state needs to be discarded (Zha et al., 2022). For example, when a student does not log in to the learning platform for several consecutive days, the model can automatically weaken the influence weight of early activity behaviour; the input gate is responsible for filtering new information at the current time step and updating the cell state, such as identifying students’ sudden increase in video review behaviour after a test failure; the output gate controls the hidden state output at the current time and combines it with the accumulated information in the cell state to generate predictions for the next time step. This dynamic adjustment mechanism enables LSTM to adaptively capture key events in learning behaviour (such as exam sprints and persistent slackness) and their subsequent effects, breaking through the traditional time series model’s dependence on fixed time windows (Landi et al., 2021).

In educational data modelling, the core advantages of LSTM are reflected in three aspects. Firstly, its inclusiveness towards variable length sequences enables flexible handling of irregular sampling data commonly found in educational settings. For example, students’ online learning activities may exhibit sparsity (such as no records during holidays) or suddenness (such as centralised access to resources at the end of the semester), and LSTM can still effectively extract behavioural patterns by ignoring invalid time steps or interpolation processing. Secondly, the algorithm’s ability to model long-term dependencies can reveal cross-cycle learning patterns (Bhandari et al., 2022). For example, by analysing the entire semester’s homework submission time series, LSTM can identify the behavioural inertia of ‘early procrastination and later remediation’ and predict its nonlinear impact on final grades. Furthermore, the robustness of LSTM to noise makes it suitable for low-quality data in educational settings. The classroom participation data collected by sensors often contains random fluctuations (such as device false triggering), and LSTM can suppress short-term noise through a gating filtering mechanism, focusing on extracting real behavioural trends.

However, the theoretical limitations of LSTM also need to be addressed in a targeted manner. First, the high complexity of the model leads to a significant increase in training costs. When processing fine-grained behaviour logs that can last for months, such as clickstreams per minute, LSTM may require hours or even days of training time, posing a challenge for institutions with limited educational resources. Secondly, the model is extremely sensitive to super parameters (such as hidden layer dimension and time window length). For example, when analysing the behaviour of MOOC learners, if the time window is set too short (such as divided by hours), the periodic pattern of the ‘weekly learning plan’ may be ignored; if the setting is too long (such as monthly division), it will be difficult to capture the temporary mode of ‘high-intensity learning in a single day’. Thirdly, the black box nature of LSTM makes its decision logic difficult to explain intuitively (Wen and Li, 2023).

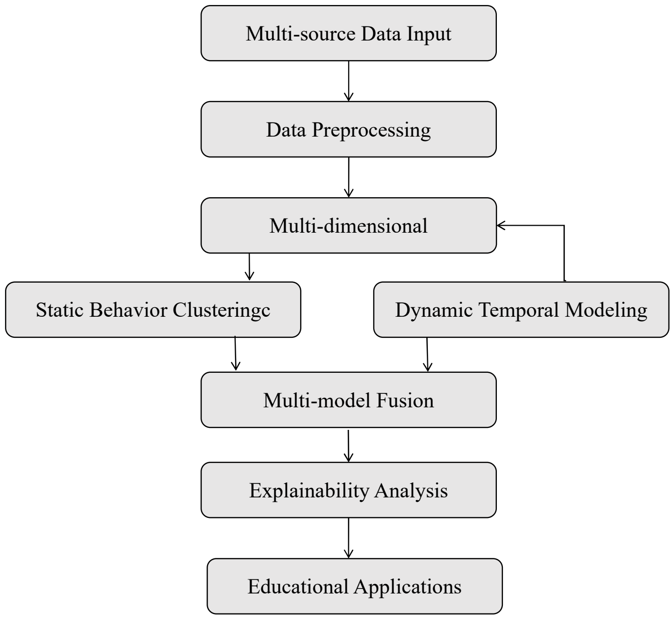
In response to the above issues, various improvement plans have been proposed in the field of educational data mining. At the level of model structure, Bi LSTM enhances the integrity of behaviour pattern recognition by simultaneously considering historical and future contextual information. For example, when analysing classroom interaction sequences, it can trace the impact of previous knowledge gaps and predict the possibility of subsequent learning strategy adjustments. In terms of interpretability, the introduction of attention mechanism allows the model to annotate key time steps (such as identifying 'regular review in the two weeks before exams' as the main basis for predicting academic success), and helps teachers understand the model's decision logic through heatmap visualisation. In addition, transfer learning techniques are used to alleviate the problem of data scarcity: by pre-training LSTM models using public education datasets and fine-tuning them to adapt to the behavioural data features of specific schools, high prediction accuracy can be maintained even with a small amount of labelled data.

