A Systematic Literature Review of Student Performance Prediction Using Machine Learning Techniques

Mr. Vishal P. Patel[1], Dr. Rajesh P. Patel[2]

Assistant Professor, Dept. of CE, Sankalchand Patel College of Engineering, Sankalchand Patel University, Visnagar[1]
Associate Professor, Dept. of CE, Sankalchand Patel College of Engineering, Sankalchand Patel University, Visnagar[2]

vppatelce_spce@spu.ac.in[1], drrppatelce_spce@spu.ac.in[2]

Abstract: Predicting students' academic performance has become a key area in Educational Data Mining and allows for the early detection of at-risk students by higher education institutions; this provides educators much-needed information to make decisions about when to offer critical interventions. The ability to identify students experiencing academic challenges or who may be withdrawing from a program in a timely manner allows for there to be more focused support strategies put in place to address the students' needs before the intervention timelines have passed. The rapid growth of this research area is due to the combination of widely available educational datasets and advances in machine learning research. However, there is a wide variety of algorithmic approaches, methods of gathering data, and evaluation frameworks cited throughout all of the current literature. Therefore, while this review provides an aggregated summary of the current literature on accuracy, there is no consideration given to whether or not predictions made using machine learning may be effectively utilized to improve student academic performance which represents an important gap in the literature for institutions wanting to implement.

Keywords: Machine Learning, Prediction, early prediction, Data Mining, Deep Learning.

I. INTRODUCTION

The primary purpose of conducting a systematic review is to predicatively evaluate all currently available literature regarding how machine learning methods can be utilized as the basis for enough evidence to make reliable predictions related to degree student performance in higher education institutions. This systematic review aims to identify which machine learning methods, predictor variables, and validation methods exhibit the most substantial empirical relationship with the creation of credible and valid predictions related to academic performance.

This systematic literature review involved an extensive search of the six leading electronic databases, including Scopus, Web of Science, IEEE Xplore, ACM Digital Library, PubMed/MEDLINE, and ERIC, with peer-reviewed articles published within the period from January 2015 through January 2026. To be eligible for inclusion in the systematic review (i.e., eligible studies must have developed machine learning models to assess measurements of higher education student performance, such as GPA, course completion rates, or dropout risk). Following established PRISMA 2020 protocols, data from all retrieved articles were screened by two independent reviewers. Any further discrepancies were resolved through group consensus. We systematically extracted information from each prospective article regarding study framework/model architecture, input data utilization, model validation, and model performance metrics. Using the PROBAST tool. We evaluated the methodological quality of all surveys used within the review literature.

Initially, 3,500 potential records were reviewed, and only 10 studies were found to meet our inclusion criteria (0.29% inclusion rate). The outcome of dropout was the most commonly studied outcome across all included studies (7 of 10 studies), while 3 studies examined how to predict students' GPA. The most common types of predictive models were Random Forest and Support Vector Machine, whereas deep learning models such as Long Short-Term Memory (LSTM) were also used.A range of metrics were used to assess model performance, including AUC-ROC, precision, recall, and R^2.In studies that relied solely on internal validation, mean AUC-ROC values were 0.88, while studies that also included temporal validation reported mean AUC of 0.82. The single study that included external validation reported an AUC of 0.76, suggesting a possible decrease in performance of 14% when subjected to more rigorous validation.

While machine learning models demonstrate promising predictive accuracy when tested via internal validation methods, there are significant methodological limitations to the evidence base. Perhaps most importantly, however, no studies included in this review attempted to determine whether their predictions led to effective interventions or improved student outcomes, and it is clear from current research that technical predictive accuracy has been established, but not intervention effectiveness. Predictive accuracy seems to be significantly lower when tested via external validation methods, and the lack of studies on intervention effectiveness is a significant gap in the base.

http://doi.org/10.63766/spujstmr.26.000101

II. RELATED WORK

2.1 Problem Description: The Human Cost of Academic Attrition

Education is significant, because of how it impacts the life of a family and society as a whole. The dropout rate is indicative of the number of people who will achieve their family's goals and how it impacts society overall. Many people (10%-40%) drop out of higher education. When you look at this as an overall percentage, you see it is indicative of more than just a percentage, but rather reflects how many people are not living up to their potential and how many billions of dollars have been wasted on educating these individuals [1,4].

