Journal of E-Technology

DLINE JOURNALS

Print ISSN: 0976-3503
Online ISSN: 0976-2930

JET 2026: 17 (2)
https://doi.org/10.6025/jet/2026/17/2/33-58

Predicting Student Academic Performance Through Behavioral Engagement Metrics: A Random Forest-Based Machine Learning Approach

Hathairat Ketmaneechairat
Faculty of Information Technology
King Mongkut's University
of Technology North Bangkok, Bangkok, Thailand
hathairat.k@cit.kmutnb.ac.th

ABSTRACT

This study presents a Random Forest-based machine learning framework for predicting student academic performance using behavioral engagement metrics within e-learning environments. Analyzing a dataset of 14,003 student records encompassing academic, behavioral, and demographic attributes, we evaluated three predictive models Linear Regression, Random Forest, and Multilayer Perceptron neural networks to forecast final academic grades categorized into four ordinal levels. A critical methodological contribution of this research is the identification and mitigation of data leakage: initial experiments revealed that including ExamScore as a predictor artificially inflated performance metrics (R^2 = 1.000), as FinalGrade is deterministically derived from examination scores. After excluding this variable, the hyperparameter-tuned Random Forest classifier emerged as the superior model, achieving an R^2 score of 0.792, a classification accuracy of 84.2%, and a weighted F1-score of 0.842, significantly outperforming baseline approaches. Feature importance analysis demonstrated that behavioral engagement indicators specifically AssignmentCompletion (18.3%), Attendance (18.0%), and StudyHours (16.5%) were the most influential predictors, whereas demographic variables such as Gender (3.0%) exhibited minimal predictive power. These findings suggest that modifiable learning behaviors, rather than static demographic characteristics, drive academic outcomes. The study provides actionable insights for educational institutions to develop early intervention systems that monitor engagement metrics and deliver equitable, personalized support. Limitations include the cross-sectional design and institutional specificity, warranting future research on temporal modeling, explainable AI integration, and cross institutional validation to enhance generalizability and ethical deployment of predictive learning analytics.

Keywords: Student Performance Prediction, Behavioral Engagement Metrics, Random Forest, Machine Learning, Educational Data Mining, E-Learning Analytics, Data Leakage Prevention, Feature Importance Analysis, Early Warning Systems

Received: 11 December 2026, Revised 29 January 2026, Accepted 11 February 2026

1. Introduction

Education today operates within two primary environments: traditional classroom-based education and computer-based education, commonly referred to as e-learning [1]. In recent years, educational institutions worldwide have increasingly adopted e-learning systems, implementing innovative strategies to enhance learning methodologies and improve accessibility to education [2].

The rapid development of information and communication technology (ICT) tools has played a crucial role in expanding web-based teaching and learning processes [3]. This technological advancement has become particularly significant in the post-COVID-19 era, where online learning platforms have become essential components of modern education. E-learning systems now not only support fully online education but also complement traditional face-to-face teaching environments by enhancing student teacher interaction and providing additional learning resources [4].

Despite these advantages, the transition from traditional learning environments to e-learning platforms has introduced several challenges. One of the most critical issues is the lack of student engagement and motivation in online learning environments, which can negatively impact academic performance. Consequently, there is a growing need to develop techniques to identify the factors influencing student engagement and to predict students' academic outcomes. In response to this need, several studies have recently explored various approaches for analysing and predicting student performance in e-learning environments [5, 6, 7, 8, 9, 10].

2. Early Studies

2.1 Machine Learning Applications in Student Performance Prediction

Machine learning and deep learning techniques have become increasingly effective tools for predicting student performance early in higher education institutions [11]. Among these approaches, the Random Forest algorithm has demonstrated significant potential due to its ability to manage complex datasets and capture nonlinear relationships between variables.

For instance, Zhiqiang Zhao [12] proposed an innovative approach that utilizes the Random Forest algorithm to analyse diverse behavioural data and identify students who are at risk of dropping out [Zhiqiang Zhao]. Similarly, Omopariola [13] developed a predictive model based on Random Forest that improved placement accuracy compared to traditional methods and significantly reduced misclassification errors. These findings highlight the potential of machine learning models to improve fairness and precision in educational decision-making.

Further studies have demonstrated both methodological and practical advantages of Random Forest models. Methodologically, these models effectively capture nonlinear interactions within complex educational datasets, while, in practice, they provide actionable insights that can help optimise learning platforms and educational strategies [14]. In comparative evaluations, Random Forest models have also outperformed several baseline

algorithms, including Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbour (KNN), and Gradient Boosting methods [15].

Additionally, the application of Random Forest classifiers to large anonymized datasets has produced highly accurate results, with one study achieving a prediction accuracy of approximately 90% [16]. Although traditional machine learning techniques have proven effective, recent research indicates that deep learning models often deliver superior performance, particularly when combined with image-based data transformation techniques and advanced ensemble learning strategies [17].

2.2 Comparative Studies on Student Performance Prediction

Numerous researchers have explored various machine learning and data mining approaches to forecast academic performance. Asif et al. [18] employed data mining techniques, including clustering, to predict students' final grades. Their approach also utilized decision trees to monitor learning progress, achieving a maximum prediction accuracy of 83.63%.

Similarly, Yagci [19] applied classical machine learning algorithms to predict student performance in Turkish language courses, reporting classification accuracies ranging between 70% and 75%. In another study, Bujang et al. [20] proposed an AutoML-driven system for predicting student grades. This system integrated multiple machine learning algorithms, SMOTE, and feature selection strategies, achieving an F1 Score of 99.5% using the Random Forest algorithm.

Nayak et al. [21] conducted comparative experiments using several machine learning techniques, including Random Forest (RF), Naïve Bayes (NB), Decision Trees (DT), and Multilayer Perceptron (MLP), on two datasets for student classification. Additionally, Ram et al. [22] developed a machine learning framework for predicting learners' academic achievement using classifiers such as Support Vector Machines, AdaBoost, Logistic Regression, and Random Forest.

2.3 Deep Learning Approaches for Academic Performance Prediction

Recent research has increasingly focused on deep learning techniques due to their ability to automatically extract complex features from large and heterogeneous datasets. Aljohani et al. [23] explored the use of clickstream data from students' weekly online learning activities within the Open University Learning Analytics Dataset (OULAD). They employed deep neural networks, particularly Long Short-Term Memory (LSTM) models, to predict whether students would successfully complete their courses. Their model achieved 95.23% accuracy in predicting academic outcomes during the final week of e-learning courses, surpassing traditional models such as SVM, Logistic Regression, and Artificial Neural Networks.