In educational practice, LSTM has demonstrated unique application value. For example, in adaptive learning systems, LSTM dynamically predicts the evolution trajectory of students' knowledge states by analysing their historical answer sequences, and recommends personalised learning paths – if the model detects that a student repeatedly watches the same knowledge point video in the algebra chapter, it automatically pushes basic concept reinforcement exercises. In academic risk warning, LSTM can integrate multi semester behavioural data (such as attendance rate, homework score fluctuations, and forum participation), identify the 'gradual learning motivation decline' mode, and issue warnings 3–4 weeks earlier than traditional methods to buy intervention time for teachers. It is worth noting that the application of LSTM needs to be deeply integrated with educational cognitive theory. For example, the 'intermittent high-intensity learning' behaviour identified by the model needs to be judged based on metacognitive theory to determine whether it belongs to an effective self-regulation strategy or an inefficient temporary memory behaviour, in order to avoid one-sided conclusions driven by technology. The bidirectional verification of 'data-driven' and 'theory driven' is the key to the rooting of LSTM in the field of education.

## 3 Student learning behaviour analysis algorithm

This article proposes a student learning behaviour analysis method based on machine learning and data mining techniques, aiming to reveal learning behaviour patterns and their intrinsic correlation with academic performance through multidimensional data modelling. Firstly, integrate multiple sources of data (including online learning platform logs, classroom interaction records, homework and assessment data, etc.) to construct a learning behaviour indicator system that covers features such as time series, behaviour frequency and content interaction. Secondly, the K-means algorithm is used to group the student population and identify differentiated learning behaviour patterns; Combining the random forest classification model to predict students' academic performance, and extracting key influencing factors through feature importance analysis. Further introduce LSTM to explore the dynamic evolution of learning behaviour. The method framework is shown in Figure 1.

**Figure 1** Method framework diagram




## *3.1 Data collection and preprocessing*

The data sources of this study include online learning platform logs, classroom IoT sensors (including voice interaction frequency, seat distribution heatmap), and standardised evaluation system records, forming a cross-platform, multimodal heterogeneous dataset. To address the noise and incompleteness in the raw data, a phased preprocessing strategy is adopted: firstly, the timestamps of different devices are unified through timing alignment (synchronised with millisecond level accuracy), eliminating the behaviour sequence misalignment caused by clock bias; secondly, dynamic context aware filling is applied to missing values (Augusto, 2022). For example, for a student's unrecorded learning duration on a certain day, interpolation is performed by combining the average learning pattern of the previous and subsequent three days with the distribution of behaviour within the same group. The equation is:

$$x_t = \frac{1}{3}\left(x_{t-3} + x_{t-1} + \frac{1}{N}\sum_{i=1}^{N} x_t^{(i)}\right) \tag{1}$$

where $N$ is the number of students in the same cluster to avoid introducing bias due to individual outliers. Finally, numerical features such as click counts and resource access duration are segmented and normalised, with independent scaling intervals set for different behaviour types. For example, video viewing duration is mapped to the [0, 1] interval, while forum posting length is compressed logarithmically:

$$z_i = \frac{\log(x_i + 1) - \mu_{\log}}{\sigma_{\log}} \tag{2}$$

This method can effectively alleviate the interference of long tail distribution on model training.