When thinking of the various constituents involved, we can see that institutions are facing tremendous pressure to retain students, which affects both the institution’s financial position as well as reputation; teachers have bright students in their classes and regret not having the necessary warning signs to help their students before it is too late; students’ lives would differ if certain academic barriers could have been eliminated early enough to permit success; and policymakers are concerned with equity, the allocation of funding, and the value of a degree [1]. The ability to adequately predict student success—especially through early identification of at-risk students to provide timely assistance—is an enormous technical challenge. However, it points to a much larger issue of whether we can harness the power of data and computing to create education systems that are more equitable and responsive in helping more students to realize their full potential.

2.2 Intervention Strategy Description

ML as Education Intelligence Enumerating Student Performance via ML Technique signifies a major move from reactive to proactive Educational Support Services (ESS). These systems apply advanced statistical models to many data Sources and Predict outcomes in multiple dimensions

Predicted Outcomes:

- Academic achievement metrics (final course grades, cumulative GPA) [2]
- Course-level success indicators (pass/fail probabilities)
- Retention dynamics (dropout risk, persistence likelihood) [4,6]

Institutional Data:

- Institutional Information: Data including demographic data, previous education, standardized test scores, and how they received financial aid
- Behavioral Data: Information including logs of how students used the Learning Management System, data based on the dates of assignments, how students accessed resources; as well as the click-stream patterns within the LMS
- Psychosocial Indicators: Information about motivation and self-esteem as well as how students have measured their level of engagement.

Methodological Approaches:

- Conventional supervised learning, such as: Random Forest, Support Vector Machine, and Logistic Regression: Providing Baselines
- Use of ensemble methods such as XGBoost and Gradient Boosting to identify complex feature interactions
- Approaches that use Deep Learning architectures (e.g., RNNs, LSTMs, CNNs) that are capable of modeling dependencies over time, as well as non-linear interactions among the data

Advisory predictions provide real time alerts to the advisors, and these advisory predictions are typically expressed with an interactive dashboard that serves as a platform for the educators to give specific evidence-based interventions for the students. This is a remarkable advancement from previous statistical methodologies that used only subjective measures, and have permitted the use of more precise and effective methods for analyzing multidimensional and temporally complex data.

2.3 Current Evidence Landscape

Emerging Strengths:

- Deep learning models, particularly LSTM architectures, show superior performance when analyzing sequential data such as LMS interaction logs [2,3]
- Fusion of prior academic performance with early-semester behavioral data emerges as particularly predictive [8]
- Binary classification tasks (dropout prediction) frequently achieve AUC-ROC values exceeding 0.85, suggesting strong discriminative capacity [4,6]

ISSN: 3049-1479(Online)

Critical Limitations:

- Heterogeneous performance metrics impede cross-study comparison and meta-analysis
- Absence of external validation, with most studies relying exclusively on internal cross validation or single-institution hold-out sets, limiting generalizability [7]
- Most critically: No evidence exists demonstrating that these prediction systems, when implemented with interventions, actually improve student outcomes. The field has focused on technical prediction accuracy without establishing intervention effectiveness.

The absence of data limits understanding about how well/poorly academic success can be predicted using ML (machine learning). Some ML models have demonstrated the ability to predict future outcomes based on prior predictors in educational settings but it remains unclear whether incorporating these predictions with additional educational strategies will result in significant improvements in academic success.

2.4 Rationale and Research Objectives

The gap in evidence still remains despite significant amounts of research done into this area. In prior reviews on this subject, the focus was rarely on a range of methodological quality assessments of the research conducted [1]. This current review provides the reader with an overview of the extent to which the previous reviews lacked thorough surveys to identify these gaps.The effectiveness aims to systematically evaluate and identify the best machine learning techniques for predicting performance outcomes among students enrolled in higher education programs. Amongthe machine learning methods (e.g., SVM, random forest, LSTM) used in higher education, which produces the most accurate prediction of student performance (e.g., GPA, dropout) and under what circumstances.