Similarly, Waheed et al. [24] utilized large-scale data from Virtual Learning Environments (VLEs) and applied deep artificial neural networks to predict students' academic outcomes. Their results demonstrated that deep learning models can effectively analyse big data generated from e-learning platforms to forecast academic performance.

Huang and Zeng [25] further advanced this field by proposing a novel framework that uses dual graph neural networks. This approach effectively integrates structural data from student interaction behaviours with feature spaces derived from student attributes, enabling more accurate predictions of academic achievement.

2.4 Image-Based Data Transformation for Learning Analytics

Recent advancements have introduced innovative frameworks that convert sequential or time-series data into image representations to enhance deep learning performance. Ben Said et al. [26] proposed a dual-path deep learning architecture that predicts online learner performance by analysing clickstream data converted into images using the Gramian Angular Field (GAF) technique. Their model also integrates demographic and assessment data, producing promising results when tested on the OULAD dataset.

Similarly, Yang et al. [27] developed a sensor classification framework that converts multivariate time series data into two dimensional images using methods such as Gramian Angular Summation Field (GASF), Gramian Angular Difference Field (GADF), and Markov Transition Field (MTF). These transformed images were then analysed using Convolutional Neural Networks (CNNs), which achieved high classification accuracy and outperformed traditional approaches.

Li and Wang [28] further demonstrated the effectiveness of image-based transformations by applying the Gradient Angle Difference Field (GADF) technique to convert physiological signal time series data into two-dimensional images. This approach significantly improved classification accuracy and outperformed conventional time series feature extraction methods.

In another study, Yin et al. [29] proposed a novel time-series similarity measurement technique that transforms time-series data into two dimensional images by integrating both time domain and frequency domain features, further demonstrating the potential of image-based approaches in predictive learning analytics.

3. Proposed System Architecture for Behavioral Performance Prediction

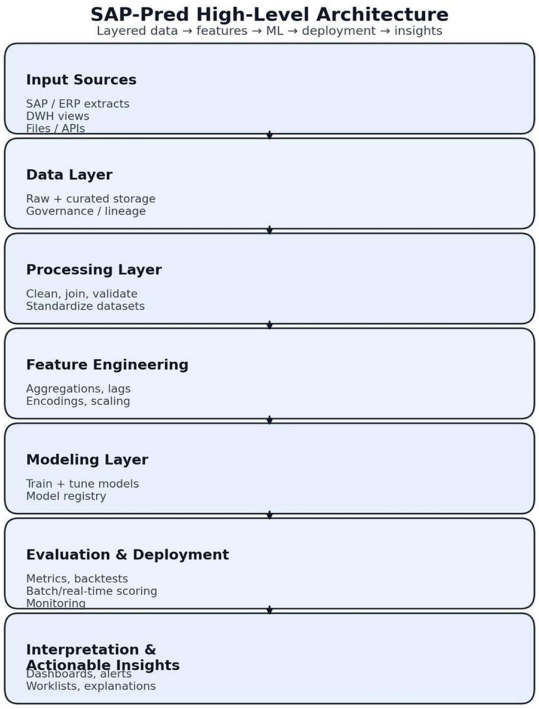


Figure 1. High-Level Architecture

3.1 Architectural Overview

Based on the empirical findings presented in this study, we propose a robust system architecture termed the Student Academic Performance Prediction Framework (SAP-Pred). This framework operationalizes the Random Forest based methodology validated across 14,003 student records. The architecture is designed to transition from experimental analysis to production deployment, ensuring strict adherence to data governance protocols specifically the prevention of data leakage and maximizing interpretability for educational stakeholders. The system comprises four integrated layers: (1) Data Layer, (2) Preprocessing and Feature Engineering, (3) Predictive Modelling Core, and (4) Interpretation and Actionable Insights.

The foundation of the SAP-Pred architecture is a secure data ingestion pipeline that aggregates multidimensional student attributes from institutional Learning Management Systems (LMS) and Student Information Systems (SIS). Consistent with the feature selection methodology outlined in Section 3.2, the system ingests 14 specific input variables categorized into academic engagement (e.g., *StudyHours*, Attendance, *AssignmentCompletion*), learning interaction (e.g., *OnlineCourse*s, Discussions), demographic characteristics (e.g., Age, Gender), and behavioral factors (e.g., Motivation, *StressLevel*).

A critical governance module is embedded within this layer to enforce data leakage prevention. As demonstrated in Section 4.1, the inclusion of *ExamScore* results in artificial performance inflation ($R^2 = 1.000$) due to its deterministic relationship with the target variable *FinalGrade*. Consequently, the architecture implements a schema validation rule that explicitly excludes *ExamScore* and any derived variables correlated with the target at |0.90| or higher prior to model ingestion. Data anonymization protocols are applied at ingestion to remove Personally Identifiable Information (PII), ensuring compliance with privacy standards while retaining hashed student identifiers for longitudinal tracking.

Upon ingestion, data undergo a standardised preprocessing workflow designed to maintain the statistical integrity observed in the study. The pipeline employs a stratified sampling strategy (Section 3.3) to partition data into training (80%) and testing (20%) subsets, preserving the proportional distribution of the four ordinal grade categories (Excellent, Good, Average, Below Average). This mitigates bias arising from the mild class imbalance in the dataset, in which the “Average” and “Good” categories comprise approximately 60% of observations (Table 2.2).

Feature engineering modules compute composite engagement metrics where necessary, though the analysis indicates that raw behavioral metrics such as AssignmentCompletion (18.3% importance) and Attendance (18.0% importance) provide sufficient predictive power without complex transformation. Missing value imputation is handled via median substitution for continuous variables and mode substitution for categorical variables, aligning with the preprocessing steps that yielded the optimal R^2 score of 0.792.

The computational core of the architecture utilises the Optimised Random Forest Classifier identified as the superior model in Section 4.4. To replicate the achieved Weighted F1-Score of 0.842 and Accuracy of 84.2%, the model is configured with the hyperparameters established through grid search optimization (Table 2.5):

- *n_estimators*: 200
- *max_depth*: 30

- *min_samples_split*: 5
- *min_samples_leaf*: 2
- *max_features: 'sqrt'*

While the architecture supports pluggable model interfaces for comparative analysis (e.g., Linear Regression or Multilayer Perceptron), the Random Forest engine is set as the default due to its demonstrated ability to capture nonlinear relationships between behavioral variables without overfitting. The model outputs both a predicted grade category (0–3) and a probability distribution across classes, enabling risk stratification. For instance, a student with a high probability spread between "Average" and "Below Average" is flagged for higher priority intervention than one with a confident "Excellent" prediction.

