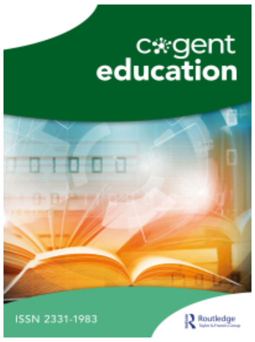
DIEKUU, J.-B., MEKALA, M.S., ABONIE, U.S., ISAACS, J. and ELYAN, E. 2025. Predicting student next-term performance in degree programs using AI-based approach: a case study from Ghana. *Cogent education* [online], 12(1), article number 2481000. Available from: https://doi.org/10.1080/2331186X.2025.2481000

Predicting student next-term performance in degree programs using AI-based approach: a case study from Ghana.

DIEKUU, J.-B., MEKALA, M.S., ABONIE, U.S., ISAACS, J. and ELYAN, E.

2025



Taylor & Francis
Taylor & Francis Group

Cogent Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/oaed20

Predicting student next-term performance in degree programs using AI-based approach: a case study from Ghana

John-Bosco Diekuu, M. S. Mekala, Ulric Sena Abonie, John Isaacs & Eyad Elyan

To cite this article: John-Bosco Diekuu, M. S. Mekala, Ulric Sena Abonie, John Isaacs & Eyad Elyan (2025) Predicting student next-term performance in degree programs using AI-based approach: a case study from Ghana, Cogent Education, 12:1, 2481000, DOI: 10.1080/2331186X.2025.2481000

To link to this article: https://doi.org/10.1080/2331186X.2025.2481000

Published online: 25 Mar 2025.

COGENT EDUCATION
2025, VOL. 12, NO. 1, 2481000
https://doi.org/10.1080/2331186X.2025.2481000

INFORMATION & COMMUNICATIONS TECHNOLOGY IN EDUCATION | RESEARCH ARTICLE
OPEN ACCESS Check for updates

Predicting student next-term performance in degree programs using AI-based approach: a case study from Ghana

John-Bosco Diekuu[a,b], M. S. Mekala[a], Ulric Sena Abonie[c], John Isaacs[a] and Eyad Elyan[a]

[a]School of Computing, Robert Gordon University, Aberdeen, UK; [b]ICT Directorate, University of Health and Allied Sciences, Ho, Ghana; [c]Department of Sport Exercise and Rehabilitation, Northumbria University, Newcastle, UK

ABSTRACT
Student performance can fluctuate over time due to various factors (e.g. previous assignment grades, social life and economic conditions). Temporal dynamics, such as semester-to-semester variations and changes in students' academic achievements, behaviors and engagement over time, can be critical factors in designing predictive models. It can be said that most existing work focuses on one-time forecasting of student performance in specific semesters, subjects or short online courses without considering temporal elements. In this paper, we present a student performance-based temporal dynamic approach to progressively predict semester-wise performance. Eight semesters of data representing 3,093 undergraduate Health Sciences students was collected from a public university in Ghana, analyzed, pre-processed and transformed into a time-series format. Then a dynamic experimental framework utilizing four different machine learning methods to predict student performance was created. This includes Random Forest, Support Vector Machine, Long Short-Term Memory and Bidirectional Long Short-Term Memory to predict student performance semester-wise over eight semesters. The results indicate that utilizing past students' performance records obtained over time enhances the accuracy of forecasting their performance in future semesters. Moreover, the results evident that high school grades and semester GPAs are the most powerful discriminant features influencing the models' performance, emphasizing the importance of consistent in-course performance.

ARTICLE HISTORY
Received 19 August 2024
Revised 11 February 2025
Accepted 10 March 2025

KEYWORDS
AI in Education; higher education; next-term performance; machine learning; LSTMs architecture

SUBJECTS
Artificial Intelligence; CAD CAE CAM - Computing & Information Technology; Information & Communication Technology (ICT); Higher Education Management

1. Introduction

Predicting students' future performance provides critical information to higher education (HE) institutions for academic decision-making. Academic achievements are a vital indicator of a student's knowledge, progress and future opportunities (Albreiki et al., 2021). However, accurately predicting students' future performance remains a significant concern for HE, as they continue to experience high attrition rates, delays in graduation timelines and inconsistencies in the quality of educational outcomes (Pelima et al., 2024). Extended periods to graduation or dropout place financial burdens on students and their families and strain the university's limited resources (Indicators, OECD, 2023). Additionally, students who graduate with weak grades and classes tend to experience increased stress and anxiety, encounter obstacles to further education, limited job prospects and potentially earn less over their lifetimes (Eisenberg et al., 2009; Tinto, 2012). These issues are particularly acute in developing countries, such as Ghana, where monitoring, forecasting and timely identifying struggling students and providing the necessary support is often limited (Tinto, 2012). Thus, AI-based solutions can be explored to fill this gap. This is imperative, as facilities are now available to digitize all student records throughout their academic

CONTACT John-Bosco Diekuu j.diekuu@rgu.ac.uk School of Computing, Robert Gordon University, Aberdeen, UK

journey, enabling AI-based methods to analyze and identify potential issues in real-time for timely interventions.

In recent times, AI has seen notable progress, especially in machine learning (ML), deep learning (DL) and data mining. These advancements have significantly enhanced prediction accuracy across various fields (Goodfellow et al., 2016). The education sector has particularly benefited from AI, utilizing it to address numerous educational challenges effectively. These include predicting student grades (Baashar et al., 2022; Karlos et al., 2020), analyzing images and videos to assess students levels of participation (Ashwin & Guddeti, 2020; Santoni et al., 2023), performing sentiment analyzes on student feedback (Dzikovska et al., 2014) and automating the creation and grading of assessments (Ahmed et al., 2022; Shehab et al., 2016). These solutions have demonstrated AI effectiveness in comprehending and analyzing student data, enabling more informed decision-making.

Despite the significant advancements, AI applications in HE remain predominantly traditional, focusing mainly on predicting student performance in specific semesters, subjects, exams or short online courses. Thus, the temporal dynamics of student performance over time in degree completion remains underexplored (Arqawi et al., 2022). Predicting student achievement in a degree program is different and presents unique challenges, including the need to consider the evolving academic development of students and the varied impact of their backgrounds.

Students' diversity in backgrounds (academic and non-academic), socioeconomic factors, and areas of interest significantly influence their performance. For example, the choice of student programs might vary greatly, leading to various course selections and sequencing, with different weights or difficulties (Adnan et al., 2021). Students from various locations with different resources, economic conditions, and access to information might enroll in the same course (Rodríguez-Hernández et al., 2021). Age contributes to performance with older students, particularly women, often employed deep learning strategies, positively affecting retention rates (Suleiman & Anane, 2022). Previous academic performance, including high school achievements and university grades, has consistently been identified as a key determinant of student success (Duong et al., 2023; Polyzou & Karypis, 2018; Xu et al., 2017). There is a strong need to identify critical factors that influence student performance to ensure effective targeting of resources and interventions in the Ghanaian cultural context. These issues are complex and interplay in degree studies.

In this study, we propose an AI-driven approach for progressive prediction of students' next-term performance in health-related degree programs using actual data from a public university in Ghana. It further aims to identify critical *factors influencing temporal dynamics*[1] that contribute to student performance and provide insights into their behavior within the Ghanaian cultural context. This method emphasizes the continuous nature of academic success by analyzing students' semester GPAs and background information, highlighting the importance of long-term temporal patterns. The main contributions are outlined as follows:

- A dataset representing students' information over a 5-year period has been utilized for this study. The data includes students' demographic and actual academic performance records of 3,093 undergraduate students.
- A feature-engineering method was designed to identify common potential features that impact model accuracy along with their correlation matrix. The results clearly show that considering semester-wise common identical features has drastically enhanced student performance over time.
- A dynamic experimental framework was created that extensively trains four different ML models using time-series semester data. The framework utilizes sequence-based models, namely Long Short-Term Memory (LSTM) and bidirectional LSTM (BiLSTM), alongside traditional ML-based methods, such as Random Forest (RF) and Support Vector Machine (SVM), to progressively predict student semester-wise performance in eight semesters of undergraduate studies.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 outlines the proposed methodology, including the dataset, models and evaluation metrics. In Section 4, we present and discuss the experimental results in context and outline the limitations of the study. Section 5 concludes the paper and recommends future research areas.

2. Related work

Recent advancements in AI for education have leveraged various machine learning (ML) techniques to predict student performance and identify at-risk students. Methods such as Decision Trees (DT) (Trakunphutthirak & Lee, 2022), Artificial Neural Networks (ANN) (Rodríguez-Hernández et al., 2021), Deep Neural Networks (Riestra-González et al., 2021) and RF (Zhang et al., 2022) have demonstrated varying success across different contexts.

Yağcı (2022) developed an ML-based model to predict students' final exam grades using midterm scores, department information and faculty data. Evaluating multiple classifiers, including RF, SVM, KNN, Logistic Regression and Naïve Bayes, on 1,854 students from a Turkish Language course, the study achieved an accuracy of 70–75%. The findings highlight midterm grades as significant predictors of final performance, reinforcing the role of educational data mining in early intervention.

