Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: https://www.science-gate.com/IJAAS.html

Integrating machine learning practicality into learning analytics: A framework for reproducible and actionable dropout prediction models

Sabine Berger [1], Abeer Alsadoon [1, 2], Oday D. Jerew [1, 2], Ahmed Hamza Osman [3, *], Albaraa Abuobieda [4], Abubakar Elsafi [5], Azhari Qismallah [6]

[1]*Higher Education Leadership Institute (HELI), Melbourne, Australia*
[2]*Asia Pacific International College (APIC), Sydney, Australia*
[3]*Department of Information Systems, Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah, Saudi Arabia*
[4]*Department of Computer Science, University of Tabuk, Tabuk, Saudi Arabia*
[5]*College of Computer Science and Engineering, Department of Software Engineering, University of Hafr Al Batin, Hafr Al Batin, Saudi Arabia*
[6]*Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia*

ARTICLE INFO

Article history:
Received 10 October 2025
Received in revised form
20 March 2026
Accepted 27 March 2026

Keywords:
Learning analytics
Dropout prediction
Model practicality
Interpretability
Computational feasibility

ABSTRACT

Interest in learning analytics for predicting student dropout has increased in recent years; however, a clear gap remains between research findings and their practical implementation in educational settings. This study systematically reviews 34 recent practical studies on learning analytics and dropout prediction to identify key gaps that limit the transfer of research into practice. Based on these gaps, we propose a novel six-layer framework that integrates research design, model optimization, and deployment considerations, providing a structured approach to conducting practical learning analytics research. The framework addresses critical issues, including reproducibility, generalizability, interpretability, actionability, and computational feasibility. We map the existing literature onto this framework using structured evaluation tables and find that most studies lack comprehensive attention to model practicality. Our framework contributes to the field by integrating theoretical and operational considerations at the research stage, thereby helping to bridge the gap between published research and real-world application. Furthermore, we introduce reproduction studies as a mechanism to promote innovation, particularly in improving model interpretability, generalizability, and actionability. Finally, we recommend adopting training time as a standard evaluation metric to strengthen the focus on practical feasibility in future research.

1. Introduction

Like many other industries, the higher education sector has increasingly adopted standard data collection and storage practices. These data support decision-making by improving understanding of underlying processes, benefitting both institutions and students (Sailer et al., 2024; Topali et al., 2025). This work falls under the field of learning analytics (LA), which links data collection, analysis, interpretation, and use within a closed-loop process informed by educational and psychological theories (Sailer et al., 2024). When examining student dropout, a foundational theoretical lens is provided by Tinto's theory of student dropout (Tinto, 1975), which outlines how students' experiences and circumstances throughout their studies shape their decision to complete or withdraw. This theory underpins LA research on dropout by guiding the identification of common factors associated with student attrition and enabling the prediction of future behavior. While LA aims to inform teaching and student support through analytic insights, educational data mining (EDM) focuses on extracting patterns from data (Barbeiro et al., 2024). Classification and regression algorithms are frequently applied to historical datasets to uncover predictors of student outcomes and apply these insights to new cohorts (Nabil et al., 2021; Barbeiro et al., 2024). In essence, applying Tinto's theory to

* Corresponding Author.
Email Address: ahoahmad@kau.edu.sa (A. H. Osman)
https://doi.org/10.21833/ijaas.2026.04.003
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-8512-578X

EDM and LA suggests that machine learning (ML) models can be used to predict a student's likelihood of dropping out based on established behavioral and contextual trends.

The use of EDM using ML models has grown rapidly within LA research (Fahd et al., 2022; Topali et al., 2025), with increasing attention directed toward predicting student dropout and academic performance. However, a clear divide remains between theoretical research and practical implementation. Much of the existing work prioritizes maximizing predictive accuracy rather than evaluating whether models are suitable for real-world use. As LA is intended to enhance the learning experience (Siemens, 2013), accuracy alone is insufficient; computational feasibility, interpretability, and actionability are also essential for effective deployment. Neglecting these factors often results in ML models that perform well in controlled research settings but fail in practice. For example, highly complex models may require extensive computational resources and long training times (Ortigosa et al., 2019), while unclear or non-actionable outputs can hinder the ability of support staff to respond effectively. This discrepancy between research and practice remains a significant barrier to wider adoption of predictive models in higher education and warrants closer examination.