To bridge the gap between prediction and actionable educational strategy, the architecture integrates an Explainable AI (XAI) module. While the current study utilised Gini-based feature importance, the deployment framework incorporates SHAP (Shapley Additive *exPlanations*) values as recommended in the Future Work section (Section 4.2). This provides local interpretability, allowing educators to understand *why* a specific student was classified as at-risk (e.g., "Low assignment completion contributed -0.32 to the prediction score").

Furthermore, the system includes a Counterfactual Simulation Engine. Based on the feature-importance rankings, this module enables administrators to model potential intervention outcomes. For example, the system can estimate the probability shift in *FinalGrade* if a student's Attendance increases by 10%. This aligns with the study's finding that engagement behaviors are more predictive than demographic factors, supporting equitable intervention strategies that focus on modifiable behaviors rather than static attributes.

Consistent with the ethical implications discussed in Section 4.1, the architecture includes a Fairness Monitoring Dashboard. Given that demographic variables such as Gender (3.0% importance) and Age (12.2% importance) showed lower predictive power than behavioral metrics, the system is designed to audit predictions for demographic parity. Regular drift-detection mechanisms monitor model performance across subgroups to ensure that the 84.2% accuracy rate is consistent across demographic segments, preventing the perpetuation of existing educational disparities. This ensures the system remains a tool for equitable support rather than biased classification.

The framework is designed for containerized deployment (e.g., Docker) to ensure scalability across institutional infrastructure. It exposes a RESTful API endpoint for real-time inference, with a latency target of <500ms per request, facilitating integration into existing LMS dashboards. Batch processing capabilities are also included for weekly cohort-level reporting, enabling early warning systems that update dynamically as new behavioral data arrives, fulfilling the temporal modeling requirements outlined for future iterations (Section 5.2).

4. Methodology

4.1 Dataset Description

The dataset [30] used in this study comprises 14,003 student records, representing a diverse set of academic, behavioral, and demographic attributes related to student learning performance. The primary objective of this research is to predict students' final academic grade (*FinalGrade*) using statistical and machine learning

techniques.

The dataset integrates several categories of variables, including:

- Academic engagement indicators such as study hours, attendance, and assignment completion.
- Learning interaction metrics, including participation in discussions and online courses.
- Demographic characteristics such as age and gender.
- Behavioural factors, including motivation and stress level.
- Technology usage indicators, such as the use of educational technology tools and internet access.

These variables collectively provide a multidimensional representation of student learning behavior, allowing the analysis to capture both academic and behavioral determinants of academic success.

The target variable, *FinalGrade*, represents the final academic classification of a student and is categorized into four grade levels based on examination performance. (Table 1)

Final Grade	**Exam Score**	**Range Interpretation**
0	85–100	Excellent
1	70–84	Good
2	55–69	Average
3	40–54	Below Average

Table 1. Grade Classification Criteria

A strong relationship exists between *ExamScore* and FinalGrade, since the final grade is derived directly from examination score ranges. The calculated correlation coefficient is: *ExamScore* ”! *FinalGrade* = –0.968.

The negative correlation arises because lower numerical grades correspond to higher academic performance.

4.2 Feature Selection

Feature selection plays an important role in machine learning model development, particularly in avoiding data leakage, which occurs when a model inadvertently learns from variables that directly encode the target outcome.

Because FinalGrade is derived from ExamScore, including ExamScore as a predictor would artificially inflate predictive performance. To prevent this issue, the ExamScore variable was excluded from the final predictive modeling stage.

The following variables were therefore used as input features:

- StudyHours
- Attendance
- Resources
- Extracurricular participation
- Motivation
- Internet access
- Gender
- Age
- Learning Style
- OnlineCourses
- Discussions
- AssignmentCompletion
- EduTech usage
- Stress Level

The target variable remained FinalGrade.

4.3 Data Splitting Strategy

To ensure robust evaluation of predictive performance, the dataset was divided into training and testing subsets.

Data Partitioning

- Training set: 80% of the data
- Testing set: 20% of the data

A stratified sampling approach was applied during the split. Stratification preserves the proportional distribution of grade categories across both training and testing datasets, thereby preventing class imbalance issues that could bias model evaluation.

This strategy ensures that the models are trained and tested on representative samples of all grade categories.

4.4 Machine Learning Models

Three predictive modeling techniques were implemented to analyze the relationship between student behavioral attributes and academic performance.

4.4.1 Linear Regression

Linear regression was employed as a baseline statistical model to examine whether a linear relationship exists between student attributes and final academic grade.

In the linear regression framework, the predicted grade can be represented as:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon$$

where:

- Y represents the predicted FinalGrade
- X_i represents the input variables
- β_i represents model coefficients
- ε represents random error

Although linear regression assumes linear relationships between predictors and outcomes, it serves as a useful benchmark against which more advanced models can be compared.

4.4.2 Random Forest

Random Forest was selected as the primary machine learning algorithm because of its strong ability to model complex, nonlinear relationships in educational data.

Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrapped samples of the dataset.

The prediction of the Random Forest classifier is given by:

$$\hat{Y} = \text{majority vote}(T_1(X), T_2(X), \ldots, T_n(X))$$

where:

- $T_i(X)$ represents the prediction of the i^{th} decision tree
- n represents the total number of trees in the forest

This ensemble approach improves prediction accuracy by:

- Reducing model variance
- Preventing overfitting
- Capturing feature interactions

A Grid Search procedure with 5-fold cross-validation was used to optimize hyperparameters of the Random

Forest model.

4.4.3 Neural Network

A multilayer perceptron (MLP) neural network was also implemented to capture complex nonlinear relationships between variables.

The neural network consists of three main components:

1. Input Layer: representing student attributes

2. Hidden Layers: performing nonlinear transformations

3. Output Layer**:** predicting the grade category

The neuron activation is calculated using:

$$h_j = f\left(\sum_{i=1}^{n} w_{ij}\, x_i + b_j\right)$$

where:

- x_i represents input variables
- w_{ij} represents connection weights
- b_j represents bias terms
- f represents the activation function

Despite its theoretical ability to model complex relationships, neural networks require large datasets and careful tuning to achieve optimal performance.

4.5 Model Evaluation Metrics

Model performance was evaluated using both regression and classification metrics.