Alhazmi and Sheneamer (2023) proposed an ML framework for early student performance prediction by applying clustering and classification techniques. Using t-SNE for dimensionality reduction, the study analyzed admission scores, first-year course grades, and standardized test scores to predict GPA. Evaluating models such as XGBoost, RF and SVM, the study found that incorporating early course performance significantly enhanced prediction accuracy, underscoring the importance of early academic indicators.

Similarly, Cruz-Jesus et al. (2020) examined academic performance prediction in Portuguese public high schools, leveraging data from 110,627 students. Comparing multiple ML models, the study found that RF outperformed others, emphasizing the influence of demographic, socioeconomic and academic variables on student success. These findings suggest that AI-driven models offer valuable insights for policymakers and educators in mitigating dropout rates.

Fernandes et al. (2019) conducted a predictive analysis of student performance in Brazil's Federal District, employing Gradient Boosting Machines (GBM) on demographic and academic datasets. While grades and absences were the strongest predictors, factors such as neighborhood, school and age also influenced performance. This highlights the necessity of incorporating both academic and socioeconomic variables in predictive models.

Xu et al. (2019) examined the relationship between internet usage behaviors and academic performance using ML methods. The study extracted features such as online duration, internet traffic volume and connection frequency from a dataset of 4,000 university students. DT, neural networks and SVM were employed to predict student performance. The findings revealed that disciplined internet behavior is a strong predictor of academic success, with internet connection frequency positively correlating with academic performance, while internet traffic volume showed a negative correlation.

Waheed et al. (2020) developed a deep learning model to predict student performance using Virtual Learning Environment (VLE) big data. The study applied deep ANN to clickstream data to identify at-risk students for early intervention. Results showed that the ANN achieved an accuracy of 84%–93%, outperforming logistic regression and SVM. The study concluded that students who accessed past lecture materials performed better, highlighting the importance of legacy data and assessment-related information in predicting student success.

Beyond static student performance prediction, some studies integrate temporal analysis to track student progress. Asif et al. (2017) applied educational data mining techniques to undergraduate student performance over a 4-year degree program. Using DT and other ML classifiers, the study demonstrated how pre-admission marks and early university coursework predict final academic achievement. By clustering students based on performance progression, key courses were identified as early indicators for timely intervention.

Shou et al. (2024) introduced a multidimensional time-series approach for student performance prediction. The model employed a multi-head self-attention mechanism and LSTM networks to enhance accuracy by integrating learning behaviors, assessment scores and demographic information. Experiments on the Open University Learning Analytics Dataset (OULAD) achieved a 74% accuracy and 73% F1-score for multi-class prediction, with early risk detection reaching 99.08% accuracy. The study highlights the effectiveness of attention mechanisms in capturing relationships between factors affecting student outcomes.

Deeva et al. (2022) proposed a sequence-based classification model leveraging behavioral data from online and blended learning environments. By implementing time-based windows to capture the temporal aspects of student interactions, the study found that course-specific models achieved up to 90% accuracy, outperforming generalized approaches.

Similarly, Delianidi et al. (2021) formulated a dynamic neural network model for student performance prediction using a sequence-learning framework. The study compared Time-Delay Neural Networks (TDNN) and Recurrent Neural Networks (RNN), demonstrating that RNN-based models outperformed state-of-the-art methods in knowledge-tracing tasks.

Despite these advancements, most studies fail to fully exploit the sequential nature of knowledge acquisition or adaptively scale models to expanding input spaces. This study introduces a progressive AI-based prediction architecture designed to integrate temporal behavioral indicators, ensuring that each predictor performs at least as well as or better than previous models. By accounting for academic progression, the proposed method enhances early intervention strategies and improves long-term student success predictions.

3. Materials and methods

In this section, we present the materials and methods used in our study. We begin by describing the dataset and its characteristics, followed by the steps involved in the data pre-processing, learning models and performance evaluation. As can be seen in the schematic diagram provided in Figure 1, four different ML algorithms were trained. Each model performed both regression and classification tasks. The regression method was designed to predict grade point averages (GPAs) and the final GPA, while the classification method was designed as a multiclass problem, forecasting the class of the students (i.e. first class, second class upper, second class lower, third class and pass).

3.1. Dataset

The dataset was constructed from several disjoint databases related to students enrolled in health-related undergraduate degree programs at a public university in Ghana, covering students enrolled in the 2015, 2016, 2017, 2018 and 2019 academic years. The dataset captured records of 3,239 anonymized students with 40 features, covering 8 semesters of each cohort. The students were enrolled in 14 different 4-year degree programs. See the data distributions in Table 1.

The dataset captures various information: Demographics (gender, date of birth, nationality, disability, bursary, location); high school traits (program, grade, number of times student attempted exams); university program details (program, whether the student was offered their preferred program, enrollment year, study type); and academic records including level, semester, grade point averages (GPAs), number of failed courses per semester, total credit hours per semester, final grade point average (FGPA) and graduating class. The graduating class of students includes First class, Second class upper, Second class lower, Third class, Pass and Withdrawn. These variables are presented in Table 2.

The Ghanaian education system adopts a semester-based approach; each academic term is a semester. There are two semesters in a year; we considered all eight (8) semesters in the 4-year cycle. We

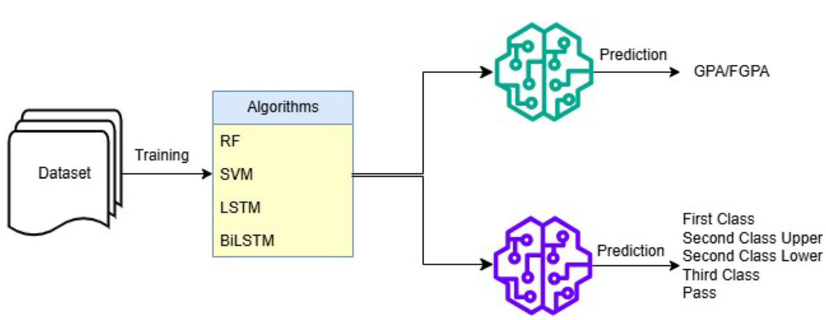


Figure 1. Schematic diagram illustrating the methodology.

Table 1. Frequency and percentage of students per program.

Program	Number ↓	Percent%
Nursing	904	30.10
Midwifery	366	11.30
Medical Laboratory Sciences	353	10.90
Physician Assistantship	299	9.23
Disease Control	296	9.14
Biomedical Sciences and Molecular Biology	180	5.56
Health Nutrition	177	5.46
Dietetics	176	5.43
Health Promotion	137	4.23
Health Information	126	3.89
Physiotherapy	125	3.86
Diagnostic Imaging	56	1.73
Speech, Language and Hearing Sciences	44	1.36
Total	3239	100.00

Table 2. Base data description.

Feature class	Feature name	Type	Description	Values
Demographics	Age	Numeric	Age at enrollment	Continuous Min (16), Max (38)
	Sex	Categorical	Gender of student	male, female
	Nationality	Categorical	Country of origin	Ghana, Nigeria, Togo etc
	Disability	Binary	Physical disability	yes, no
	Bursary	Binary	Government fees subsidy	yes, no
	Location	Categorical	Student home residential region	Ashanti, Brong Ahafo, Central, Eastern, Greater Accra, Northern, Upper East, Upper West, Volta, Western
High school traits	Program	Categorical	Program studied	science, Arts, Home economics, etc
	Grade	Numeric	Admission grade	min (36), max (6)
	No_attempt	Nominal	No. of exam attempts	1 = 1 attempt 2 = 2 attempts 3= above 2 attempts
University program	Program	Categorical	Program studied	Nursing, Midwifery, Disease control, Nutrition, Health Promotion, Health information, Physician assistantship, Medical laboratory, Diagnostic imaging, Physiotherapy, Speech and language therapy, Dietetics, and Biomedical and molecular biology
	Is_choice	Binary	Is the program student's preferred choice?	yes, no
	Year_group	Nominal	Year admitted	2015, 2016, 2017, 2018, 2019
	Semester	Nominal	academic terms	1 - 8
	Study type	Categorical	Type of studies	Regular, sandwich, weekends
Academic records	Levels	Nominal	academic stage	100, 200, 300, 400
	CR (Sem 1–8)	Nominal	Credit hours per semester	min (16), max (26)
	Fails (Sem 1–8)	Numeric	Count of failed courses per semester	1 = 1 course, 2 = 2 courses, 3= above 2 courses
	Sem_GPA (1–8)	Numeric	GPA per semester	min (0.0), max (4.0)
	Yearly_GPA (1–4)	Numeric	Average GPA per year	min (0.0), max (4.0)
	FGPA	Numeric	Final Cumulative GPA for the 4-year	min (0.0), max (4.0)
Student standing	Graduating class	Categorical	Student status at the end of 4 years	3.6–4.0 = First Class, 3.0–3.59 = Second Class Upper, 2.5–2.99 = Second Class Lower, 2.0–2.49 = Third Class, 1.0–1.9 = Pass

utilized the semesters' GPAs but not the letter grades awarded in courses (e.g. A, B, C, etc.). This is because all the students have studied different courses, sometimes different courses within the same program with different weights. This made it difficult to use the scores obtained in individual courses. However, GPA is the weighted average of the scores in all courses studied in a particular semester, reflecting the equal representation of all students' performance (Xu et al., 2017).