This study evaluates recent literature on student dropout prediction using a structured, implementation-focused lens. In particular, it examines the practicality of proposed models by assessing their reproducibility, generalizability, interpretability, and actionability, as well as their computational feasibility. Each reviewed study is summarized in structured tables that highlight the extent to which these criteria are addressed. Insights from the review inform the development of a novel six-layered framework that guides a practically oriented research approach. The framework illustrates how data-centered EDM processes connect with the practicality of the final model, underscoring the need to integrate these considerations throughout the entire research workflow. The study's contributions are threefold: (1) It demonstrates the importance of embedding practicality criteria at every stage of model development and identifies current shortcomings; (2) it proposes a novel six-layer framework to support the systematic inclusion of these factors in EDM and LA research; and (3) it outlines required actions through recommendations and an applied operational framework.

The paper is organized as follows: Section 2 outlines the methodology for the literature review and classification. Section 3 presents the critical gaps identified across reproducibility, generalizability, interpretability, actionability, and computational efficiency. Section 4 introduces and explains the proposed layered framework. Section 5 discusses the findings and maps the reviewed literature onto the framework. Section 6 provides recommendations for future research, and Section 7 summarizes the study's conclusions and implications for practice.

2. Systematic review methodology and evaluation criteria

The systematic review was conducted using searches in Google Scholar and ERIC databases. Of the 183 screened records, 49 studies met the inclusion criteria, with 34 selected for the evaluation tables and literature mapping. The selection process is illustrated in Fig. 1. Most included studies were published between 2020 and 2025, with a small number of earlier works incorporated due to their distinctive contributions. Fig. 2 presents the publication-year distribution of the reviewed articles, while Fig. 3 shows their categorization based on the SCImago Journal and Country Rank (SJR). The review primarily targeted journals ranked Q1 or Q2, with scientific reports and conference papers noted separately. Two scientific reports and one conference paper were included because they document real-world implementations of ML models for predicting student academic progress in higher education.

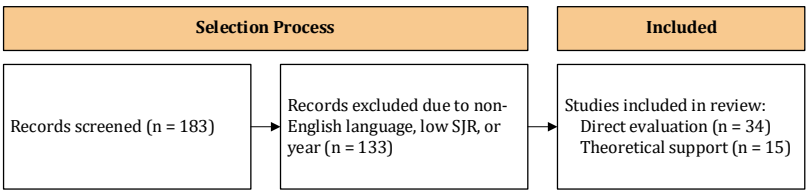

Fig. 1: Process of screening and selecting studies

Search keywords included: "Student dropout prediction," "early warning systems in higher education," "predicting at-risk students," "predicting student academic performance," "machine learning in higher education," "learning analytics for student dropout prediction," "educational data mining," "reproducing learning analytics model," "explainable learning analytics," "interpretable learning analytics," "implementation of learning analytics," and "implementation of student dropout prediction." Only studies in the English language using educational datasets and applying ML methods for dropout or academic performance prediction were selected. After determining the research aims, four research questions were formulated to structure the review and guide development of the proposed layered framework:

- RQ1: What reproducibility information is provided in the reviewed studies?

- RQ2: How do the studies assess the generalizability of their models?
- RQ3: To what extent do the studies consider model interpretability and actionability of the output?
- RQ4: How many studies report computational demand when evaluating model performance?

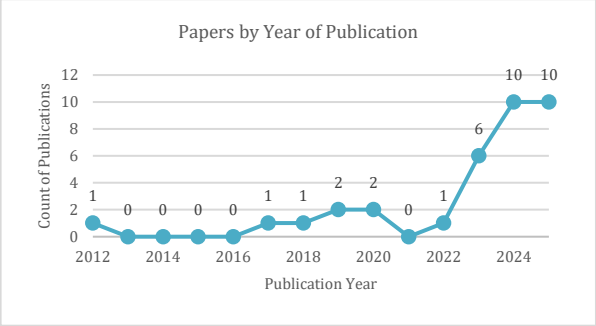

Fig. 2: Distribution of our reviewed papers by their year of publication

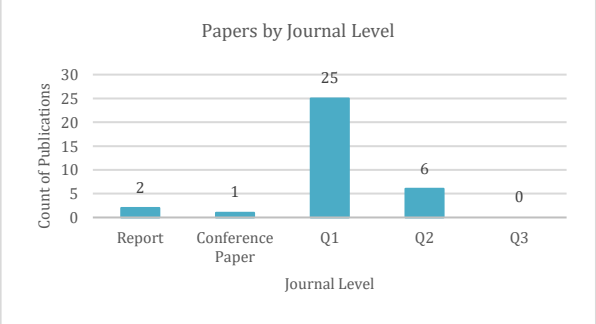

Fig. 3: Distribution of our reviewed papers by their SCImago SJR rating

Each reviewed study was evaluated for its reproducibility, interpretability, and actionability, generalizability, and inclusion of computational performance measures. For the literature review, the studies were grouped according to their primary focus or contribution. However, all were assessed across every criterion in the evaluation tables as follows:

- Reproducibility: Each study was assessed on the level of methodological and mathematical detail provided, as well as the availability of datasets and code. Table 1 summarizes this assessment, categorizing studies as fully, partially, or not reproducible. Full reproducibility was assigned only when the methodology was described in sufficient detail, the dataset was publicly available or deposited in an online repository, and the final code was accessible through a public repository or freely available supplementary material. Partial reproducibility was attributed to studies that provided enough methodological detail to allow independent reproduction of the model, often including hyperparameters or pseudocode. This classification draws on Raghupathi et al. (2022), who assessed reproducibility across method, data, and experiment. Their "method," representing minimal methodological rigor, is classified here as non-reproducible, while their "data" and "experiment" categories align with partial and full reproducibility, respectively. Dataset and code availability were evaluated in the same manner, with full availability requiring open access, while partial availability included limitations such as code available only on request or provision of pseudocode or partial scripts.
- Generalizability: A model was considered fully generalizable when its performance was tested across different contexts, either through multiple datasets or by examining various combinations of feature types within the same dataset. Partial generalizability was attributed to studies that tested models using only widely available features or assessed scalability in a limited way.
- Interpretability and actionability: Interpretability and actionability were assessed independently. Interpretability was defined as the degree to which a study provided insight into the model's decision-making processes. Full consideration was identified in studies using white-box models, either directly or as surrogate models, or employing interpretability tools such as explainable artificial intelligence (XAI). Studies relying solely on statistical analyses for feature impact or correlations were judged to have partial interpretability. Actionability was defined as the extent to which outputs directly support decision-making. Full actionability was assigned to studies producing visualizations, dashboards, or automated alerts. Activities requiring manual interpretation or advanced technical knowledge, such as rule extraction or feature-impact analysis, were treated as partial actionability.
- Computational feasibility: Studies that explicitly reported training time, inference time, or prediction time were classified as fully addressing computational feasibility. Partial consideration was assigned when adjustments were made to reduce computational load, even if the effects were not formally measured.
- Mapping literature to framework: Finally, we assessed overall operational feasibility using the results from all prior criteria and our proposed layered framework. Reproducibility enables independent validation and extension of proposed models; interpretability supports fairness and ethical deployment; actionability ensures that outputs meaningfully inform staff responses and institutional decision-making; generalizability allows a model to function across diverse providers and educational contexts; and excessive computational demand limits real-world uptake due to system constraints and slower processing of updated information.

3. Critical gaps in learning analytics literature

3.1. Reproducibility

Previous research has highlighted a persistent trend of withholding data and code associated with dropout prediction models. This lack of transparency

has contributed to the development of highly specialized models tailored to individual providers, with limited opportunities for validation or extension. Table 1 summarizes the disadvantages of this practice from an implementation-oriented perspective and outlines our findings across the reviewed studies. The literature examined here underscores the broader reproducibility problem and demonstrates the value of reproduction and benefaction studies.

Reproducibility challenges are well documented across multiple disciplines, with information systems research experiencing particularly low reproducibility rates (Raghupathi et al., 2022). Raghupathi et al. (2022) classified reproducibility across method, data, and experiment, reporting that only 30% of studies were fully reproducible. Similarly, Gundersen (2021) proposed a hierarchy of description, code, data, and experiment, emphasizing that verification through reaching similar conclusions rather than replicating exact results plays a critical role in advancing research. This is illustrated by Zhidkikh et al. (2024), who applied a model originally developed by Van Petegem et al. (2023) to a different institutional dataset. By leveraging an existing approach, they were able to extend the model through the addition of self-report features and a larger dataset, demonstrating the efficiency and innovation supported by reproducible work. Collberg and Proebsting (2016) further highlighted these issues through attempts to retrieve the source code from reviewed studies using online searches and direct author contact. Despite these efforts, approximately 57% of the studies exhibited no reproducibility, with the remainder demonstrating only limited reproducibility. Nevertheless, their findings point to the value of partial reproducibility for benefaction, which is defined as the ability to build upon prior work when sufficient methodological detail is available. Even when experimental code is private, detailed methodological descriptions, hyperparameter information, pseudocode, and accessible datasets can substantially improve reproducibility. Studies by Huang et al. (2022), Latif et al. (2023), Masood et al. (2024), Mustofa et al. (2025), Roy and Farid (2024), Tong and Li (2025), Wen and Juan (2023), and Zhang et al. (2025) fall into this category, as shown in Table 1. Conversely, an ongoing emphasis on novelty can impede progress, with Molla-Esparza et al. (2025) noting that repeated reinvention limits cumulative knowledge building. Our findings suggest that partially or fully reproducible studies play a crucial role in driving innovation by enabling researchers to reliably validate, compare, and extend existing models rather than starting from scratch.