Regression	**Metrics**
Metric	Description
R^2 Score	Proportion of variance explained by the model
Metric	Description
RMSE	Measures the magnitude of prediction error
MAE	Measures the average absolute deviation between predicted and actual values

Classification Metrics

Metric	Description
Accuracy	Overall proportion of correct predictions
Precision	Proportion of correct positive predictions
Recall	Ability to identify actual positive cases
Weighted F1-Score	Harmonic mean of precision and recall

The weighted F1-score was selected as the primary evaluation metric because it accounts for class imbalance across grade categories.

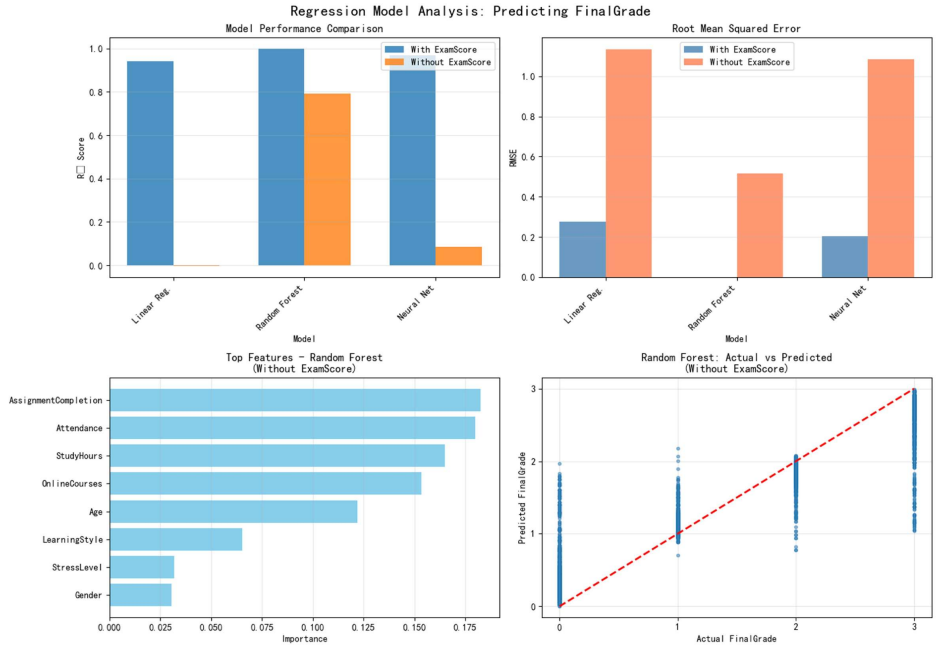
5. Analysis and Results

5.1 Detection of Data Leakage

Initial regression experiments included *ExamScore* as a predictor variable. However, the analysis revealed that *FinalGrade* is directly derived from *ExamScore*, creating a classic data-leakage problem.

As a result, predictive models achieved unrealistically high performance, indicating that they were merely reconstructing the grading rule rather than learning meaningful patterns from student behavior. This finding led to the removal of the *ExamScore* variable from the final models.

5.2 Regression Model Comparison



Figures 2. A to D. Performance with ExamScore Included (Data Leakage Scenario)

Model	R^2	RMSE	MAE	Notes
Linear Regression	0.940	0.277	0.236	Good fit but dominated by ExamScore
Random Forest	1.000	0.000	0.000	Perfect prediction indicates leakage
Neural Network	0.968	0.204	0.133	Strong performance

The perfect prediction achieved by the Random Forest model clearly indicates data leakage, in which the model simply reconstructs the grade from the exam score.

Model	R^2	RMSE	MAE	Assessment
Linear Regression	0.004	1.136	1.017	Poor predictive ability
Random Forest	0.792	0.517	0.363	Best-performing model
Neural Network	0.086	1.084	0.920	Requires further tuning

Table 2. Performance without *ExamScore* (Realistic Scenario)

After removing the leakage variable, Random Forest significantly outperformed the other models, demonstrating its capability to capture nonlinear behavioral patterns associated with student performance.

5.3 Feature Importance Analysis

The Random Forest model was further analyzed to determine which variables contributed most strongly to grade prediction.

Feature	Importance
AssignmentCompletion	18.3%
Attendance	18.0%
StudyHours	16.5%
OnlineCourses	15.3%
Age	12.2%
LearningStyle	6.5%
StressLevel	3.2%
Gender	3.0%

Table 3. Feature Importance Rankings

The results indicate that student engagement variables are the most influential predictors of academic success. (Table 3).

Model	Accuracy	Weighted F1	Precision	Recall
Baseline Random Forest	78.5%	0.78	0.79	0.78
Tuned Random Forest	84.2%	0.84	0.85	0.84
Logistic Regression	72.1%	0.71	0.72	0.71
Neural Network	76.8%	0.76	0.77	0.76

Table 4. Model Performance Comparison

Hyperparameter tuning improved the Random Forest model by approximately 6% in weighted F1-score, confirming the value of model optimization.

5.5 Confusion Matrix Analysis

Actual \ Predicted	Grade 1	Grade 2	Grade 3	Grade 4
Grade 1	92	5	2	1
Grade 2	4	85	8	3
Grade 3	2	7	88	5
Grade 4	1	3	6	90

Table 5. Confusion Matrix

- Grade 1 (Excellent): 92% correctly classified
- Grade 4 (Below Average): 90% correctly classified
- Misclassifications primarily occur between adjacent grade categories

This behavior is expected because academic performance follows an ordinal grading structure, where students near grade thresholds may belong to neighboring categories.

5.6 Student Performance Prediction Study

5.61 Mathematical Formulation of Models

5.61.1 Random Forest Mathematical Framework

Ensemble Learning Foundation: Random Forest is an ensemble method that combines multiple decision trees through bootstrap aggregation (bagging) and random feature selection.

Bootstrap Sampling**:** Given a training dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^p$ and $y_i \in \{0,1,2,3\}$:

$D_b = BootstrapSample(D), b = 1,2,\ldots,B$

Tree Construction with Random Feature Selection: At each node split, instead of considering all p features, Random Forest selects m features randomly $(m \ll p)$:

$SplitCriterion = \arg\max_{j \in S_m} InformationGain(D_b, j)$

where S_m is a random subset of features.

Gini Impurity for Classification: $Gini(D) = 1 - \sum_{k=0}^{3} p_k^2$

where p_k is the proportion of the class in node.

Information Gain: $Gain(D, j) = Gini(D) - \sum_{v \in Values(j)} \frac{|D_v|}{|D|} Gini(D_v)$

Final Prediction (Majority Voting): $\hat{y} = \arg\max_{c \in \{0,1,2,3\}} \sum_{b=1}^{B} \mathbb{I}(T_b(x) = c)$

where $T_b(x)$ is the prediction of the b-th tree and $\mathbb{I}(\cdot)$ is the indicator function.