The University categorized students with a GPA below 2.5 in any semester as underperforming. The GPA range of 0.0–2.49 consists of third class, pass or fail categories, with a substantial number of such students either withdrawing or graduating with weak grades. It was observed that over 4.5% of the students were withdrawn, while 16% completed with a pass or third class, consistent with the high global dropout rates in higher education institutions (Araque et al., 2009; OECD, 2023). Thus, Our target groups within this study are students who belong to the 'third class' and 'pass' categories. These categories of students stand the risk of not graduating on time or completing with weak passes. They are particularly at higher risk of not getting sustainable jobs in regions like Ghana, where the unemployment rate among tertiary graduates stands at a notably high of 22.3% (GSS, 2024).

3.1.1. Data pre-processing

The dataset was pre-processed to align with this study's primary objective of accurately predicting students' next-term performance. Given the nature of the dataset, each student is represented by multiple rows corresponding to different semesters and levels of their academic journey, leading to redundant information. To address this, we transformed the dataset into a time-series format, consolidating each student's data into a single row. Table 3 illustrates examples of the transformed dataset, showcasing a few records for each student. This transformation facilitates a more comprehensive analysis of individual academic performance over time. Capturing the temporal progression of performance provides valuable insights into long-term dependencies and factors influencing students' learning outcomes throughout their studies.

Rows in the dataset with more than 70% missing or invalid values in the semester GPA columns were removed. Additionally, records of withdrawn students at various stages of their academic trajectory were excluded. These students lacked a complete set of results to track the progression of academic performance. Hence, their inclusion in the analysis would not have been meaningful. This resulted in the removal of 146 records from the initial 3,239 rows, leaving a final dataset of 3,093 students who successfully completed their studies and graduated from the university.

We excluded columns in the dataset with more than 70% missing or invalid values, as well as columns containing identical input values. This step eliminated 14 columns out of the initial 40, minimizing the reliance on synthetic data. To further optimize the feature space and ensure low correlation among features, any feature in a pair with more than 80% correlation was removed, following (Manigandan et al., 2024) work. These steps reduced the dimensionality of the dataset, leaving 17 features for analysis. The correlation between the selected features is illustrated in Figure 2. Finally, the K-Nearest Neighbors (KNN) Imputation technique (Pujianto et al., 2019) was applied to handle the remaining missing values in the dataset.

Additionally, feature importance was conducted to determine the impact of each variable on the model performance. It was noticed that all features with importance scores below 1.0% have minimal impact on performance. Hence, 5 features were further excluded from the final dataset. Overall, 12 features were included in the final experiment.

Demographic and high school features were utilized as static and common features across all prediction levels. However, the framework is flexible and can accommodate additional variables if available. Each dataset was sequentially structured to include information up to the target term. For example, Set_1, used to predict semester 2 performance, combines static features with semester 1 GPAs. Similarly, Set_2 (for predicting semester 3) incorporates static variables alongside semester 1 and 2 GPAs. This sequential logic continues for all subsequent datasets, with each set as the foundation for predicting the semester's performance immediately following its timeframe, as outlined in Table 4.

For classification purposes, the target GPAs of each semester were encoded into five categories, as detailed in Table 5, while the actual GPA values from preceding semesters were used as predictors. This approach ensures a structured and consistent progression in predicting academic performance across multiple levels.

This study explores whether a classifier trained to detect the confusion or frustration affective states may achieve higher overall accuracy than engagement detection alone, even with the poor class imbalance.

3.2. Learning models

RF, SVM, LSTM and BiLSTM were chosen as the learning algorithms. Reviews in (Albreiki et al., 2021; Namoun & Alshanqiti, 2020) have shown that RF and SVM are considered to be the most widely used learning methods in the prediction of educational learning outcomes. Therefore, RF and SVM were used as baseline models. On the other hand, LSTM and BiLSTM architectures are effective and widely used in

Table 3. Example dataset rows for each unique student.

Gender	Age	Location	Grade	Program	Exam_attempt	Is_Choice	Bursary	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Sem7	Sem8	FGPA	Class
Male	20	Greater Accra	25	Dietetics	1	True	Yes	3.55	3.67	3.25	3.43	3.91	3.76	3.89	3.50	3.64	First Class
Female	18	Volta	23	Health Information	3	False	Yes	2.97	2.80	3.24	2.31	2.98	3.63	3.28	3.70	3.12	Second Class Upper
Male	22	Greater Accra	15	Medical Lab Sciences	1	True	Yes	3.48	3.64	3.63	3.84	3.08	3.57	3.67	3.82	3.60	First Class
Female	18	Ashanti	17	Medical Lab Sciences	1	True	Yes	3.36	3.77	3.76	3.74	3.17	3.73	3.47	3.82	3.61	First Class
Female	21	Upper East	27	Health Promotion	2	True	No	2.86	2.71	2.24	2.45	2.52	3.00	2.89	3.26	2.73	Second Class Lower
Male	22	Brong Ahafo	22	Physiotherapy	1	False	Yes	3.60	3.75	2.95	3.76	3.74	3.85	3.68	3.91	3.66	First Class
Male	20	Greater Accra	19	Midwifery	1	False	Yes	3.66	3.88	3.73	3.67	3.67	3.84	3.58	3.55	3.70	First Class
Male	19	Northern	23	disease Control	1	True	Yes	2.75	1.84	2.34	2.19	1.80	2.39	2.35	2.64	2.27	Third Class
Female	20	Volta	16	Nursing	1	True	Yes	3.93	3.75	3.77	3.80	3.71	3.74	3.89	3.11	3.72	First Class
Female	21	Eastern	13	Nursing	1	False	Yes	1.23	1.30	1.22	1.26	2.16	1.98	2.18	2.76	1.73	Pass

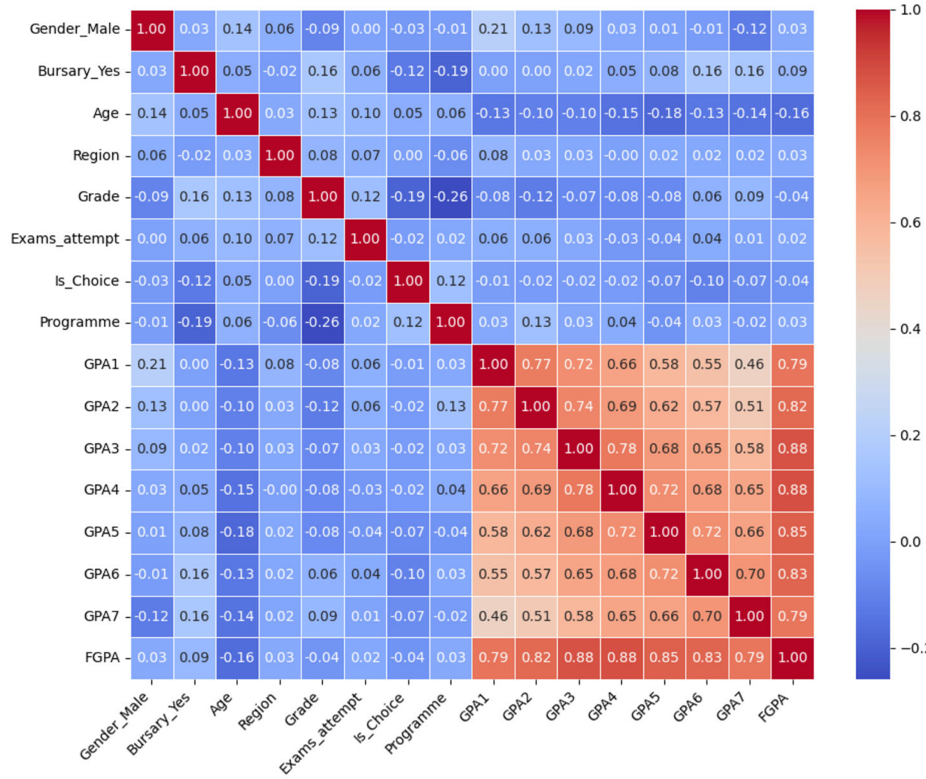


Figure 2. Pearson correlation coefficient between the input features.

Table 4. Overview of next-term prediction datasets.

Initials	Included semesters	Sample size	Prediction
Set_1	1	3093	Semester 2 performance
Set_2	1–2	3093	Semester 3 performance
Set_3	1–3	3093	Semester 4 performance
Set_4	1–4	3093	Semester 5 performance
Set_5	1–5	3093	Semester 6 performance
Set_6	1–6	3093	Semester 7 performance
Set_7	1–7	3093	Semester 8 performance

Table 5. Statistics of students' final class distribution.

GPA range	Number of students	Label	Risk level
3.6–4.0	126	First Class	No Risk
3.0–3.59	1463	Second Class Upper	No Risk
2.5–2.99	1009	Second Class Lower	No Risk
2.0–2.49	401	Third Class	Low Risk
1.0–1.99	94	Pass	High Risk

modeling time-series data (Cheng et al., 2022). It should be noted that our data have been transformed into time-dependent sequences, making the selection of LSTM and BiLSTM particularly suitable for this study. These models are described in detail below.