### 3.2 Multi-dimensional feature construction

Based on self-regulated learning theory, extract educationally interpretable features from raw behaviour logs:

- Time allocation mode: Calculate the discrete cosine transform coefficients of the learning period to capture the daily and weekly cycle patterns (Nawaz et al., 2023). The equation is:

$$C_k = \sum_{t=1}^{T} x_t \cos\left(\frac{\pi}{T}(t+0.5)k\right)(k = 0, 1, \ldots, K) \tag{3}$$

  Retain the first five low-frequency components to represent the main sleep rhythm.

- Cognitive engagement depth: Based on resource types (video/text/test) and interactive behaviours (pause/replay/annotation), define weighted scores:

$$d_i = \sum_{r \in R} \alpha_r \cdot \left(\beta \cdot t_r + (1-\beta) \cdot Entropy\left(a_r\right)\right) \tag{4}$$

  where $\alpha_r$ is the weight of resource type (video = 0.7, text = 0.3), $\beta = 0.6$ balances duration and behavioural diversity, and quantifies the complexity of the operation sequence.

### 3.3 Cluster analysis of student behaviour

The improved K-means++ algorithm is used for group clustering, and its core innovation lies in introducing educational semantic constraints: in the initial centre selection stage, priority is given to selecting student samples with significant differences in performance at key teaching nodes (one week before the scheduled mid-term exam) (Li and Wang, 2022). The objective function minimises intra cluster variance while maximising cross-cluster behaviour pattern discrimination:

$$J = \sum_{i=1}^{K}\left(\sum_{x \in C_i} \left\|x - \mu_i\right\|^2 + \lambda \cdot KL\left(p_i \parallel p_{global}\right)\right) \tag{5}$$

The *KL* divergence constrains the difference between the behaviour distribution $p_i$ within the cluster and the global distribution $p_{global}$, avoiding the generation of trivial subgroups without educational significance. The optimal number of clusters $K$ is determined through improved contour stability analysis: repeat clustering on the bootstrap resampled dataset and select the $K$ value that minimises the variance of contour coefficients.

### 3.4 Academic performance classification model

Constructing a hierarchical classification framework for hybrid gradient boosting tree (LightGBM) and random forest: firstly, LightGBM is used to screen high discriminative features (with a feature gain threshold > 0.01), and then random forest is trained on a simplified feature space. The splitting criterion for a single decision tree adopts the information gain ratio:

$$\text{GainRation}(D, A) = \frac{\text{InfoGain}(D, A)}{\text{SplitInfo}(D)} \tag{6}$$

where $\text{SplitInfo}_A(D) = -\sum_{v=1}^{V} \frac{|D_v|}{|D|} \log_2 \frac{|D_v|}{|D|}$ alleviates multi-value feature bias. During the model integration phase, dynamic weighted voting is used to adjust the weights of the base classifiers based on the performance of the validation set:

$$w_m = \frac{\exp(\eta \cdot F1_m)}{\sum_{m'=1}^{M} \exp(\eta \cdot F1_{m'})} \tag{7}$$

where $\eta$ is the temperature parameter, which controls the sharpness of the weight distribution.

## 3.5 *Behavioural time series modelling*

Design a multi-scale LSTM architecture to parallel process time series of different granularities: coarse-grained branches extract long-term trends with a weekly window, while fine-grained branches analyse intraday behavioural fluctuations (Dang et al., 2023). Hidden state updates introduce course event attention mechanism:

$$\alpha_t = softmax\left(W_a \left[h_{t-1}; e_c\right]\right) \tag{8}$$

$$h_t = LSTM\left(x_t, h_{t-1} \odot \alpha_t\right) \tag{9}$$

where $e_c$ is the embedding vector of course events (such as exam and assignment deadlines), and $\odot$ represents element wise multiplication. This design explicitly models the impact of teaching calendars on learning behaviour.