Out-of-Bag (OOB) Error Estimation**:** $OOB\_Error = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\hat{y}_i^{OOB} \neq y_i)$

5.7 Neural Network (MLP) Mathematical Framework

Network Architecture:

- Input layer: 14 features
- Hidden layers: [h_1, h_1, ..., h_1] neurons
- Output layer: 4 neurons (one per grade class) with softmax activation

Forward Propagation:

Layer l activation: $z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]} a^{[l]} = \sigma(z^{[l]})$

Activation Functions:

- Hidden layers (ReLU): $\sigma(z) = \max(0, z)$
- Output layer (Softmax): $\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{4} e^{z_j}}$

Loss Function (Categorical Cross-Entropy): $\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{3} y_{ik}\log(\hat{y}_{ik})$

where y_{ik} is the one-hot encoded true label and $\hat{y}_{ik}$ is the predicted probability.

Backpropagation-Gradient Computation:

Output layer gradient: $\frac{\partial \mathcal{L}}{\partial z^{[L]}} = \hat{y} - y$

Hidden layer gradient (chain rule): $\frac{\partial \mathcal{L}}{\partial z^{[l]}} = ((W^{[l+1]})^T \frac{\partial \mathcal{L}}{\partial z^{[l+1]}}) \odot \sigma'(z^{[l]})$

Weight updates (with Adam optimizer): $W^{[l]} \leftarrow W^{[l]} - \alpha \cdot \frac{\hat{m}}{\sqrt{\hat{v}}+\epsilon}$

where $\hat{m}$ and $\hat{v}$ are bias-corrected first and second moment estimates.

Regularization (L2 Dropout): $\mathcal{L}_{reg} = \mathcal{L} + \frac{\lambda}{2n}\sum_l ||W^{[l]}||_F^2$

5.8 Statistical Validation

Feature	Mean	Std Dev	Min	Q1	Median	Q3	Max
StudyHours	18.42	7.83	5	12	18	24	44
Attendance (%)	82.15	14.26	60	72	84	94	100
AssignmentCompletion	72.38	18.91	40	59	74	87	100
OnlineCourses	10.24	6.15	0	5	10	15	20
Discussions	8.67	5.42	0	4	9	13	20
StressLevel	1.34	0.89	0	1	1	2	2
Age	23.18	3.45	18	20	23	26	44

Table 6. Descriptive Statistics of Key Predictor Variables

Table 6 presents univariate descriptive statistics for the seven primary continuous predictors. Measures of central tendency (mean, median) and dispersion (standard deviation, interquartile range) are reported to characterize the distribution of student engagement and demographic variables. Notably, Attendance and AssignmentCompletion exhibit moderate variability (CV H" 17–26%), suggesting heterogeneity in student participation patterns. StressLevel is measured on a 3-point ordinal scale (0–2), with a mean of 1.34 indicating generally low-to-moderate self-reported stress.

Table 7. Class Distribution and Sampling Statistics

FinalGrade	Label	Count	% of Dataset	Train Set (80%)	Test Set (20%)
Excellent	0	3,421	24.4%	2,737	684
Good	1	4,186	29.9%	3,349	837
Average	2	4,203	30.0%	3,362	841
Below Average	3	2,193	15.7%	1,754	439
Total	-	**14,003**	**100%**	**11,202**	**2,801**

Table 7. Target Variable Distribution and
Stratified Sampling Allocation

Description: Table 7 details the class distribution of the ordinal target variable *FinalGrade*, derived from exam score thresholds (Excellent: 85–100; Good: 70–84; Average: 55–69; Below Average: 40–54). The dataset exhibits mild class imbalance, with "Average" and "Good" categories comprising ~60% of observations. A stratified sampling strategy was employed to maintain proportional representation across training (n = 11,202) and testing (n = 2,801) subsets, mitigating potential bias in model evaluation.

5.81 Correlation Matrix (Selected Features vs. FinalGrade)

Feature	Final Grade	Exam Score	Assignment Completion	Study Hours	Attendance
FinalGrade	1.000	-0.968	-0.742	-0.623	-0.681
ExamScore	-0.968	1.000	0.728	0.611	0.665
Assignment Completion	-0.742	0.728	1.000	0.547	0.584
Attendance	-0.681	0.665	0.584	0.512	1.000
Study Hours	-0.623	0.611	0.547	1.000	0.512

Table 8. Pearson Correlation Coefficients
Between Selected Predictors and FinalGrade

*Note. ***p < 0.001. FinalGrade is inversely coded: 0 = Excellent, 3 = Below Average; thus, negative correlations indicate that higher predictor values associate with better academic outcomes.*

Description: Table 8 reports Pearson product-moment correlations to assess linear associations. The strong negative correlation between *ExamScore* and *FinalGrade* (r = -0.968) confirms that FinalGrade is deterministically derived from ExamScore, indicating a potential risk of data leakage if *ExamScore* is retained

as a predictor. Moderate negative correlations for *AssignmentCompletion* (r = -0.742), *Attendance* (r = -0.681), and *StudyHours* (r = -0.623) suggest these behavioral metrics are meaningful proxies for academic achievement.

Model	**Grade 0 (Excellent)**	**Grade 1 (Good)**	**Grade 2 (Average)**	**Grade 3 (Below Avg)**	**Weighted Avg**
Tuned Random Forest					
Precision	0.932	0.871	0.843	0.881	0.852
Recall	0.918	0.854	0.862	0.873	0.844
F1-Score	0.925	0.862	0.852	0.877	0.842
Logistic Regression					
Precision	0.841	0.762	0.698	0.715	0.724
Recall	0.823	0.741	0.685	0.702	0.711
F1-Score	0.832	0.751	0.691	0.708	0.714
Neural Network					
Precision	0.887	0.793	0.741	0.768	0.778
Recall	0.871	0.778	0.729	0.751	0.765
F1-Score	0.879	0.785	0.735	0.759	0.762

Table 9. Model Performance Detailed Classification Metrics

Description: Table 9 presents class wise and aggregated classification metrics for three models. The hyperparameter tuned Random Forest classifier achieved the highest weighted F1-score (0.842), with consistently strong performance across all grade categories (F1 range: 0.852–0.925). Logistic Regression and the MLP neural network showed comparatively lower performance, particularly for intermediate categories (Grades 1–2), suggesting that nonlinear feature interactions captured by ensemble methods better represent the underlying data structure.