Table 1: Overview of reviewed learning analytics studies' reproducibility assessment

Study	Methodology	Dataset availability	Code availability	Reproducibility level	Context transferability
Ortigosa et al. (2019)	✗	✗	✗	✗	✗
Adejo and Connolly (2018)	✗	✗	✗	✗	✗
Wong et al. (2025)	✗	✗	✗	✗	✗
Nagy and Molontay (2024)	✗	✗	✗	✗	≈
Mosia (2025)	≈	✗	✗	✗	≈
Hoca and Dimililer (2025)	✗	✗	✗	✗	≈
Alamuddin et al. (2019)	✗	✗	✗	✗	✗
Arnold and Pistilli (2012)	✗	✗	✗	✗	✗
Du et al. (2020)	✗	✗	✗	✗	✗
Maniyan et al. (2024)	✗	✗	✗	✗	≈
Zanellati et al. (2024)	✗	✗	✗	✗	✗
Rebelo Marcolino et al. (2025)	✓	✗	≈	≈	≈
Delen et al. (2024)	✓	✗	✗	≈	✗
Pek et al. (2023)	✓	≈	≈	≈	✗
Olaya et al. (2020)	✓	✗	✗	≈	✗
Skittou et al. (2024)	≈	✗	≈	≈	✗
Sonnleitner et al. (2025)	≈	≈	✗	≈	✗
Rabelo and Zárate (2025)	✓	✗	✗	≈	✗
Zhang et al. (2025)	✓	✓	✗	≈	≈
Vives et al. (2024)	✓	≈	≈	≈	✗
Van Petegem et al. (2023)	✓	✗	✗	≈	✗
Zhidkikh et al. (2024)	✓	✗	✗	≈	✗
Wen and Juan (2023)	✓	✓	≈	≈	≈
Hoffait and Schyns (2017)	≈	✗	≈	≈	≈
Romero and Liao (2025)	✓	✓	✗	≈	✗
Huang et al. (2022)	≈	✓	≈	≈	✗
Latif et al. (2023)	≈	✓	✗	≈	✗
Masood et al. (2024)	✓	✓	≈	≈	≈
Mustofa et al. (2025)	≈	✓	≈	≈	≈
Roy and Farid (2024)	≈	✓	≈	≈	✓
Pan et al. (2024)	✓	✓	✓	✓	✓
Tong and Li (2025)	✓	✓	≈	≈	✗
Matz et al. (2023)	✓	✓	✓	✓	✓

Reproducibility level: Dataset and full or partial code available (✓), detailed methodology provided (≈), no reproducibility (✗); Methodology: Detailed methodology (✓), mathematical approach only or less detailed methodology (≈), no details (✗); Dataset available: Public dataset or online repository linked (✓), upon request only (≈), not available (✗); Code available: Online repository or supplementary material (✓), upon request only, pseudocode only, partial code (≈), not available (✗); Context transferability: Assessed performance on multiple datasets (✓), considerations for generalizability made (≈), not considered (✗)

3.2. Generalizability

Developing models within isolated contexts also restricts their practical applicability. Evaluating models across datasets that differ in scale and composition is essential for identifying algorithmic limitations and assessing real-world feasibility (Roy and Farid, 2024). For example, in an effort to improve accuracy while reducing training time, Roy and Farid (2024) introduced an Adaptive Feature Selection Algorithm (AFSA), which achieved performance comparable to Forward Selection (FS). When tested on four distinct datasets, however, AFSA exhibited substantial variability: its accuracy fell 12.2% below FS on the SSP dataset, but only 0.8% on the significantly smaller WOC2 dataset. A broader perspective is offered by Matz et al. (2023), who compared the performance of Elastic Net and a Random Forest (RF) classifier using data from three large universities and an additional combined dataset from 16 community colleges. Their findings showed RF outperforming Elastic Net on Universities 1 and 4 by 11% and 12% in F1-score and by 0.09 and 0.12 in AUC, whereas improvements for Universities 2 and 3 were minimal (1% and 0% F1-score, and 0.05 and 0.01 AUC). They also examined the effect of feature types, noting that including all available features generally improved performance.