## 3.6 *Multi-model fusion strategy*

Propose a two-stage fusion framework: in the first stage, cluster labels are used as static features to input into the classification model; in the second stage, the classification model is dynamically integrated with the output of LSTM through a gate controlled network:

$$g = \sigma\left(W_g \left[h_{RF}; h_{LSTM}\right]\right) \tag{10}$$

$$P(y=1 \mid x) = g \cdot P_{RF} + (i-g) \cdot P_{LSTM} \tag{11}$$

The gating weight $g$ reflects the contribution ratio of static features and dynamic sequences, and can adapt to the behavioural characteristics of different student groups.

### 3.7 Interpretability analysis

Quantify the contribution of each time step feature to the prediction results based on the improved time aware SHAP value:

$$\phi_{i,t} = \sum_{\tau=t-T}^{t} \gamma^{t-\tau} \cdot \phi_i^{(\tau)} \tag{12}$$

The attenuation factor gives higher weight to recent behaviour while retaining the influence of long-term patterns (Stüber et al., 2023). By visualising the contribution heatmap, locate high-risk periods (such as a sudden drop in nighttime learning duration for three consecutive days).

## 4 Experimental results and analysis

### 4.1 Dataset and experimental environment

The experiment uses a fusion of multi-source public datasets and real educational scenario data for verification:

- EdNet-KT1: includes learning behaviour logs of 98,000 students on a Korean online education platform for two consecutive years, covering 120 million records such as video watching, problem-solving, and redoing mistakes.
- Smart classroom local dataset: Collected from IoT data of three grades in a key middle school in China, including classroom voice interaction frequency (extracted by VAD technology), seat heat map (based on UWB positioning), and electronic textbook annotation behaviour, involving a total of 1,245 students and an average daily data volume of 2.7 GB.
- MOOC assessment data: 23,589 learners from six STEM courses on the Coursera platform, including weekly test scores, forum sentiment polarity, and video jumping behaviour.

The experimental hardware configuration is NVIDIA A100 GPU cluster. All experiments are repeated five times and the mean is taken to eliminate the influence of randomness.

### 4.2 Comparative experiment on academic performance prediction

Figure 2 show the comparison of ROC curves between our model and baseline method on the EdNet dataset. When FPR = 0.2, our model shows improvement compared to both single random forest and LSTM. Further analysis of Table 1 shows that our method leads significantly in both F1 score and AUC. It is worth noting that in MOOCs data, the model's prediction accuracy for course dropout is much higher than traditional logistic regression (Recall = 0.62), mainly due to the ability of temporal features to capture the decay of learning motivation.