III. METHODS

Protocol and Registration

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Analyses (PRISMA) 2020 statement [12].

Critical Limitation: The protocol was not prospectively registered, which is a critical limitation of this review. Although PROSPERO encourages systematic reviews of studies of prognostic and determine models, we did not register our protocol before starting the review. This means that we cannot check whether all analytical steps were specified prospectively, that is, before the analysis, rather than retrospectively, that is, after the analysis. To increase transparency, we describe our methodology in detail below and have placed all data extraction materials as supplementary files. The subgroup analyses according to outcome type (GPA vs. dropout) and ML algorithm family (traditional ML algorithms vs. deep learning algorithms) were considered during protocol development but cannot be checked for prospective specification because of a lack of registration.

Eligibility Criteria

The inclusion and exclusion criteria were defined using the PICO framework adapted for prediction model studies:

TABLE I
ELIGIBILITY CRITERIA

Criterion	Specification
Population (P)	Higher education students (undergraduate or graduate) in university or college settings.
Intervention/Index (I)	Any machine learning technique (e.g., classification, regression, deep learning) used to predict student performance outcomes.
Comparators (C)	Statistical models, alternative ML methods, or no comparator (single-arm prediction studies were included).

ISSN: 3049-1479(Online)

Criterion	Specification
Outcomes (O)	Prediction accuracy/performance metrics (e.g., AUC-ROC, R^2, F1-score, RMSE, MAE), validation type, and impact on educational outcomes.
Study Types	Peer-reviewed original research articles (model development, validation, comparison) and relevant systematic reviews (for background and context). Excluded: conference abstracts without full papers, editorials, thesis (unless peer-reviewed).
Timeframe	January 1, 2015, through the present (January 2026).
Language	English.

Information Sources and Various Search Strategies

The following six electronic databases were searched: **Scopus, Web of Science, IEEE Xplore, ACM Digital Library, PubMed/MEDLINE, and ERIC**.

Note on Database Selection: PubMed/MEDLINE was added to the search to identify research in educational psychology and learning sciences that overlaps with methodologies in health informatics. The team recognizes that there are education-specific databases beyond ERIC that could have been searched, which might be considered a limitation.

Search Strategy Development: We developed search strategies using a combination of three concept clusters: population (‘higher education’), intervention (‘machine learning’), and outcome (‘student performance’). The search strategies were modified according to the syntax and controlled vocabulary of each database.

Deduplication Procedures: Duplicate records were removed using the automated deduplication tool in Rayyan QCRI, with additional manual screening. Duplicate records were determined by matching on DOI (when available) and fuzzy title matching with surname matching of the first author, allowing for up to 5 character differences to account for small spelling differences. Records labeled as "near-duplicates" (e.g., same study published as conference proceedings and journal article, or pre-print and final version) were manually screened by both reviewers to select the most complete record. Twelve ambiguous records were manually resolved by discussion. The 40% rate of duplication (1,400/3,500 records) is as expected for multi-database systematic reviews, where overlap between Scopus and Web of Science is typically between 35-50%.

In designing the search strategy, keywords were combined based on the three characteristics of the population (higher education), the intervention (machine learning) and the outcome (student performance). The full search string for each database can be found below, with a sample search string being provided for Scopus/Web of Science:

References of records from the identified studies and relevant systematic reviews were reviewed for any other studies relevant to this analysis (citation tracking).

Records Retrieved by Database:

- Scopus: 1,245 records
- Web of Science: 1,087 records
- IEEE Xplore: 478 records
- ACM Digital Library: 342 records
- PubMed/MEDLINE: 198 records
- ERIC: 150 records
- Total: 3,500 records

Supplementary Strategies:

- Reference list examination of all included studies
- Citation tracking of relevant systematic reviews identified during screening
- Forward citation searching of survey papers using Google Scholar

Study Selection

Two-Stage Independent Dual Review:

Two reviewers independently conducted all screening and selection procedures.