3.2.1. LSTM

LSTM networks are a specialized type of recurrent neural network where the hidden layer updates are replaced by purpose-built memory cells (Huang et al., 2015). These memory cells are designed to better capture and exploit long-range dependencies in the student semester-wise data, enhancing the network's ability to learn from and make predictions based on distant historical information. Figure 3 illustrates a single LSTM memory cell (Graves et al., 2013). Given that the input data includes both static

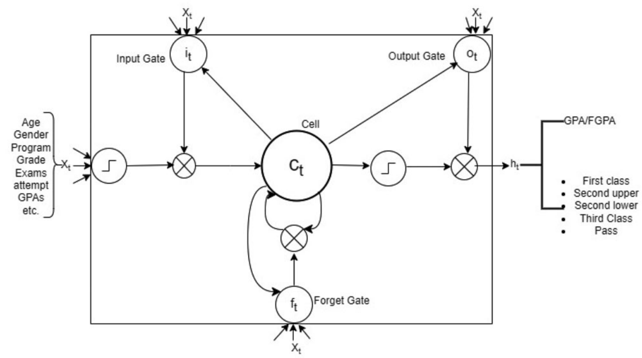


Figure 3. Long short-term memory cell.

demographic data and dynamic data (GPA for each semester), the LSTM memory cell is implemented as follows:

$$i_t = \sigma(W_{x_i} \cdot x_t + W_{h_i} \cdot h_{t-1} + W_{c_i} \cdot c_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{x_f} \cdot x_t + W_{h_f} \cdot h_{t-1} + W_{c_f} \cdot c_{t-1} + b_f) \tag{2}$$

$$c_t = f_t \cdot c_{t-1} + i_t \tanh(W_{x_c} \cdot x_t + W_{h_c} \cdot h_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{x_o} \cdot x_t + W_{h_o} \cdot h_{t-1} + W_{c_o} \cdot c_t + b_o) \tag{4}$$

$$h_t = o_t \tanh(c_t) \tag{5}$$

where i, f, o and c are the input gate, forget gate, output gate and cell vectors, all the same size as the hidden vector h. While σ is the logistic sigmoid function used in the gates to determine how much training data should be retained, forgotten or passed on. At each time step t (semester), the input Sequence (x_t) concatenates the student static demographic features and sequence of GPA records over semesters. Each element x_t represents the GPA for the t-th semester. The input gate (i_t) determines how much current input data should be added to the cell state. While input bias (b_i) is the bias vector (b) added to the input gate to adjust its sensitivity.

The forget gate (f_t) controls how much information from previous semesters should be retained or discarded. While the forget bias (b_f) is added to adjust its tendency to forget past information. The forget gate weight matrix (W_f) controls the influence of the input and previous hidden state on the forget gate's operation.

Cell state (C_t) is the long-term memory of the LSTM that stores accumulated student static data and GPAs over time. It gets updated every time step t based on the input and forget gates. The cell vector sequence (c_{t-1}) acts as the sequence of cell states over time, representing the evolving memory as the model processes each semester's GPA.

Hidden state (h_t) is the short-term memory of the LSTM used to predict the semester-wise GPAs, FGPA or the Class. Hidden vector sequence (h_{t-1}) shows how the model's short-term memory evolves with each semester's GPA. Hidden-input gate matrix (W_{h_i}) is the weights that determine how the previous hidden state influences the input gate.

Output gate (o_t) determines how much of the cell state's information should be used in the hidden state for predicting the student's future performance. Output bias (b_o) is the bias added to the output gate to adjust its operation. The weight matrices (e.g. W_x, W_h, W_c) are learned parameters that transform the input data, hidden states and gates. They define how much influence the GPA of each semester has on future predictions. The weight matrices from the cell to gate vectors (e.g. W_{c_i}) are diagonal, so element m in each gate vector only receives input from element m of the cell vector.

LSTM is included in this study because of its prowess in modeling time-dependent or sequential data (Hochreiter & Schmidhuber, 1997), making them particularly effective for analyzing student academic progress over time. Their memory mechanism, which includes forget, input and output gates, enables the retention of long-term dependencies, making it possible to capture trends and anomalies in longitudinal student data (Sundermeyer et al., 2012). Additionally, LSTMs are adept at learning complex nonlinear relationships, such as those inherent in student behavioral or performance patterns. LSTM can be used for various tasks, including classification, regression and forecasting, depending on the specific requirements of the student dataset. However, it should be noted that LSTMs can be computationally intensive to train, particularly when working with large datasets, as they require substantial resources to optimize their complex architecture. They are also prone to overfitting, especially in small datasets common in specific educational contexts, unless regularization techniques are applied effectively (Huang et al., 2015).

3.2.2. Bidirectional LSTM

BiLSTM networks were utilized to process student semester-wise performance records forward and backward through two separate LSTM networks connected to the same output layer (Huang et al., 2015). This approach is particularly effective in predicting future student performance, where the model can leverage past and future semester GPAs for a specific time. The BiLSTM allows the model to utilize past semester performances and static demographic features through forward states while incorporating insights from future semesters via backward states (Huang et al., 2015). This dual perspective enables a more comprehensive understanding of a student's performance trajectory. During the training process, the forward and backward passes over the network were conducted similarly to traditional LSTM networks, with the added complexity of unfolding hidden states across all time steps. Special handling is applied at the start and end of each input data sequence; specifically, the hidden states are reset to zero at the beginning to ensure that the model processes each student's data independently. Additionally, a batch implementation was employed, allowing multiple student records to be processed simultaneously. This efficiently utilizes the available data to enhance the model's accuracy and robustness.

However, it should be noted that BiLSTM presents notable limitations. They are computationally more intensive than standard LSTMs, requiring double the memory and processing power, which can make them inefficient for large datasets. Additionally, backward processing may not add significant value in all scenarios, potentially increasing computational overhead without substantial performance gains. Furthermore, BiLSTMs can struggle with sparse data, where their advantages over unidirectional LSTMs are less pronounced (Greff et al., 2017).

3.2.3. Random forest

RF is a popular and efficient algorithm for classification and regression tasks, introduced by Breiman (2001). It is an ensemble method composed of multiple decision trees (Kukkar et al., 2024), where each tree independently predicts the student's semester-wise performance. The overall prediction output for each semester is made by compiling and averaging the predictions from all the trees, resulting in a more robust and accurate decision. RF randomly selects a subset of the static demographic features and dynamic features (semesters' GPA) for each decision tree node, which enhances the diversity of the trees and reduces the likelihood of overfitting. This technique makes RF robust to noise and effective even when the training sample is imbalanced, as is often the case with student data, where the failed or at-risk students are always underrepresented. RF is simple to implement by relying on two primary input parameters throughout the creation process: the number of decision trees and the number of attributes considered at each node.

It's worth noting that the ensemble of decision trees with random feature selection enhances robustness against overfitting and improves generalization (Breiman, 2001), making RF a valuable model for analyzing student data. RF also provides insights into feature importance, making identifying key factors influencing student performance easier. Additionally, RF is scalable and efficient for moderately large datasets, as it can parallelize training across multiple trees. Notwithstanding, RF has some limitations. It struggles to capture temporal dependencies or sequential patterns, making it less suitable for analyzing longitudinal data. Furthermore, its performance is sensitive to hyperparameters, such as the number of trees and maximum depth, which require careful tuning for optimal results.

3.2.4. Support vector machine

SVM is a discriminative classifier that works by fitting a boundary to a set of points belonging to one class (Suthaharan & Suthaharan, 2016). In the context of student data, these points represent individual students, each characterized by multiple features such as demographic information and semester-wise performance. The boundary, supported by specific data points called support vectors, is critical in determining the separation between classes of student performance. In a two-dimensional space, this boundary could be a straight line or a curve, making visualization relatively straightforward. However, our adaptive framework expands as the number of features increases at any time step (semester), making visualizing the boundary challenging. SVM aims to maximize the margin, which is the distance between the closest data points and the boundary, ensuring that the model is as robust (Lee & Shin, 2020). This margin maximization may not always be achievable with complex, non-linear datasets. Our dataset comprises static features and dynamic GPAs for each semester making it not linearly separable. A non-linear SVM is typically employed in such a scenario, and the kernel function is used to map the original data into a higher-dimensional space where it becomes linearly separable. The kernel function can be represented as:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \tag{6}$$

where K is the kernel function, x_i and x_j are n-dimensional input vectors representing student data, and ϕ is a mapping function that transforms the data from n-dimensional space to a higher m-dimensional space.

SVM have notable strengths that make them effective for analyzing student data. SVM performs well with small datasets, where its margin maximization principle helps to prevent overfitting, especially when combined with appropriately chosen kernels (Cortes & Vapnik, 1995). Its versatility with custom kernels, such as the radial basis function, enables SVM to capture complex nonlinear patterns in student data (Namoun & Alshanqiti, 2020). However, SVM has some weaknesses. It is computationally inefficient for large datasets, as its complexity scales poorly with the number of samples. SVM requires careful hyperparameter tuning, such as selecting the kernel type and regularization term, which can be computationally expensive and time-consuming.