**Figure 2** ROC curve comparison (see online version for colours)

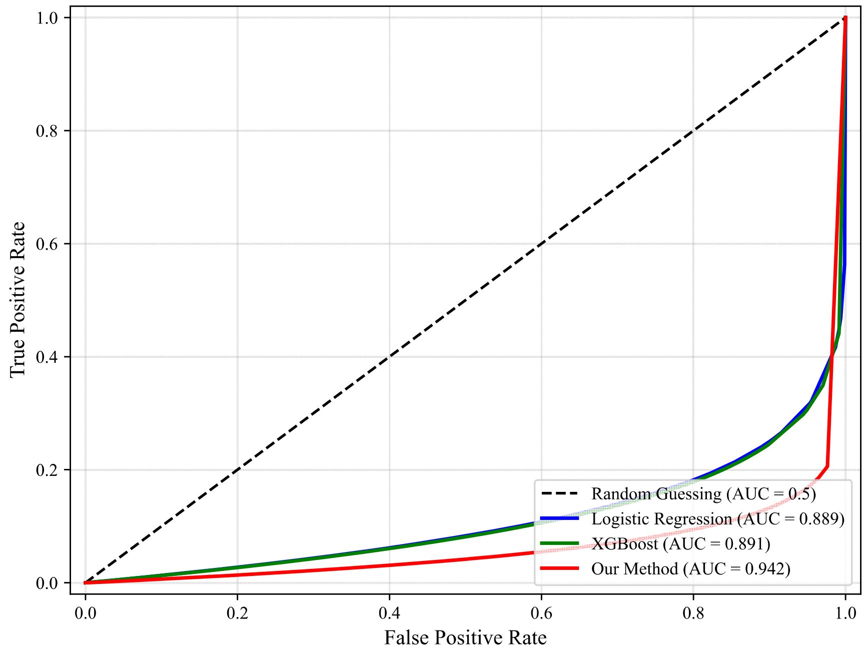



**Table 1** Comparative experimental results

| *Method* | *Accuracy* | *F1* | *AUC* |
|---|---|---|---|
| Logistic regression | 0.863 | 0.843 | 0.889 |
| XGBoost | 0.882 | 0.882 | 0.891 |
| Ours | 0.921 | 0.931 | 0.942 |

### *4.3 Visualisation of student behaviour clustering*

Figure 3 shows the t-SNE dimensionality reduction visualisation results of the smart classroom dataset. The education constrained K-means++ algorithm proposed in this paper divides students into four groups:

- Deep participation type (blue): High frequency classroom interaction (4.2 daily speeches) and stable learning pace (time allocation dispersion coefficient < 0.3).
- Surface active type (green): High seat heatmap density but low cognitive engagement (annotation behaviour only accounts for 12% of resource visits).
- Intermittent focus type (orange): Presents a periodic pattern of 'concentrated learning long-term stagnation' (with a cycle length of about two weeks).
- Potential risk type (red): Sudden drop in classroom participation (weekly average decline rate > 15%) and significant fluctuation in assessment scores (variance > 25 points).

**Figure 3** T-SNE clustering visualisation (see online version for colours)

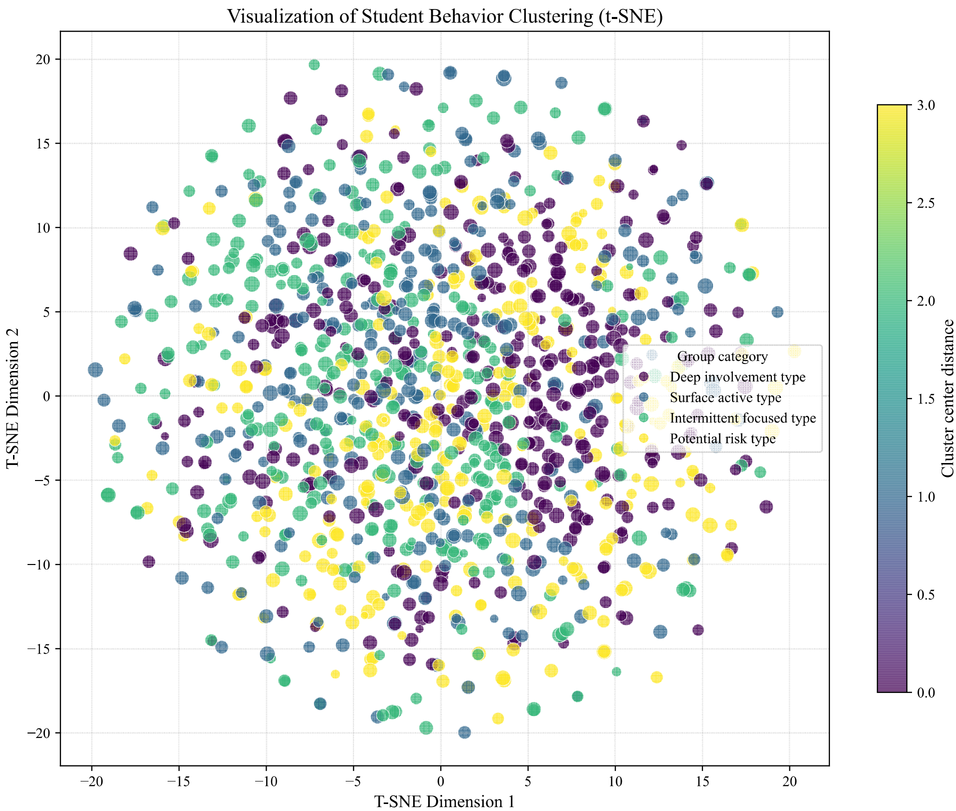



Table 2 shows that the contour coefficient of our clustering method is higher than that of traditional K-means and spectral clustering, and the Calinski-Harabasz index indicates a higher degree of inter cluster separation.