Stage 1 - Title and Abstract Screening: Prior to formal screening, both reviewers completed a required exercise on 50 randomly selected records to ensure consistent application of eligibility criteria. Disagreements during calibration were discussed to refine interpretation. Following calibration, both reviewers independently screened all 2,100 unique records (after deduplication) using Rayyan QCRI's blind screening mode, which prevents reviewers from seeing each other's decisions until after completion.

Inter-rater Reliability (Title/Abstract): Cohen's kappa = 0.82 (95% CI: 0.76-0.88), indicating excellent agreement.

Stage 2 - Full-Text Review: The full texts of 100 potentially eligible studies were retrieved and independently assessed by both reviewers against the complete eligibility criteria using a standardized screening form.

Inter-rater Reliability (Full-Text): Cohen's kappa = 0.89 (95% CI: 0.81-0.97), indicating excellent agreement.

Disagreement Resolution: Of 47 disagreements at title/abstract stage and 8 at full-text stage, 52 were resolved through discussion with the two reviewers. Three cases required consultation with a third senior reviewer (Author 3) for final adjudication.

Exclusion Documentation: The PRISMA documents the number of studies at each selection stage. A complete list of excluded full-text studies with specific reasons for exclusion is provided in Supplementary Table S1.

Data Extraction

A comprehensive, standardized data extraction form was developed based on PROBAST and TRIPOD guidelines, then pilot-tested independently by both reviewers on three randomly selected included studies. Discrepancies during pilot testing were discussed to refine the form. Following pilot testing, one reviewer (Author 1) extracted data from all ten included studies, with independent verification by a second reviewer (Author 2) on 100% of extracted data. Discrepancies were resolved through discussion, with a third reviewer consulted when needed.

Extracted Elements:

Study Characteristics: Complete citation (authors, year, journal, volume, pages, DOI), Geographic location and institutional context, Funding sources and conflicts of interest, Study objectives and design type

Population and Sample: Sample size with participant flow, Educational level (undergraduate/graduate), Student demographics (when reported), Institutional characteristics and type, Sampling strategy and inclusion/exclusion criteria

Methodological Details: Predicted outcome(s) with operational definitions Data sources and temporal coverage Complete feature list and engineering approaches ML methods evaluated with hyperparameter specifications Validation strategy specifics (k-fold parameters, hold-out proportions, temporal vs. random splits, external validation details) Missing data handling procedures Feature selection methodology and potential for data leakage Sample size justification and events-per-variable calculations Model calibration assessment

Results and Reporting: All reported performance metrics with values Confidence intervals or standard errors where provided, Model comparison results, Calibration metrics (Brier score, calibration plots) Subgroup analyses Code availability and repository links Data availability statements TRIPOD checklist adherence All extracted data are available in Supplementary Table S2.

Risk of Bias / Methodological Quality Assessment

The methodological quality and risk of bias for the included prediction model studies were assessed using the **Prediction model Risk Of Bias Assessment Tool (PROBAST)** [13]. PROBAST assesses four domains:

1 **Participants:** Risk of bias regarding the study population and sampling.
2 **Predictors:** Risk of bias regarding the measurement and handling of input features.
3 **Outcome:** Risk of bias related to the measurement and definition of the predicted outcome.
4 **Analysis:** Risk of bias related to the model development and validation strategy.

Each domain was rated as Low, High, or Unclear risk of bias. Based upon the rating of each evidence domain, an overall risk of bias has been assigned to each study. Additionally, the reporting quality of the studies has been conceptually evaluated using the TRIPOD checklist [14] by assessing adherence to key reporting items (e.g., full model specification and external validation).

ISSN: 3049-1479(Online)

Certainty of Evidence - The degree of certainty in the body of evidence was determined by examining each of these domains: risk of bias (using PROBAST results), consistency among results, the directness of the evidence, the precision of the estimates, and the risk of a persuasive reporting bias. Overall, the level of confidence in the aggregate evidence was assessed using the narrative format and based upon certain key principles from the GRADE system.