3.2.5. Model evaluation methods

Five metrics were used to evaluate the models' performance, including accuracy, precision, recall and f1-score for the classification method and Mean Square Error (MSE) for the regression method. They are computed as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

where TP = number of true positive predictions; FP = number of false positive predictions; TN = number of true negative predictions; FN = number of false negative predictions.

MSE is the average squared difference between the estimated and actual values. It is a measure of the quality of an estimator. MSE is always non-negative, and values closer to zero are better. It is computed using:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11}$$

where n is the number of observations, y_i is the actual value for the i-th observation and $\hat{y}_i$ is the predicted value for the i-th observation.

3.3. Experimental setup

The experiment was designed to forecast students' next-term performance recursively and their final performance. Each semester's absolute GPA (regression) and class (classification) were predicted. The study utilized records from 3,093 students with 12 features. To enhance the accuracy of our predictions, base predictors for a given semester (t) are trained using student backpack data accumulated up until the previous semester (t-1). Each set was partitioned into 80% training and 20% testing sets. The experiments were conducted using 5-fold cross-validation to ensure robustness. Additionally, we utilized the GridSearchCV functionality in scikit-learn to fine-tune the hyperparameters and optimize the models' performance. The selected hyperparameters are presented in Table 6.

Figure 4 shows the distributions of students' final performance, indicating the imbalanced nature of the classes, with the red line representing the average score. This unequal representation of the classes was also observed in the semesters' performance. This imbalance is a problem for the classification method. It should be noted that no advanced multiclass data imbalance method was explored in this work. However, cost-sensitive learning with a balanced weight was applied to ensure a balanced distribution of each semester's target class weights. Cost-sensitive learning minimizes the total cost by emphasizing the minority (positive) classes through awarding higher misclassification costs (Ling & Sheng, 2008).

4. Results and discussions

The results have been segmented based on the sequential predictions, starting in semester 2, when the students have covered 12.5% of study time, and continuing through to when they complete. We presented and compared the results based on classification and regression methods.

4.1. Next-term performance

4.1.1. Classification method

For the classification, each model was evaluated with the key metrics: precision, recall, f1-score and accuracy. The results are presented in Table 7, where the top result in each metric is highlighted in **bold**.

Table 6. Hyper parameters settings.

Model	Hyper parameters
RF	n_estimators = 500, max_depth = 20, min_samples_leaf = 1, min_samples_split = 5.
SVM	kernel = rbf, C = 10.0, gamma = scale,.
Bi/LSTM	optimizer = adam, epochs = 150, batch_size = 32, activation functions = ReLU, Softmax regularization = dropout, memory cells = 50.

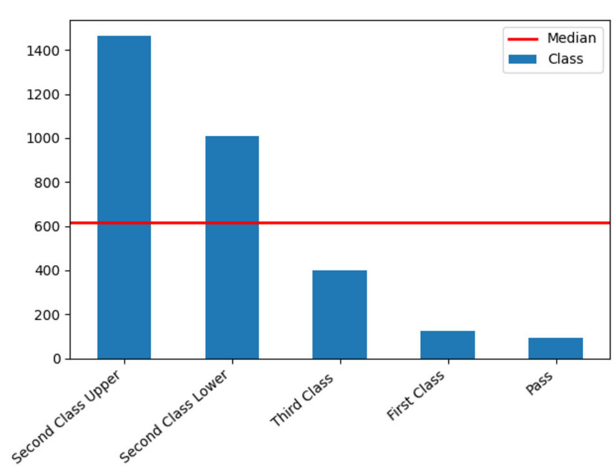


Figure 4. Students' performance distribution.

Table 7. Models' performance in next-term prediction - classification method.

Semester	Algorithm	Precision	Recall	F1-score	Accuracy
2	RF	0.56	0.56	0.52	0.56
	SVM	0.43	0.49	0.44	0.47
	LSTM	0.49	0.45	0.46	0.45
	BiLSTM	**0.57**	**0.58**	**0.56**	**0.58**
3	RF	0.56	0.56	0.54	0.56
	SVM	0.43	0.49	0.44	0.47
	LSTM	0.51	0.47	0.47	0.47
	BiLSTM	**0.56**	**0.57**	**0.54**	**0.57**
4	RF	0.57	0.56	0.54	0.56
	SVM	0.50	0.58	0.52	0.53
	LSTM	0.55	0.53	0.53	0.53
	BiLSTM	**0.58**	**0.58**	**0.57**	**0.58**
5	RF	0.54	0.54	0.52	0.54
	SVM	0.44	0.53	0.46	0.46
	LSTM	0.50	0.47	0.47	0.47
	BiLSTM	**0.54**	**0.55**	**0.53**	**0.55**
6	RF	0.60	0.61	0.58	0.61
	SVM	0.49	0.59	0.52	0.54
	LSTM	0.58	0.53	0.54	0.53
	BiLSTM	**0.61**	**0.62**	**0.60**	**0.62**
7	RF	0.58	0.59	0.57	0.59
	SVM	0.45	0.54	0.47	0.49
	LSTM	0.56	0.51	0.51	0.51
	BiLSTM	**0.60**	**0.61**	**0.59**	**0.61**
8	RF	0.59	0.61	0.58	0.61
	SVM	0.44	0.52	0.46	0.53
	LSTM	0.57	0.51	0.52	0.51
	BiLSTM	**0.60**	**0.62**	**0.60**	**0.62**

The bold values are the top results recorded under each metric semester-by-semester.

SVM achieved the lowest performance across semesters in most metrics, with accuracies consistently below 55%. Notably, SVM's accuracy ranged from 47% in the early semesters (2 and 3), peaking at 53% by Semester 4. However, its performance experienced fluctuations thereafter, dipping to 46% in Semester 5, rebounding to 54% in Semester 6, declining again to 49% in Semester 7 and finally marginally increasing to 53% in Semester 8. Despite this, SVM showed a modest increase in recall from 0.49 in Semester 2 to 0.59 in Semester 6, indicating the systematic improvement in identifying more at-risk students and the potential to capture relevant cases over time.

RF demonstrated consistently high accuracy levels over SVM, with 56% accuracy in Semesters 2 through 4. It then recorded a slight dip to 54% in Semester 5 and peaked at 61% in Semester 6, followed by 59% in Semester 7 and returning to 61% in Semester 8. Though exhibiting a lower performance than BiLSTM, RF's robust handling of feature interactions makes it a strong candidate for less complex datasets with less pronounced temporal dynamics.

LSTM has shown moderate stability in its performance, with accuracy consistently around 53%. Although designed to capture temporal data dependencies, LSTM's effectiveness did not significantly outperform RF in this context. LSTM recorded an accuracy of 45% in Semester 2, increasing gradually to 47% in Semester 3, reaching 53% in Semester 4, maintaining at 47% in Semester 5, rising again to 53% in Semester 6 and stabilizing at 51% in Semesters 7 and 8. It has, however, shown a consistent increase in precision over time, demonstrating its ability to learn the dependencies in student data in identifying the positive classes.

BiLSTM architecture achieved the top performance across all metrics in all semesters. Starting from an accuracy of 58% in Semester 2, it rose to 62% by Semester 8. The BiLSTM processes the student semester-wise performance records by running two separate LSTM networks in forward and backward directions, which are then merged at a common output layer (Schuster & Paliwal, 1997). This dual-directional approach allows the model to capture both past and future context, providing a more comprehensive understanding of the sequential data. We observed that this mechanism is particularly beneficial in capturing subtle temporal patterns, which likely contributes to the stand-out model's improved performance compared to the others. Notable in Semester 8, BiLSTM achieved consistently high values compared to the other models: precision (60%), recall (62%) and F1-score (60%).

It can be observed that while BiLSTM shows better performance, overall models' accuracies are relatively low (less than 70%). This is problematic for sensitive sectors, such as education, where high recall is critical for decision-making. The low performance of the model may be attributed to the huge uneven distribution of instances in the various data categories (El-Deeb et al., 2022). This skewed distribution typically biases most standard ML algorithms toward the majority class, leading to misclassification of examples from minority classes (Vuttipittayamongkol & Elyan, 2020), ultimately resulting in poor model performance. Although cost-sensitive learning was applied to balance the training sample, it seems less effective in the multiclass scenario.

Multiclass imbalance problem requires an unconventional approach because most resampling techniques, including cost-sensitive learning, were designed and tested in binary classification (Lango & Stefanowski, 2022; Li et al., 2020). The imbalance in multiclass datasets can appear in various forms, presenting unique challenges not encountered in binary classification, such as multi-majority and multi-minority (Zhou & Liu, 2005). Therefore, more research is needed to address multiclass imbalance learning in the education sector to improve the accuracy of classification models.

4.1.2. Regression approach

Table 8 presents the MSE scores of the various models.

BiLSTM exhibited superior performance in the early and middle semesters of the academic journey, consistently achieving the lowest MSE values from Semester 2 through Semester 6, recording 0.126, 0.128, 0.104, 0.119 and 0.101, respectively. A slight increase was observed in Semester 7, with an MSE of 0.136, but it improved again to 0.120 in Semester 8. LSTM, while generally showing higher MSE values in the early semesters compared to BiLSTM, recorded the best performance in the latter semesters, achieving the lowest MSEs of 0.123 in Semester 7 and 0.118 in Semester 8. The scores for LSTM were 0.138, 0.139, 0.110, 0.123 and 0.106 from Semesters 2 through 6. The consistent prediction errors of less than 0.15 signified the effectiveness of both BiLSTM and LSTM in forecasting future performance, indicating over 85% confidence.