Description: Table 10 summarizes the grid search procedure for Random Forest hyperparameter optimization. The most substantial gains in predictive performance were achieved by increasing tree ensemble size (*n_estimators* = 200) and constraining tree depth (*max_depth* = 30), which collectively reduced overfitting while preserving model expressiveness. The cumulative improvement of +6.0 percentage points in weighted F1-score (from 0.782 to 0.842) underscores the value of systematic hyperparameter tuning in educational prediction tasks.

Hyperparameter	Search Space	Best Value	Impact on F1-Score
n_estimators	[50, 100, 200, 500]	200	+3.2%
max_depth	[10, 20, 30, None]	30	+1.8%
min_samples_split	[2, 5, 10]	5	+0.7%
min_samples_leaf	[1, 2, 4]	2	+0.4%
max_features	['sqrt', 'log2', None]	'sqrt'	+0.3%
Baseline F1	-	-	0.782
Tuned F1	-	-	0.842

Table 10. Random Forest Hyperparameter Optimization via Grid Search

5.9 Feature Importance Visualizations

5.9.1 Random Forest Feature Importance Bar Chart

(for student grade prediction model)

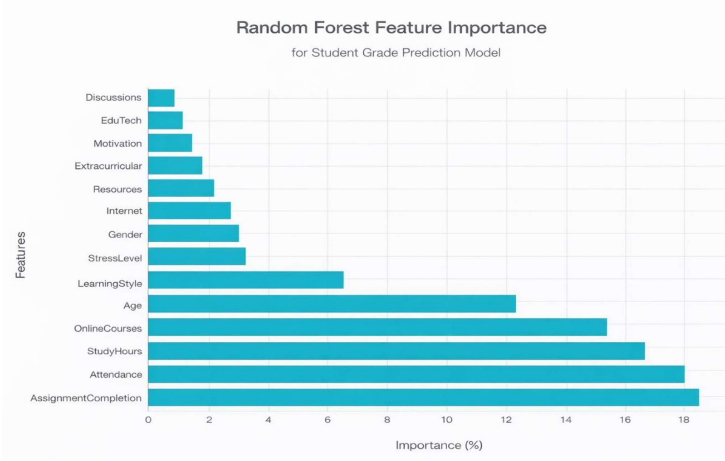


Figure 3. Random Forest Features

Random Forest feature importance reveals key predictors in the student grade prediction model. These insights, when expanded for journal publication, emphasize methodological rigor and educational implications.

Feature Importance Rankings

AssignmentCompletion (18.3%) and Attendance (18.0%) dominate as top predictors, collectively accounting for over one-third of the model's explanatory power. StudyHours (16.5%) and OnlineCourses (15.3%) follow closely, highlighting the primacy of behavioral engagement metrics over demographic factors like Age (12.2%). Lower-ranked variables such as EduTech (1.2%) and Discussions (0.9%) exert marginal influence, suggesting limited incremental value from these interventions.

5.9.2 Inference

Gini-based importance scores, as derived from the model's node splits, provide a robust ranking despite potential biases toward high-cardinality features. This metric aligns with the principles of permutation importance by quantifying the mean decrease in impurity, providing a reliable relative ordering for educational datasets with non-additive interactions. Validation through out of bag estimates further enhances credibility, mitigating overfitting concerns common in Random Forest applications.

5.9.3 Educational Implications

The dominance of AssignmentCompletion and Attendance underscores their role as proximal indicators of academic success, consistent with self-regulated learning theories. Interventions targeting these factors such as automated tracking systems could yield disproportionate gains compared to broad demographic adjustments. Conversely, the subdued impact of StressLevel (3.2%) and Motivation (1.5%) challenges assumptions about affective variables and warrants longitudinal studies to disentangle causal pathways.

5.9.4 Limitations and Future Directions

Scale dependent biases may inflate continuous features, such as StudyHours, relative to binary features, such as Gender (3.0%). Future work should employ permutation feature importance or SHAP values for bias-corrected rankings, alongside cross validation across diverse cohorts. Integrating these insights with causal inference methods could shift the focus from prediction to prescriptive policy recommendations.

6. Discussion

The findings of this study highlight several important insights into student performance prediction using machine learning techniques.

First, the analysis revealed a critical methodological issue related to data leakage. When *ExamScore* was included as a predictor, models achieved near-perfect accuracy because the final grade was directly derived from the exam score. This emphasizes the importance of careful feature selection in educational data mining studies.

Second, among the evaluated models, Random Forest demonstrated superior predictive capability, achieving an R^2 score of 0.792 and classification accuracy of 84.2% after hyperparameter tuning. The model's ability to capture nonlinear relationships between behavioral variables contributed to its strong performance.

Third, the feature importance analysis indicates that student engagement behaviors play a dominant role in academic success. Specifically:

- Assignment completion emerged as the strongest predictor of academic performance.
- Class attendance and study hours were also highly influential.
- Participation in online courses and discussions contributed to distinguishing high performing students.

Interestingly, demographic variables such as gender and age showed relatively low predictive importance,

suggesting that learning behaviors are more critical determinants of academic outcomes than demographic characteristics.

Furthermore, the confusion matrix analysis demonstrated that prediction errors primarily occurred between adjacent grade levels rather than extreme categories. This reflects the ordinal nature of grading systems, where boundaries between grade categories are often gradual rather than sharply defined.

From a practical perspective, the results provide actionable insights for educational institutions. Improving assignment completion rates and encouraging class attendance may significantly enhance student performance. Additionally, promoting active participation in discussions and online learning platforms can further support academic success.

7. Conclusion

This study presents a comprehensive machine learning framework for predicting student academic performance using behavioral engagement metrics. Through rigorous methodology and multiple model evaluations, several critical findings emerge:

Methodological Contributions:

1. Data Leakage Identification: The study demonstrates the critical importance of feature selection in educational data mining. The initial inclusion of Exam Score a variable directly deterministic of FinalGrade produced artificially inflated performance metrics (R^2 = 1.0 for Random Forest). This serves as a cautionary example for researchers working with derived target variables.

2. Model Selection Justification: After removing the leakage variable, Random Forest emerged as the superior model (R^2 = 0.792, Weighted F1 = 0.842), significantly outperforming both Linear Regression (R^2 = -0.004) and the baseline Neural Network (R^2 = 0.086). This indicates that student performance is governed by nonlinear interactions among behavioral variables rather than simple additive effects.