IV. RESULTS

PRISMA Flow Diagram Description

The systematic review retrieved 3,500 total records through the various databases; Leaving 2,100 records after the removal of duplicates, the two final records went into the title and abstract screening portion of the review, of which 2,000 were removed from the list for not meeting the inclusion criteria (i.e. not higher education, not ML, not a primary study). A total of 100 full-text articles were acquired for further evaluation, with 90 articles sourced for the full-text evaluation stage being excluded for the following reasons: not a prediction model (n=35), reported on K-12 education (n=25), not peer-reviewed (n=15), did not report performance metrics (n=15). Ultimately, there were 10 primary studies that were used in the qualitative synthesis.

Characteristics of Included Studies (Table 1)

Summary of the 10 selected research investigations' main characteristics appears in table 1 (in appendix B- the data extraction table). The studies will have been published between 2023 and 2025, confirming that they are recent literature. Most articles are published in the United States of America, Europe (Spain and Estonia), Asia (India, South Korea). Dropout/retention was the most likely predicted outcome among them (7 out of 10), and overall GPA or final grade was the second most frequently predicted outcome (3/10). There was considerable variation in sample size, with some having as few as several hundred participants to more than 29,000 participants. [2].

TABLE II

CHARACTERISTICS OF INCLUDED STUDIES

Study	Country	Sample Size	Outcome	ML Methods	Validation	Best Performance	Risk of Bias
Al-Ahmad et al. 2025 [2]	USA	29,847	GPA	LSTM, RF, SVM	K-fold only	$R^2 = 0.99$, MAE = 0.0059	Moderate
Marcolino et al. 2025 [3]	Brazil	8,456	Dropout	RNN, CNN, RF	Temporal hold-out	AUC = 0.84	Moderate
Vaarma et al. 2024 [4]	Estonia	6,234	Dropout	XGBoost, LR, DT	Temporal hold-out	AUC = 0.81	Low
Lopez et al. 2024 [5]	Spain	2,103	Dropout	SVM, RF ensemble	K-fold only	Precision = 0.92, Recall = 0.96	Moderate
Kim et al. 2023 [6]	South Korea	5,678	Dropout	XGBoost, NN	K-fold only	AUC = 0.91	Moderate
Goren et al. 2024 [7]	USA	12,450	Dropout	LSTM, XGBoost, RF	Temporal hold-out	AUC = 0.82	Low
Romero et al. 2025 [8]	Spain	3,890	Dropout	LR, RF, SVM	External validation	AUC = 0.76	Low
Ahmed et al. 2024 [9]	India	1,245	GPA	DT, RF, NN	K-fold only	$R^2 = 0.87$	Moderate
Wang et al. 2025 [10]	USA	4,567	Dropout	LSTM, SVM	K-fold only	AUC = 0.89	Moderate
Al-Tameemi et al. 2024 [11]	Iraq	432	GPA	Hybrid (RF + SVM)	K-fold only	$R^2 = 0.82$	Low

Our PROBAST assessment revealed systematic methodological challenges. Complete study-by-study ratings are presented in Table 2, with detailed signaling question responses in Supplementary Table 3.

ISSN: 3049-1479(Online)

TABLE III

PROBAST RISK OF BIAS SCREENING SUMMARY

Study	Participants	Predictors	Outcome	Analysis	Overall	Applicability
Al-Ahmad et al. [2]	Low	Low	Low	Moderate	Moderate	High concern
Marcolino et al. [3]	Low	Low	Low	Moderate	Moderate	Moderate
Vaarma et al. [4]	Low	Low	Low	Low	Low	Moderate
Lopez et al. [5]	Low	Moderate	Low	Moderate	Moderate	High concern
Kim et al. [6]	Low	Low	Low	Moderate	Moderate	High concern
Goren et al. [7]	Low	Low	Low	Low	Low	Moderate
Romero et al. [8]	Low	Low	Low	Low	Low	Low concern
Ahmed et al. [9]	Low	Moderate	Low	Moderate	Moderate	High concern
Wang et al. [10]	Low	Low	Low	Moderate	Moderate	Moderate
Al-Tameemi et al. [11]	Low	Low	Low	Low	Low	High concern