SVM and RF, on the other hand, exhibited higher MSE values across all semesters. SVM showed gradual improvement from an MSE of 0.319 in Semester 2 to 0.244 in Semester 8, reflecting some progression in model adaptation. RF also showed a steady decrease in MSE from 0.291 in Semester 2 to 0.220 in Semester 8, indicating effective learning and adaptation over time. However, both recorded a high prediction error of over 0.3, showing that their prediction confidence is less than 70% in some semesters.

Consequently, results presented in Figure 5 compared the predicted GPA with the actual GPA for each semester to affirm the efficiency of the various models in forecasting next-term performance. The actual GPA is represented by the dotted line, while the predicted GPA is represented by solid lines. Each color represents a model. The closer the predicted GPA is to the actual GPA, the better the model. Additionally, the lines below the actual GPA line recorded lower prediction errors, while those above the dotted line had higher prediction errors. Hence, the lines below the dotted line performed better. It can be seen that BiLSTM consistently obtained predictions closed to the actual GPAs.

The results showed that BiLSTM is the most effective algorithm in predicting student next-term performance. This proficiency underscores the significance of advanced neural network architectures for educational data analytics, particularly in scenarios demanding a nuanced understanding of students' performance over time.

It can be observed that both classification and regression results showed that each semester's performance is unique, with models showing varying performance across semesters in an unconventional way. For example, in Table 7, the results for semester 3 are lower than those for semester 2, and a similar pattern is observed between semesters 5 and 4. Likewise, in Table 8, semester 2 MSE scores are lower than those of

Table 8. MSE measure of next-term performance prediction.

Semester	BiLSTM	LSTM	SVM	RF
2	**0.126**	0.138	0.319	0.291
3	**0.128**	0.139	0.342	0.315
4	**0.104**	0.110	0.323	0.303
5	**0.119**	0.123	0.308	0.280
6	**0.101**	0.106	0.275	0.257
7	0.136	**0.123**	0.282	0.262
8	0.120	**0.118**	0.244	0.220

The bold values are the top results recorded semester-by-semester.

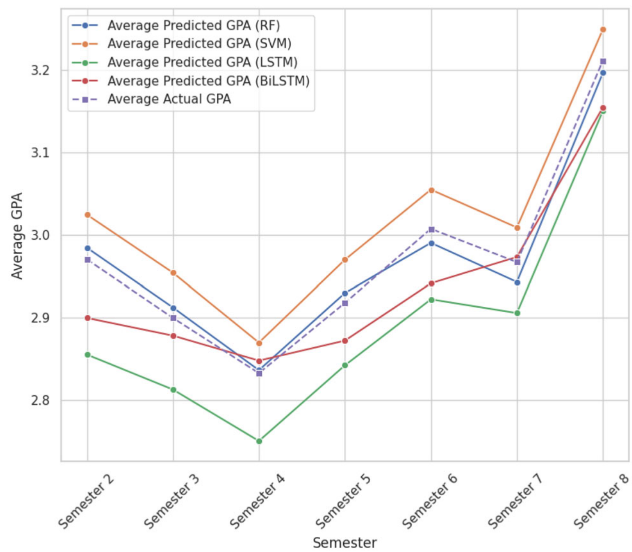


Figure 5. Predicted GPA vs actual GPA.

semester 3, and semester 6 scores are lower than those of semester 7. The observed trend of ML models performing better in the second semester compared to the first semester of every academic year can be attributed to several factors supported by both pedagogical theories and ML concepts.

Studies show that students often require time to adapt to new academic environments, course structures and expectations, particularly after a break (Biggs et al., 2022; Clark & Linn, 2003). This adjustment period during the first semester frequently leads to lower performance compared to the second semester, where familiarity with academic routines, study materials and teaching styles fosters better outcomes (Clark & Linn, 2003). Similarly, Second semester courses often build on the foundations of the first semester, allowing students to apply their knowledge more effectively, leading to better performance (Biggs et al., 2022). Moreover, student motivation and engagement typically improve as the academic year progresses. The second semester benefits from this momentum as students consolidate their efforts to achieve higher grades (Schunk & et al., 2014). For those who underperform in the first semester, the second semester provides an opportunity to intensify their efforts, recover their GPA and meet their academic goals. Behavioral studies highlight that students often experience stress and anxiety at the beginning of an academic year, which can negatively impact performance (Pekrun, 2006).

From an ML perspective, typically, model accuracy improves with more data. The fluctuations in performance can be attributed to concept drift (Žliobaitė et al., 2016), a change in the underlying data distribution over time. Semester GPAs and performance metrics are influenced by factors such as curriculum adjustments, teaching styles, assessment methods and cohort-specific dynamics (Clark & Linn, 2003). As noted by Žliobaitė et al. (2016), such shifts can impact predictive models, resulting in variations in their accuracy across semesters. However, the changes in the relationship between input data and the target variable were not accounted for in this study, requiring further investigation. Similarly, ML issues typically associated with multiclass imbalance datasets, such as model overfitting on specific semester data or sensitivity to outliers or class overlap behavior were not extensively studied in this work (Santos et al., 2022). These factors have the potential to influence the performance of the model.

Comparing the regression and classification methods, the MSE scores obtained in regression significantly outperformed the accuracy levels of the classification method, despite the latter being the most popular approach (Albreiki et al., 2021). These results provide a valuable choice for educational institutions looking to implement predictive analytics for academic performance monitoring.

4.2. FGPA prediction

Each model was used to forecast the FGPA at different semesters. Figure 6 provides a visual representation of the performance of various models (color-coded) in predicting FGPA, illustrating the comparative

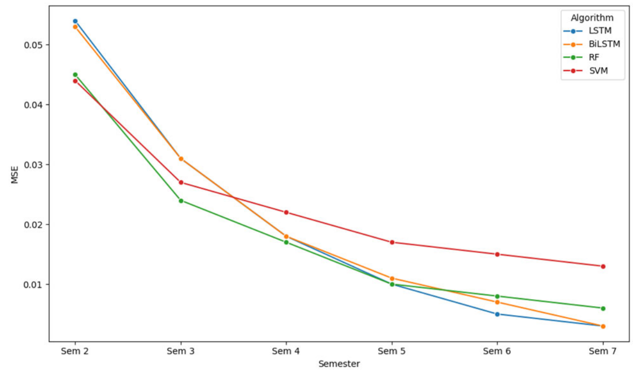


Figure 6. Comparing models performance in predicting FGPA from different semesters.

Table 9. Recall measure for the various performance categories in graduating status prediction.

Semester	Category	RF	SVM	LSTM	BiLSTM
2	First Class	0.30	0.65	0.64	0.59
	Pass	0.29	0.42	0.61	**0.62**
	Second class lower	0.59	0.58	0.40	0.30
	Second class upper	0.80	0.72	0.63	0.62
	Third class	0.48	0.51	**0.58**	0.45
3	First Class	0.52	0.65	0.80	0.70
	Pass	0.42	0.42	**0.73**	0.72
	Second class lower	0.64	0.58	0.49	0.51
	Second class upper	0.86	0.72	0.70	0.68
	Third class	0.53	0.51	**0.62**	0.58
4	First Class	0.70	0.87	0.84	0.65
	Pass	0.67	0.75	**0.85**	0.73
	Second class lower	0.76	0.78	0.58	0.62
	Second class upper	0.88	0.78	0.71	0.73
	Third class	0.61	**0.73**	0.69	0.70
5	First Class	0.65	0.91	0.76	0.77
	Pass	0.71	0.79	**0.89**	0.88
	Second class lower	0.82	0.78	0.65	0.75
	Second class upper	0.92	0.86	0.74	0.78
	Third class	0.74	**0.84**	0.70	0.66
6	First Class	0.78	0.96	0.86	0.84
	Pass	0.83	**0.96**	0.93	0.75
	Second class lower	0.85	0.84	0.70	0.78
	Second class upper	0.92	0.88	0.78	0.78
	Third class	**0.82**	**0.82**	0.81	**0.82**
7	First Class	0.70	0.96	0.85	0.83
	Pass	0.83	0.96	**0.97**	0.94
	Second class lower	0.89	0.88	0.79	0.83
	Second class upper	0.96	0.89	0.83	0.80
	Third class	0.81	**0.86**	0.83	0.83

The bold values are the top recall results for struggling students (Pass and Third class) semester-by-semester.

effectiveness of each model. It can be seen that RF and SVM significantly outperformed the sequence models in the early semesters simply because the predictions were treated as static one-time predictions, signifying their robustness where the temporal dynamic is less. In contrast, though BiLSTM and LSTM performed poorly in the early semesters, they progressively improved to become top-performing in the middle and latter semesters. The graph shows an improved performance of all the models over time, indicating their abilities to learn the evolving nature of students' performance.