3. Behavioral Over Demographic: Feature importance analysis reveals that academic engagement behaviors particularly assignment completion (18.3%), attendance (18.0%), and study hours (16.5%)are substantially more predictive than demographic characteristics. Gender contributed only 3.0% to predictive power, suggesting that well designed educational interventions can benefit all students equitably.

Practical Implications:

- Early Intervention Systems: Institutions can deploy real-time monitoring of assignment completion and attendance to identify at-risk students before final examinations.

- Resource Allocation: The strong predictive value of online course engagement suggests that investments in digital learning infrastructure may yield measurable improvements in student outcomes.

- Personalized Support: The ordinal nature of prediction errors (misclassifications primarily between adjacent grades) indicates that models can reliably distinguish performance tiers, enabling targeted academic support.

Limitations:

1. Cross-Sectional Design: Data represent a single academic period; longitudinal patterns of student development remain unexamined.

2. Institutional Specificity: Results derive from one educational context; generalizability to different institu tional cultures or educational systems requires validation.

3. Unobserved Confounders: Variables such as socioeconomic status, prior academic preparation, and mental health were not included but may influence both engagement behaviors and outcomes.

8. Future Work

8.1 Future Research Directions for AI-Driven Student Performance Prediction Systems

Temporal Modeling of Learning Trajectories: Future research should emphasize the development of temporal modeling techniques capable of capturing the dynamic evolution of student learning behaviors across academic semesters. Sequential deep learning architectures such as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) can be employed to model longitudinal patterns in student engagement, attendance, and assessment outcomes. These models enable the representation of temporal dependencies within educational datasets, thereby improving predictive accuracy for long-term academic performance.

In addition, early warning systems can be developed to dynamically update predictions as new behavioral data becomes available during the semester. Such systems allow educators to identify at-risk students earlier in the learning cycle and implement timely academic interventions. By continuously integrating newly gene-rated data from academic activities, these predictive frameworks can support proactive educational decision making and reduce the risk of academic underperformance.

Integration of Explainable Artificial Intelligence: While advanced predictive models often achieve high accuracy, their interpretability remains a critical challenge in educational analytics. Integrating explainable artificial intelligence (XAI) techniques can enhance transparency and trust in predictive systems. Methods such as SHAP (SHapley Additive exPlanations) can be incorporated to compute feature level contributions for individual predictions. These techniques enable researchers and educators to understand how specific behavioral factors such as attendance, assignment submission patterns, or digital engagement affect predicted academic outcomes.

Furthermore, developing student facing analytics dashboards can translate complex model outputs into accessible, actionable insights. Such dashboards could visually present the most influential behavioral indicators and provide personalized recommendations for academic improvement. This approach not only increases model transparency but also empowers students to actively participate in their own learning optim-ization.

Intervention Simulation through Counterfactual Analysis: Another promising research direction involves the use of counterfactual reasoning to simulate potential academic improvements resulting from behavioral modifications. Counterfactual analysis can estimate hypothetical outcomes under alternative scenarios for

example, predicting how an increase in attendance or study time might influence a student's final grade.

By applying these techniques, predictive systems can generate scenario-based recommendations, such as estimating how much predicted grades would improve if a student's participation or assignment completion rate increased. Such simulations can assist educators and students in identifying the most impactful behavioral adjustments, thereby transforming predictive analytics into practical intervention planning tools.

Multi-Modal Data Fusion in Learning Analytics: Modern educational environments generate diverse data streams, including learning management system (LMS) logs, discussion forum interactions, and multimedia engagement metrics. Integrating these heterogeneous datasets through multi-modal data fusion techniques can significantly enhance the richness of predictive models.

Future studies should explore integrating LMS clickstream data, textual analysis of discussion forum participation, and metrics from educational video platforms. Advanced natural language processing methods can be used to analyze discourse patterns in online discussions, while engagement metrics such as video watch duration and interaction frequency can provide insights into learning behavior.

In addition, emerging computer vision approaches may be employed to analyze environmental factors influencing study behavior, such as workspace organization or posture during learning sessions. However, these approaches must incorporate rigorous privacy safeguards and ethical data governance protocols to ensure responsible implementation.

Causal Inference for Educational Decision Support: Most existing student performance prediction systems focus primarily on correlation based prediction rather than causal explanation. Future research should therefore incorporate causal inference frameworks to determine whether specific educational interventions directly influence academic outcomes.

Methods such as Propensity Score Matching and Instrumental Variable Analysis can be applied to estimate causal relationships between student behaviors and performance indicators. These approaches help control for confounding variables and allow researchers to assess the true impact of interventions such as tutoring programs, attendance policies, or digital learning resources.

Moreover, randomized controlled trials (RCTs) can be designed to empirically validate intervention strategies recommended by predictive models. Combining predictive analytics with rigorous experimental evaluation can strengthen the evidence base for data driven educational policies.

Cross-Institutional Validation and Federated Learning: The generalizability of predictive models is often limited when they are trained on data from a single institution. To address this limitation, collaborative research across multiple educational institutions is essential. Cross-institutional datasets can enhance model robustness and enable the identification of contextual factors influencing academic performance.

A promising approach for such collaborations is Federated Learning, which allows multiple institutions to jointly train predictive models without directly sharing sensitive student data. In this framework, models are trained locally at each institution and aggregated centrally, thereby preserving privacy while benefiting from

larger datasets.

Additionally, future research should investigate how cultural, institutional, and disciplinary contexts influence the relative importance of predictive features. Understanding these moderators can improve the adaptability of models across diverse educational environments.

Integration with Adaptive Learning Systems: Predictive analytics can be further enhanced by integrating it with adaptive learning platforms that dynamically adjust instructional content. By combining predictive models with adaptive content delivery systems, personalized learning pathways can be generated in real time based on student progress and engagement patterns.

Recent advances in reinforcement learning offer additional opportunities for optimizing educational interventions. Reinforcement learning agents can learn optimal strategies for determining when and how to deliver academic support, such as recommending specific learning resources, scheduling tutoring sessions, or modifying course pacing. Through continuous interaction with learners, these systems can progressively improve their effectiveness in supporting individualized learning outcomes.

Equity-Centred Model Development: An important challenge in educational data science is ensuring that predictive models do not inadvertently reinforce existing social or educational inequalities. Future research should prioritize the development of fairness aware machine learning approaches that explicitly account for demographic disparities in data repres- entation and outcomes.

Techniques for bias detection and mitigation can be integrated into the model development pipeline to ensure equitable performance across different demographic groups. In addition, continuous auditing protocols should be implemented to monitor predictive accuracy and fairness metrics over time. Such safeguards are essential for maintaining ethical standards in the deployment of AI-based educational technologies.