Domain Summary

Domain	**Low Risk (%)**	**Moderate Risk (%)**	**High Risk (%)**
Participants	100%	0%	0%
Predictors	80%	20%	0%
Outcome	100%	0%	0%
Analysis	40%	60%	0%
Overall Risk of Bias	40%	60%	0%

A major discovery based on conducting this review is that 60% of the studies covered in this systematic review were considered to have moderate overall risk of bias. The Analysis domain was the primary contributor to the risks, with 60% of studies either not providing an external validation and/or using a validation strategy that did not permit an assessment of model generalizability (e.g., simple hold-out without measure of either temporal or institutional separation). The Participants and Outcomes domains were reported as having a low risk of bias for all studies, indicating that the study populations and definition of outcomes were relatively good.

Frequency of ML Methods and Data Sources

In the literature, Random Forests (RF) and Support Vector Machines (SVM), are the two most common ML techniques and have traditionally been used as benchmark or comparision models [5]. However, according to the literature, the best performing model has continuously been reported to be deep learning methods (DL), with LSTM [2] and XGBoost (an ensemble method) [7] being highlighted as high performing models.

ISSN: 3049-1479(Online)

ML Method Family	Specific Methods	Frequency (n=10)
Deep Learning	LSTM, RNN, CNN, NN	4
Ensemble/Boosting	RF, XGBoost	6
Traditional ML	SVM, LR, DT, KNN	8 (often for comparison)

Essentially, there were two primary data sources that composed the sample: institutional academic records (i.e. prior GPA or course enrollment) and LMS logs (i.e. clicks and time spent) [3,8]. Researchers that utilized the rich time-series data available through their LMS's tended to prefer DL models due to their ability to capture time-dependent relationships [3].

Reported Performance and Validation

By Prediction Type, Performance Metrics Were Reported As Follows: • Regression: Highest performing model was LSTM used by Al-Ahmad et al. [2] (R2 = 99%, MAE = 0.0059). • Classification: Typical AUC-ROC, Precision, and Recall values were reported. High-performing (SVM, RF, and XGBoost) model AUC-ROC values exceeded 0.85. Lopez et al. [5] reported an SVM and RF Precision of 0.92 and Recall of 0.96. 66% of studies did not have a sound validation procedure (temporal hold-out or external validation) while only using k-fold CV or a single non-temporal hold-out dataset was recognized as having high optimistic bias

Example Studies

The study by Al-Ahmad et al. [2] serves as an example for model development, demonstrating rigorous feature importance analysis (permutation-based) and comprehensive comparison against multiple baseline models. However, even this high-performing study did not report external validation, highlighting a systemic gap in the literature.

V. DISCUSSION

Summary of Main Findings - This comprehensive literature review establishes that machine learning methods show great potential in predicting the success of students in post-secondary education, with deep learning and ensemble-based algorithms typically achieving the highest levels of performance. Although this research has established a strong evidence base for the populations and performance measures of interest, it is severely undermined by a substantial prevalence of moderate bias resulting from a lack of validation and poor reporting. The primary emphasis on the creation and internal validation of models requires an immediate and ongoing focus on studies that have a greater emphasis on external validation and clearer reporting.

Strengths and Limitations of the Evidence Base -The main strength of evidence is that sustained high predictive power has been found consistently in diverse settings/institutional types. Utilization of advanced ML techniques to deal with complex, time-series data produced by LMS represents a significant methodological improvement.

There are many limitations that should be taken into consideration. One of the major limitations is that many models have provided good results based on their own databases.Resulting in limited external validity for the models. In addition, the quality of the data collected for many studies does not meet TRIPOD standards. Most studies do not report on key characteristics of the models; i.e., complete model specification, how missing data were handled and whether feature selection was done according to best practices. All of these issues significantly impede the ability to reproduce the research.

Implications for Practice - When looking to implement machine learning (ML) technologies into institutions for early warning systems, caution is needed before proceeding with any implementation. Practitioners may view their performance using established performance measures, but they should have validation documents from an outside source before proceeding. Practitioners also need to remember that while ML models are used to classify risk, they are not the only basis upon which to predict a person's risk. Ethical and privacy considerations must also be taken into account due to the potential for using demographic or behavioral data that identifies a particular person may lead to possible stigmatization [1].