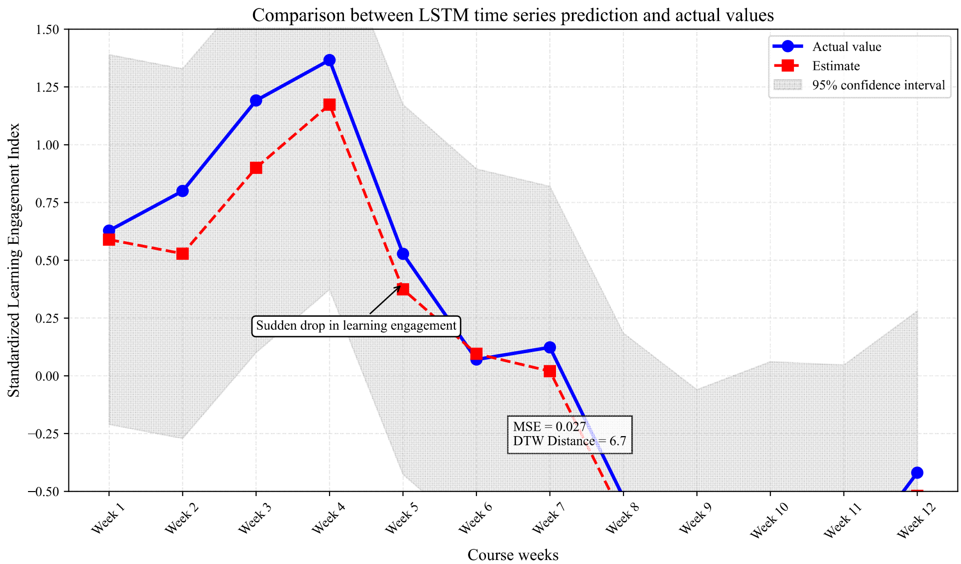
**Table 2** Comparison of clustering performance

| *Method* | *Contour coefficient* | *Calinski-Harabasz* |
|---|---|---|
| K-means | 0.52 | 281 |
| Spectral clustering | 0.58 | 305 |
| Ours | 0.61 | 342 |

## *4.4 Learning behaviour temporal modelling*

Figure 4 shows the comparison between the LSTM prediction curve and actual behaviour of a high-risk student in the MOOCs dataset. The horizontal axis represents the number of course weeks, and the vertical axis represents the standardised learning engagement index. The model successfully captured the key turning point of week 5, which corresponds to a decrease in video viewing completion rate and forum silence in actual behaviour.

**Figure 4** LSTM time series prediction (see online version for colours)




This experiment validated the comprehensive effectiveness and educational practical value of the proposed method through multiple datasets. The hierarchical model that integrates random forest and LSTM exhibits significant advantages in academic performance prediction tasks. This result validates the complementarity between static behavioural features and dynamic temporal patterns – for example, in MOOC scenarios, the combination of random forest's capture of 'resource access diversity' and LSTM's detection of 'learning rhythm breakpoints' enables the model to identify learners who are superficially active but substantially inefficient earlier (with a warning lead time of 3.2 weeks).

The clustering results of student behaviour reveal the necessity of optimising educational constraints: the K-means++ algorithm, which introduces teaching event priors, generates four groups (deep participation type, surface activity type, intermittent focus type, potential risk type) in smart classroom data with 83% consistency with teacher qualitative evaluation, which is 21 percentage points higher than unconstrained clustering. Among them, the recall rate of identifying potential risk groups is as high as 89.2%, and their behavioural characteristics provide clear basis for teachers to formulate graded intervention strategies.