Consequently, given the imbalanced nature of our dataset, we assessed the recall (sensitivity) to evaluate how well each model identified all relevant cases of 'Third class' and 'Pass' in forecasting FGPA. High recall values are desirable, as they show the model's ability to detect more struggling students, often a significant oversight in educational settings. Table 9 presents the recall measure of each class. Though RF recorded the lowest MSE in the early semesters, as shown in Figure 6, it performs poorly in identifying students who need support (pass and third-class categories). BiLSTM and LSTM, on the other hand, provided more consistent and higher recall rates across most semesters in identifying the relevant cases. Notably, BiLSTM showed substantial improvement over the semesters, starting at 62% in Semester

2 and peaking at 94% in Semester 6. Similarly, LSTM demonstrated remarkable sensitivity across semesters, especially in the last semester, with a score of 97%.

The findings offer dual benefits to students and higher education managers. For students, these predictions can proactively identify individuals at risk of underperforming or dropping out by the analyzes of the temporal patterns in academic performance. Early detection facilitates timely interventions, such as academic advising, tutoring, personalized learning plans and early-warning systems to support struggling students in improving their learning outcomes. Evidence indicates that early and targeted support significantly enhances retention and graduation rates (Albreiki et al., 2021; Tinto, 2012). Furthermore, the longitudinal AI-driven analyzes offer valuable insights and real-time feedback into long-term academic performance trends, preparing students for challenges beyond university. For instance, by identifying factors predictive of post-graduation success, institutions can incorporate skills into their curricula that enhance employability (Albreiki et al., 2021; Rodríguez-Hernández et al., 2021).

For education managers, these predictive analytics can enable a deeper understanding of how course sequences, instructional methods or extracurricular activities influence student performance (Adnan et al., 2021). For example, students with low GPAs in foundational courses can be advised to retake these courses before progressing further, a strategy supported by evidence suggesting that strong foundational skills contribute to improved long-term outcomes (Biggs et al., 2022). Additional instructional support can be provided for courses with high failure rates, or programs with higher dropout rates can be prioritized for intervention (Guanin-Fajardo et al., 2024). Moreover, predictive models allow universities to allocate resources more effectively by identifying areas of need. Such strategies help to address systemic inequities, particularly in low-resource settings like Ghana, where socioeconomic disparities may impact educational outcomes (Alhassan et al., 2021).

Similarly, AI-based predictions also support institutions in meeting accountability standards by demonstrating measurable improvements in key metrics such as retention, graduation rates, student satisfaction, and equity. These efforts align with the global goals for sustainable development in education (SDG 4) (Ebzeeva & Smirnova, 2023). By leveraging predictive insights, universities can redesign curricula, enhance instructional methods, and allocate resources where they are most needed, ultimately fostering more effective educational practices (Rodríguez-Hernández et al., 2021; Guanin-Fajardo et al., 2024).

4.3. Feature importance

The purpose of determining the importance of features is to select an appropriate subset of features that have a significant impact on student performance. RF model provides a framework to evaluate the significance of various predictors within our dataset (Genuer et al., 2010). We determined the importance of each variable using the mean decrease in impurity, and the results are presented in Table 10. We further performed feature correlations of the selected features with the FGPA and presented the results in Table 11.

The results presented in Table 10 exclude features with an importance score below 1.00%. These features, including location, bursary status, gender, the number of exam attempts, and Is_choice, were identified as the least predictive and had negligible impact on model performance. Notably, despite the

Table 10. Feature importance obtained from the best RF classifier using the mean decrease in impurity.

Variable	Importance scores (%) ↓
Semester_3	20.00
Semester_4	16.49
Semester_6	11.70
Semester_5	11.60
Semester_2	11.54
Semester_7	9.76
Semester_1	7.20
Semester_8	5.62
High school grade	1.53
Program	1.18
Age	1.08

Table 11. Feature correlation with the target.

Variable	Feature correlation with target
Semester_4	0.878
Semester_3	0.876
Semester_5	0.846
Semester_6	0.829
Semester_2	0.819
Semester_1	0.791
Semester_7	0.787
Semester_8	0.669
Program_Nursing	0.087
Program_Physiotherapy	0.051
Program_MedLab	0.047
Program_Speech and Hearing	0.039
Program_Dietetics	0.034
Program_Diagnotic Imaging	0.019
Program_Disease Control	0.006
Program_Health Information	−0.001
Program_Health Promotion	−0.035
Program_Health Nutrition	−0.036
High school grade	−0.038
Program_Midwifery	−0.055
Program_Biomedical Sciences	−0.070
Program_Physician Assistantship	−0.111
Age	−0.162

economic and resource disparities across different regions in Ghana (Osei-Assibey, 2014), the geographical location was found to have minimal influence on academic performance.

Background variables, such as 'Age' (1.08%) and 'Program' (1.18%), demonstrated a moderate impact on student outcomes. High school grades and semester GPAs, however, emerged as the most significant predictors of academic success. Correlation analyzes, as shown in Table 11, revealed varying degrees of association between semester GPAs and the FGPA, with the second and third-year semesters showing the highest correlations, aligning with existing literature (Quelopana et al., 2024). This trend is consistent with the academic structure in Ghana, where students typically begin specializing in their chosen programs during the second year, contributing directly to their FGPA.

Similarly, the results showed that 'Program' exhibited a varied degree of correlation with FGPA. Nursing, physiotherapy and medical laboratory sciences, among others, recorded positive correlations though minimal, while others showed a negative correlation. These differences may be due to inherent characteristics and academic rigor in the different programs confirming with the existing literature (Adnan et al., 2021).

4.4. Limitations of the study

This study's data was sourced exclusively from a single institution in Ghana, which may not represent other educational environments due to unique local characteristics. This could potentially restrict the broader applicability of our findings. Hence, caution should be exercised when generalizing these results to different contexts. Similarly, the dataset contained only demographics and previous academic achievements. Though these are commonly used to measure student knowledge (Albreiki et al., 2021), they may not fully capture the impact of external factors. Therefore, there is a need to add behavioral, psychometric and socioeconomic data to gauge the value addition to a student's education over time.

5. Conclusion

In this paper, four algorithms, namely RF, SVM, LSTM and BiLSTM, were trained to predict students' upcoming performance. The experiments were modeled as both regression and classification problems. The experimental framework utilized student demographic and evolving academic records in degree completion covering a 5-year period. Results clearly showed that the framework has the ability to predict students' future performance. This information could provide educators with on-time information and insights for targeted intervention development to support at-risk students.

The findings showed that BiLSTM is the most effective algorithm in forecasting the student's next-term performance. This demonstrates the significance of advanced neural network architectures over traditional ML models, particularly in scenarios demanding a nuanced understanding of data over time. Additionally, the regression methods exhibited more robustness and effectiveness in predicting future performance, consistently achieving lower error rates than the classification methods. Interestingly, the study found that key features such as semester GPAs and high school grades had strong discriminating powers in the prediction models.

5.1. Future research direction

We recommend that future research explore advanced preprocessing techniques, including noise and overlap treatment, multiclass data balancing strategies, data augmentation and optimization of the model training process to further enhance performance. Additionally, adopting adaptive learning models that account for concept drift could address fluctuations in student performance during model training, improving the robustness of predictions. We also suggest including external features such as students' behavioral, socio-economic and psychometric attributes in future studies, as these features could provide valuable insights into academic performance over time. Furthermore, extending this work to encompass multiple institutions and diverse cultural contexts would enhance the generalizability and applicability of the results. Implementing these strategies could significantly assist institutions in proactively identifying and supporting students at risk of unsatisfactory degree completion.

Note

1. Instructional Methods: Different teaching approaches, such as traditional lectures, active learning, or flipped classrooms, and how they impact student performance over time. Assessment Types: The effects of various forms of assessment (quizzes, projects, exams) on performance dynamics. External Influences: Factors outside the classroom, such as family support, socio-economic status, and extracurricular activities.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

John-Bosco Diekuu http://orcid.org/0000-0001-9862-5962

Data availability statement

The dataset supporting this study's findings is not publicly available and not easily redistributable to researchers due to reasonable privacy, ethical restrictions and the study center's data protection laws. However, an anonymized version of the dataset is available from the corresponding author upon reasonable request and with permission of the host university.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, *9*, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

Ahmed, A., Joorabchi, A., & Hayes, M. J. (2022). On deep learning approaches to automated assessment: Strategies for short answer grading. *CSEDU*, *2*, 85–94.