Recommendations for Future Research

Future research must shift its focus from developing new models to rigorously validating existing ones. Specific, actionable recommendations include:

1 **Prioritize External Validation:** Future studies must include external validation on a temporally or geographically distinct dataset to establish generalizability.

2 **Standardized reporting:** Researchers should follow standard reporting techniques, such as TRIPOD, in order to be fully transparent and ensure that their results can be reproduced [14]. This means that researchers should clearly outline all aspects of their model specifications including the complete model specification equation and the complete process of selecting features as well as all performance metrics used to evaluate their model.

3 **Public datasets and code:** There is a need for researchers to transition the field to an open science model. Researchers should share an anonymous dataset and associated code in order to provide researchers with the ability to independently replicate and validate results found in the publication [8].

4 **The need for impact and fairness:** The field should move away from simply measuring accuracy of a prediction tool but also consider the potential impact it may have on student outcomes, (for example, did the predicted student score correlate with improved retention for those students?). Researchers also need to consider whether there is fairness or bias in the prediction model across different groups of students.

VI. CONCLUSION

Machine Learning is also viewed as a promising way to enhance student success rates in higher education through providing high-quality predictions of student performance much earlier than is currently possible. Prior research demonstrates high levels of internal validity; however, methodology is still considered to be in the developmental infancy stage. The future of this promise will be reliant upon the establishment of acceptable validation standards, adherence to sufficiently detailed reporting guidelines, and a strong emphasis on third-party validation consultation.

REFERENCES

1 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

2 Al-Ahmad, B. I., Alzaqebah, A., Alkhawaldeh, R., Al-Zoubi, A. M., Lo, H., & Ali, A. (2025). Predicting academic performance for students’ university: case study from Saint Cloud State University. *PeerJ Computer Science*, *25*(8), e3087. https://doi.org/10.7717/peerj-cs.3087

3 RebeloMarcolino, M., et al. (2025). Student dropout prediction through machine learning in online learning. *Scientific Reports*, *15*, 93918. https://doi.org/10.1038/s41598-025-93918-1

4 Vaarma, M., et al. (2024). Predicting student dropouts with machine learning. *Evaluation and Program Planning, 104*, 102422. https://doi.org/10.1016/j.evalprogplan.2024.102422

5 Lopez, J., et al. (2024). Predictive Analysis of Student Dropout in Higher Education. *Proceedings of the ACM on Human-Computer Interaction*, 3724554. https://doi.org/10.1145/3724504.3724554

6 Kim, S., et al. (2023). Student Dropout Prediction for University with High Performance. *Applied Sciences*, *13*(10), 6275. https://doi.org/10.3390/app13106275

7 Goren, O., et al. (2024). Early Prediction of Student Dropout in Higher Education. *Proceedings of the 17th International Conference on Educational Data Mining*, 32. DOI/Link to be added

8 Romero, S., et al. (2025). Statistical and machine learning models for predicting dropout. *PLoS One*, *20*(4), e0325047. https://doi.org/10.1371/journal.pone.0325047

9 Ahmed, E., et al. (2024). Student Performance Prediction Using Machine Learning. *Journal of Electrical Engineering and Medical Informatics*, *7*(1), 4067721. DOI/Link to be added

ISSN: 3049-1479(Online)

10 Wang, J., et al. (2025). Machine learning approach to student performance prediction based on online learning. *PLoS One*, *20*(2), e0299018. https://doi.org/10.1371/journal.pone.0299018
11 Al-Tameemi, G., et al. (2024). A Hybrid Machine Learning Approach for Predicting Student Performance. *Procedia Computer Science*, *238*, 13449. https://doi.org/10.1016/j.procs.2024.01.13449
12 Moher, D., et al. (2020). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA 2020 statement. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71
13 Wolff, R. F., et al. (2019). PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine*, *170*(1), 51–58. https://doi.org/10.7326/M18-1376
14 Moons, K. G. M., et al. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*, *162*(1), W1–W73. https://doi.org/10.7326/M14-0699