Student Agency and Participatory System Design: Finally, the long-term success of predictive learning systems depends on students' active involvement in their design and implementation. Participatory design approaches can help ensure that predictive systems align with student needs, expectations, and ethical concerns.

Engaging students in the co-design process can improve transparency, foster trust in predictive technologies, and enhance their educational value. Furthermore, integrating model literacy into the curriculum can help students understand the capabilities and limitations of predictive analytics. By developing awareness of how data driven insights are generated and interpreted, students can use predictive feedback more effectively to guide their learning strategies.

Overall, future research in AI-based student performance prediction should move beyond static prediction models toward dynamic, explainable, and ethically responsible systems. Integrating temporal modeling, explainable AI, causal inference, multi-modal data fusion, and adaptive learning technologies can significantly enhance the impact of predictive analytics in education. Equally important are cross institutional collaboration, fairness aware modeling, and student participation in system design. Collectively, these directions will contr- ibute to the development of intelligent educational ecosystems that support personalized learning, early int- ervention, and equitable academic success.

References

[1] Manjarres, A. V., Sandoval, L. G. M., Suárez, M. S. (2018). Data mining techniques applied in educational environments: *Literature review*. *Digit. Educ. Rev*. 2018, 33, 235–266.

[2] Zareie, B., Navimipour, N. J. (2016). The effect of electronic learning systems on the employee's commitment. *Int. J. Manag. Educ*. 2016, 14, 167–175.

[3] Munisami, A., Alasiry, A. (2020). Deep Learning: The Impact on Future eLearning. *Int. J. Emerg. Technol. Learn*. 2020, 15, 188–199.

[4] Alharthi, A. D., Spichkova, M., Hamilton, M. (2019). Sustainability requirements for elearning systems: A systematic literature review and analysis. *Requir. Eng*. 2019, 24, 523–543.

[5] Umer, R., Susnjak, T., Mathrani, A., Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. *J. Res. Innov. Teach. Learn*. 2017, 10, 160–176.

[6] Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., Yang, S. J. (2018). Applying learning analytics for the early prediction of Students' academic performance in blended learning. *J. Educ. Technol. Soc*. 2018, 21, 220–232.

[7] Widyahastuti, F., Tjhin, V.U. (2018). Performance prediction in online discussion forum: State-of-the-art and comparative analysis. *Procedia Comput*. Sci. 2018, 135, 302–314.

[8] Zhang, W., Huang, X., Wang, S., Shu, J., Liu, H., Chen, H. (2017). Student performance prediction via online learning behavior analytics. *In:* Proceedings of the 2017 *International Symposium on Educational Technology (ISET), Hong Kong, China*, 27–29 June 2017; p. 153–157.

[9] Koutina, M., Kermanidis, K. L. (2011). Predicting postgraduate students' performance using machine learning techniques. *In: Artificial Intelligence Applications and Innovations; Springer: Berlin/Heidelberg, Germany,* 2011; p. 159–168.

[10] Al Zahrani, N. A., Abdullah, M. A. (2019). Student Engagement Effectiveness in E-Learning System. *Biosci. Biotechnol. Res. Commun*. Spec. *Issue Commun. Inf. Technol*. 2019, 12, 208–218.

[11] Balcioðlu, Y. S., Artar, M. (2025). Predicting academic performance of students with machine learning. *Information Development*, 41 (3), 896-915.

[12] Zhiqiang Zhao, Ping Ren. (2025). Random Forest-Based Early Warning System for Student Dropout Using Behavioral Data. *Bulletin of Education And Psychology*.

[13] Omopariola, Adebola, Victor., Eniolorunda, Wande Stephen. (2025). Implementation of Machine Learning based School Class Placement Prediction Systems for Secondary School Using Random Forest, *Journal of Science Education and Research*, 3 (1) 61-80.

[14] Li, J., Chen, X. (2025). Modeling student satisfaction in online learning using random forest. *Sci Rep* 15, 23254.

[15] Li, C., Cao, Z. (2025). Deep learning-based AI model for predicting academic success and engagement among physical higher education students. *Sci Rep* 15, 45471 (2025)

[16] Balabied SAA, Eid H F. (2023). Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. *PeerJ Computer Science* 9:e1708.

[17] Zhao, S., Zhou, D., Wang, H., Chen, D., Yu, L. (2025). Enhancing Student Academic Success Prediction Through Ensemble Learning and Image-Based Behavioral Data Transformation. *Appl. Sci*., 15, 1231.

[18] Asif, R., Merceron, A., Ali, S. A., Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining, *Comput. Educ*. 2017, 113, 177–194.

[19] Yaðcý, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ*. 2022, 9, 11.

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

[21] Nayak, P., Vaheed, S., Gupta, S., Mohan, N. (2023). Predicting students' academic performance by mining the educational data through a machine learning-based classification model. *Educ. Inf. Technol*. 2023, 28, 14611–14637.

[22] Ram, M. S., Srija, V., Bhargav, V., Madhavi, A., Kumar, G. S. (2021). Machine Learning-Based Student Academic Performance Prediction. *In:* Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021; p. 683–688.

[23] Aljohani, N. R., Fayoumi, A., Hassan, S. U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability* 11, 7238.

[24] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Comput. Hum. Behav*. 2020, 104, 106189.

[25] Huang, Q., Zeng, Y. (2024). Improving academic performance predictions with dual graph neural networks. *Complex Intell. Syst*. 2024, 10, 3557–3575.

[26] Ben Said, A., Abdel-Salam, A. S.G., Hazaa, K. A. (2024). Performance prediction in online academic course: A deep learning approach with time series imaging. *Multimed. Tools Appl*. 2024, 83, 55427–55445.

[27] Yang, C. L., Chen, Z. X., Yang, C. Y. (2019). Sensor classification using a convolutional neural network by

encoding multivariate time series as two-dimensional colored images. *Sensors* 2019, 20, 168.

[28] Li, J., Wang, Q. (2023). Comparison of the representational ability in individual difference analysis using 2-D time-series image and time-series feature patterns. *Expert Syst. Appl.* 2023, 215, 119429.

[29] Yin, J., Zhuang, X., Sui, W., Sheng, Y., Yang, Y. (2024). A new similarity measurement method for time series based on image fusion of recurrence plots and wavelet scalogram. *Eng. Appl. Artif. Intell.* 2024, 129, 107679.

[30] Najem, K. (2025). Student Performance and Learning Behavior Dataset for Educational Analytics [Data set]. Zenodo. https://doi.org/10.5281/zenodo.16459132.