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552. https://doi.org/10.3390/educsci11090552

Alhassan, R. K., Ayanore, M. A., Diekuu, J.-B., Prempeh, E. B. A., & Donkor, E. S. (2021). Leveraging e-Learning technology to enhance pre-service training for healthcare trainees in Ghana: Evidence from a pilot project and pointers to policy reforms. *BMC Health Services Research*, *21*(1), 1209. https://doi.org/10.1186/s12913-021-07224-3

Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. *IEEE Access*, *11*, 27579–27589. https://doi.org/10.1109/ACCESS.2023.3250702

Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computers & Education*, *53*(3), 563–574. https://doi.org/10.1016/j.compedu.2009.03.013

Arqawi, S. M., Zitawi, E. A., Rabaya, A. H., Abunasser, B. S., & Abu-Naser, S. S. (2022). Predicting university student retention using artificial intelligence. *International Journal of Advanced Computer Science and Applications*, *13*(9), 315–324. https://doi.org/10.14569/IJACSA.2022.0130937

Ashwin, T. S., & Guddeti, R. M. R. (2020). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies*, *25*(2), 1387–1415. https://doi.org/10.1007/s10639-019-10004-6

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Baashar, Y., Hamed, Y., Alkawsi, G., Fernando Capretz, L., Alhussian, H., Alwadain, A., & Al-Amri, R. (2022). Evaluation of postgraduate academic performance using artificial intelligence models. *Alexandria Engineering Journal*, *61*(12), 9867–9878. https://doi.org/10.1016/j.aej.2022.03.021

Biggs, J., Tang, C., & Kennedy, G. (2022). Teaching for quality learning at university 5e. McGraw-Hill Education (UK).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cheng, Y., Nunes, B. P., & Manrique, R. 2022. Not another hardcoded solution to the student dropout prediction problem: A novel approach using genetic algorithms for feature selection. In *Intelligent Tutoring Systems: 18th International Conference, ITS 2022, Bucharest, Romania, June 29–July 1, 2022, Proceedings* (pp. 238–251). Springer.

Clark, D., & Linn, M. C. (2003). Designing for knowledge integration: The impact of instructional time. *Journal of the Learning Sciences*, *12*(4), 451–493. https://doi.org/10.1207/S15327809JLS1204_1

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon*, *6*(6), e04081. https://doi.org/10.1016/j.heliyon.2020.e04081

Deeva, G., De Smedt, J., Saint-Pierre, C., Weber, R., & De Weerdt, J. (2022). Predicting student performance using sequence classification with time-based windows. *Expert Systems with Applications*, *209*, 118182. https://doi.org/10.1016/j.eswa.2022.118182

Delianidi, M., Diamantaras, K., Chrysogonidis, G., & Nikiforidis, V. (2021). Student performance prediction using dynamic neural models. *arXiv preprint arXiv:2106.00524*.

Duong, H. T.-H., Tran, L. T.-M., To, H. Q., & Van Nguyen, K. (2023). Academic performance warning system based on data driven for higher education. *Neural Computing & Applications*, *35*(8), 5819–5837. https://doi.org/10.1007/s00521-022-07997-6

Dzikovska, M., Steinhauser, N., Farrow, E., Moore, J., & Campbell, G. (2014). BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, *24*(3), 284–332. https://doi.org/10.1007/s40593-014-0017-9

Ebzeeva, Y. N., & Smirnova, Y. B. (2023). Contemporary trends in educational policy: UNESCO higher education roadmap. *RUDN Journal of Sociology*, *23*(2), 329–337. https://doi.org/10.22363/2313-2272-2023-23-2-329-337

Eisenberg, D., Golberstein, E., & Hunt, J. B. (2009). Mental health and academic success in college. *The B.E. Journal of Economic Analysis & Policy*, *9*(1). https://doi.org/10.2202/1935-1682.2191

El-Deeb, O. M., Elbadawy, W., & Elzanfaly, D. S. (2022). The effect of imbalanced classes on students' academic performance prediction: An evaluation study. *International Journal of e-Collaboration*, *18*(1), 1–17. https://doi.org/10.4018/IJeC.304373

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*, 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. http://www.deeplearningbook.org.

Graves, A., Mohamed, A-R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). IEEE.

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

GSS. (2024). Ghana statistical service. Retrieved June 29, 2024, from https://www.statsghana.gov.gh/

Guanin-Fajardo, J. H., Guaña-Moya, J., & Casillas, J. (2024). Predicting academic success of college students using machine learning techniques. *Data*, *9*(4), 60. https://doi.org/10.3390/data9040060

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Indicators, OECD. (2023). Education at a glance 2022. https://www.oecd.org/en/publications/education-at-a-glance-2022_3197152b-en.html

Karlos, S., Kostopoulos, G., & Kotsiantis, S. (2020). Predicting and interpreting students' grades in distance higher education through a semi-regression method. *Applied Sciences*, *10*(23), 8413. https://doi.org/10.3390/app10238413

Kukkar, A., Mohana, R., Sharma, A., & Nayyar, A. (2024). *A novel methodology using RNN+ LSTM+ ML for predicting student's academic performance*. Education and Information Technologies.

Lango, M., & Stefanowski, J. (2022). What makes multi-class imbalanced problems difficult? An experimental study. *Expert Systems with Applications*, *199*, 116962. https://doi.org/10.1016/j.eswa.2022.116962

Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, *63*(2), 157–170. https://doi.org/10.1016/j.bushor.2019.10.005

Li, Q., Song, Y., Zhang, J., & Sheng, V. S. (2020). Multiclass imbalanced learning with one-versus-one decomposition and spectral clustering. *Expert Systems with Applications*, *147*, 113152. https://doi.org/10.1016/j.eswa.2019.113152

Ling, C. X., & Sheng, V. S. (2008). Cost-sensitive learning and the class imbalance problem. *Encyclopedia of Machine Learning*, *2011*, 231–235.

Manigandan, E., P., Anispremkoilraj, B. S., Kumar, Shilpa, M., Satre, A., Chauhan ., & C., Jeyaganthan. (2024). An effective BiLSTM-CRF based approach to predict student achievement: An experimental evaluation. In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)* (pp. 779–784). IEEE.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), 237. https://doi.org/10.3390/app11010237

OECD. (2023). Education at a glance: Graduation and entry rates (edition 2022). https://www.oecd-ilibrary.org/content/data/4d543c0d-en

Osei-Assibey, E. (2014). Nature and dynamics of inequalities in Ghana. *Development*, *57*(3-4), 521–530. https://doi.org/10.1057/dev.2015.25

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024). *Predicting university student graduation using academic performance and machine learning: A systematic literature review*. IEEE Access.

Polyzou, A., & Karypis, G. (2018). *Feature extraction for classifying students based on their academic performance*. International Educational Data Mining Society.

Pujianto, U., Prasetya Wibawa, A., Akbar, M. I., et al. (2019). K-nearest neighbor (k-NN) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)* (pp. 83-–88). IEEE.

Quelopana, A., Keith, B., & Pizarro, R. (2024). Predictive modeling of on-time graduation in computing engineering programs: A case study from Northern Chile. *Computer Applications in Engineering Education*, *32*(5), e22767. https://doi.org/10.1002/cae.22767

Riestra-González, M., del Puerto Paule-Ruíz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, *163*, 104108. https://doi.org/10.1016/j.compedu.2020.104108

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, *2*, 100018. https://doi.org/10.1016/j.caeai.2021.100018

Santoni, M. M., Basaruddin, T., & Junus, K. (2023). Convolutional neural network model based students' engagement detection in imbalanced DAiSEE dataset. *International Journal of Advanced Computer Science and Applications*, *14*(3), 617–626. https://doi.org/10.14569/IJACSA.2023.0140371

Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., Soares, C., Wilk, S., & Santos, J. (2022). On the joint-effect of class imbalance and overlap: A critical review. *Artificial Intelligence Review*, *55*(8), 6207–6275. https://doi.org/10.1007/s10462-022-10150-3

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2014). *Motivation in education: Theory, research, and applications*. Pearson Education.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Shehab, A., Elhoseny, M., & Hassanien, A. E. (2016). A hybrid scheme for automated essay grading based on LVQ and NLP techniques. In *2016 12th International Computer Engineering Conference (ICENCO)* (pp. 65–70). IEEE.

Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting student performance in online learning: A multidimensional time-series data analysis approach. *Applied Sciences*, *14*(6), 2522. https://doi.org/10.3390/app14062522

Suleiman, R., & Anane, R. (2022). Institutional data analysis and machine learning prediction of student performance. In *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 1480–1485). IEEE.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In Interspeech, 2012 (pp. 194–197).
Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. In *Machine learning models and algorithms for big data classification: thinking with examples for effective learning* (pp. 207–235). Springer.
Tinto, V. (2012). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago Press.
Trakunphutthirak, R., & Lee, V. C. S. (2022). Application of educational data mining approach for student academic performance prediction using progressive temporal data. *Journal of Educational Computing Research*, *60*(3), 742–776. https://doi.org/10.1177/07356331211048777
Vuttipittayamongkol, P., & Elyan, E. (2020). Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. *Information Sciences*, *509*, 47–70. https://doi.org/10.1016/j.ins.2019.08.062
Waheed, H., Hassan, S.-U., Radi Aljohani, N., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*, 106189. https://doi.org/10.1016/j.chb.2019.106189
Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, *11*(5), 742–753. https://doi.org/10.1109/JSTSP.2017.2692560
Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166–173. https://doi.org/10.1016/j.chb.2019.04.015
Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z
Zhang, W., Wang, Y., & Wang, S. (2022). Predicting academic performance using tree-based machine learning models: A case study of bachelor students in an engineering department in China. *Education and Information Technologies*, *27*(9), 13051–13066. https://doi.org/10.1007/s10639-022-11170-w
Zhou, Z.-H., & Liu, X.-Y. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering18* (1): 63–77. https://doi.org/10.1109/TKDE.2006.17
Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. In Big data analysis: new algorithms for a new society (pp. 91–114). Springer.