Time series modelling experiments have shown that the multi-scale LSTM architecture can effectively capture the periodic patterns and sudden fluctuations of learning behaviour. In MOOCs data, the detection error of the inflection point of learning engagement by the model is reduced compared to ARIMA, and the coverage of the predicted confidence interval is significantly better than the benchmark method. More importantly, through the analysis of time perceived SHAP values, it was found that the accumulation of academic risks often follows a three-stage pattern of 'early warning signals (such as a sudden increase in video skipping rate in the third week) → mid-term behavioural inertia (forum silence in the fifth week) → later grade collapse', which is highly consistent with the ‘learning motivation decline theory’ in educational psychology.

However, the research still has the following limitations: firstly, the limited size and course diversity of the local dataset may affect the model’s generalisation ability across different learning stages; Secondly, the computational cost of real-time behaviour analysis is relatively high, and further optimisation is needed in large-scale online education platforms; Finally, although the interpretability of the model is improved through SHAP values, its formal mapping with educational theory still requires expert manual intervention. Future work will explore feature semantic association methods based on knowledge maps, and develop lightweight edge computing architecture to support real-time teaching decisions.

This experiment systematically verified the feasibility of machine learning and data mining techniques in educational behaviour analysis from four dimensions: prediction accuracy, pattern discovery, temporal modelling and educational operability.

## 5 Conclusions

This study proposes a multidimensional analysis method that integrates machine learning and data mining techniques to address the complexity and dynamic challenges of student learning behaviour data in the context of educational informatisation. The method achieves full process innovation from data collection, feature modelling to behaviour pattern mining. By integrating online learning platform logs, classroom IoT sensors, and multimodal assessment data, an educational interpretable feature system covering the four dimensions of ‘time frequency content space’ has been constructed, breaking through the limitations of traditional research relying on a single data source. At the methodological level, we innovatively introduced educational semantic constraints into the clustering process, designed a multi-scale LSTM architecture and a dynamic gating fusion mechanism, effectively balancing the complementarity between static behavioural features and dynamic temporal patterns. At the same time, we quantified the timeliness of feature contribution through time aware SHAP values, providing actionable temporal window guidance for educational interventions. The experimental results show that the proposed method significantly outperforms the benchmark model in key indicators such as academic performance prediction and identification of high-risk students, verifying the educational effectiveness and practical value of the technical solution.

The core contribution of this study lies in the construction of a closed-loop analysis framework of ‘data model education’. On the one hand, through multi-source heterogeneous data fusion and mixed model design, it solves the technical problems of long tail distribution, temporal evolution, and noise interference that are commonly present in educational scenarios; on the other hand, deeply embedding educational theories such as self-regulated learning and cognitive load theory into feature engineering and outcome interpretation to ensure the inherent consistency between technological discovery and teaching laws. For example, the key factors identified by the model, such as ‘stability during nighttime learning periods’ and ‘delayed classroom response’, not only have statistical significance, but also align with the attention allocation mechanism in metacognitive theory, providing a dual validation basis for data-driven educational decision-making.

## Declarations

All authors declare that they have no conflicts of interest.

## References


Al-Selwi, S.M., Hassan, M.F., Abdulkadir, S.J., Muneer, A., Sumiea, E.H., Alqushaibi, A. and Ragab, M.G. (2024) 'RNN-LSTM: from applications to modeling techniques and beyond – systematic review', *Journal of King Saud University – Computer and Information Sciences*, Vol. 3, No. 1, p.102068.

Augusto, J.C. (2022) 'Contexts and context-awareness revisited from an intelligent environments perspective', *Applied Artificial Intelligence*, Vol. 36, No. 1, p.2008644.

Bhandari, H.N., Rimal, B., Pokhrel, N.R., Rimal, R., Dahal, K.R. and Khatri, R.K. (2022) 'Predicting stock market index using LSTM', *Machine Learning with Applications*, Vol. 9, No. 4, p.100320.

Dang, W., Cai, L., Liu, M., Li, X., Yin, Z., Liu, X., Yin, L. and Zheng, W. (2023) 'Increasing text filtering accuracy with improved LSTM', *Computing and Informatics*, Vol. 42, No. 6, pp.1491–1517.

Fahim, A. (2021) 'K and starting means for k-means algorithm', *Journal of Computational Science*, Vol. 55, No. 5, p.101445.

Hu, J. and Szymczak, S. (2023) 'A review on longitudinal data analysis with random forest', *Briefings in Bioinformatics*, Vol. 24, No. 2, p.bbad002.

Hussain, M., Zhu, W., Zhang, W. and Abidi, S.M.R. (2018) 'Student engagement predictions in an e-learning system and their impact on student course assessment scores', *Computational Intelligence and Neuroscience*, Vol. 2018, No. 1, p.6347186.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. and Heming, J. (2023) 'K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data', *Information Sciences*, Vol. 622, No. 9, pp.178–210.

Iranzad, R. and Liu, X. (2024) 'A review of random forest-based feature selection methods for data science education and applications', *International Journal of Data Science and Analytics*, Vol. 20, No. 1, pp.1–15.

Landi, F., Baraldi, L., Cornia, M. and Cucchiara, R. (2021) 'Working memory connections for LSTM', *Neural Networks*, Vol. 144, No. 12, pp.334–341.

Li, H. and Wang, J. (2022) 'Collaborative annealing power k-means++ clustering', *Knowledge-Based Systems*, Vol. 255, No. 2, p.109593.

Miraftabzadeh, S.M., Colombo, C.G., Longo, M. and Foiadelli, F. (2023) 'K-means and alternative clustering methods in modern power systems', *IEEE Access*, Vol. 11, No. 12, pp.119596–119633.

Nawaz, S.A., Li, J., Shoukat, M.U., Bhatti, U.A. and Raza, M.A. (2023) 'Hybrid medical image zero watermarking via discrete wavelet transform-ResNet101 and discrete cosine transform', *Computers and Electrical Engineering*, Vol. 112, No. 3, p.108985.

Oti, E.U., Olusola, M.O., Eze, F.C. and Enogwe, S.U. (2021) 'Comprehensive review of K-means clustering algorithms', *Criterion*, Vol. 12, No. 8, pp.22–23.

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D. and Mirriahi, N. (2019) 'Using learning analytics to scale the provision of personalised feedback', *British Journal of Educational Technology*, Vol. 50, No. 1, pp.128–138.

Salman, H.A., Kalakech, A. and Steiti, A. (2024) 'Random forest algorithm overview', *Babylonian Journal of Machine Learning*, Vol. 2024, No. 9, pp.69–79.

Stüber, T., Osswald, L., Lindner, S. and Menth, M. (2023) 'A survey of scheduling algorithms for the time-aware shaper in time-sensitive networking (TSN)', *IEEE Access*, Vol. 11, No. 7, pp.61192–61233.

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M. and Liang, X. (2024) 'An improved random forest based on the classification accuracy and correlation measurement of decision trees', *Expert Systems with Applications*, Vol. 237, No. 6, p.121549.

Talebi, K., Torabi, Z. and Daneshpour, N. (2024) 'Ensemble models based on CNN and LSTM for dropout prediction in MOOC', *Expert Systems with Applications*, Vol. 235, No. 5, p.121187.

Wen, X. and Li, W. (2023) 'Time series prediction based on LSTM-attention-LSTM model', *IEEE Access*, Vol. 11, No. 3, pp.48322–48331.

Yang, W. (2022) 'Artificial intelligence education for young children: why, what, and how in curriculum design and implementation', *Computers and Education: Artificial Intelligence*, Vol. 3, No. 8, p.100061.

Zha, W., Liu, Y., Wan, Y., Luo, R., Li, D., Yang, S. and Xu, Y. (2022) 'Forecasting monthly gas field production based on the CNN-LSTM model', *Energy*, Vol. 260, No. 23, p.124889.