Our classification of studies based on their consideration for context transferability (Table 1) shows that few works evaluate their models using more than one dataset (Matz et al., 2023; Pan et al., 2024; Roy and Farid, 2024). Similarly, Adejo and Connolly (2018) investigated the impact of different data types by constructing multiple feature compositions from a single dataset. They found that online learning variables consistently produced the largest performance gains across models, while survey data had mixed effects. Collectively, these results underscore the importance of generalizability testing for identifying model boundaries and ensuring appropriate evaluation. Despite this, our analysis of data sources and feature types (Table 2) indicates that such testing remains uncommon in the reviewed literature.

3.3. Interpretability and actionability

Predictive modelling for student dropout is fundamentally tied to improving retention through timely and targeted intervention. Although studies vary in how they define the model output, ranging from binary dropout classification to estimates of academic performance or engagement (Sghir et al., 2023), these outputs alone rarely provide sufficient guidance for student support teams. Staff need more than a risk score; they need clarity about why a model produced a particular prediction and how the information can be applied in practice (Ramaswami et al., 2023). The increasing use of black-box models, whose internal logic is inaccessible or opaque (Loyola-González, 2019), heightens this challenge. While "white-box" models, such as decision trees (DT), can visualize their decision-making processes and often perform well on simpler datasets, black-box models such as artificial neural networks (ANN) or ensemble models tend to handle complex datasets more effectively (Loyola-González, 2019). For example, Roy and Farid (2024) reported marginally better performance using DT compared to black-box models when only demographic and academic, and only basic behavioral features are included. In contrast, when using complex online learning logs, Huang et al. (2022) achieved the highest accuracies with RF and long-short term memory (LSTM) models. Similarly, Adejo and Connolly (2018) compared DT, ANN, and support vector machine (SVM) models individually and as ensembles, finding that the ensemble achieved higher precision, recall, and F-score across various data combinations. Pek et al. (2023) evaluated logistic regression (LR), k-nearest neighbor (KNN), DT, SVM, Naïve Bayes (NB), RF, and AdaBoost individually and as a stacked ensemble, reporting a 0.4% increase in mean accuracy for the ensemble over LR. Rabelo and Zárate (2025) assessed LR, classification and regression trees (CART), and a feed-forward multilayer perceptron (FFMLP), with LR achieving 0.7% higher accuracy than the MLP, and 1.3% higher than the best CART result, while their ensemble exceeded all individual models by 1.8%. Overall, while black-box models frequently deliver improved performance on complex datasets, this comes at the cost of interpretability and limits the actionability of their outputs.

The literature suggests several approaches for enhancing the interpretability of black-box models. One strategy is to use white-box models as surrogates that receive the outputs of the primary model and retrace the steps leading to the same prediction (Du et al., 2020; Hoca and Dimililer, 2025). Other studies examine feature contributions through methods such as least absolute shrinkage and selection operator (LASSO) regression (Rabelo and Zárate, 2025), stepwise selection (Rabelo and Zárate, 2025), t-distributed stochastic neighboring ensemble (t-SNE) (Du et al., 2020), manual feature selection (Skittou et al., 2024), permutation importance (Nagy and Molontay, 2024), partial dependence plots (Nagy and Molontay, 2024), and various XAI tools (Mustofa et al., 2025; Nagy and Molontay, 2024; Tong and Li, 2025; Zanellati et al., 2024).

Within our proposed layered framework, we identify opportunities to strengthen interpretability and actionability at the algorithm-selection layer by adopting an output- and practicality-focused perspective. XAI tools such as SHapley Additive exPlanations (SHAP) or local interpretable model-agnostic explanations (LIME) allow for post-prediction investigation of feature contributions. Implemented at the third layer of the framework, these tools enhance the actionability of model outputs and provide opportunities to detect biases or inaccuracies. For example, Zanellati et al. (2024) found that students predicted to drop out were often

transferring to another course after retaking and passing their entry examination, highlighting how explanation techniques can reveal underlying behaviors that inform more appropriate intervention strategies. This type of insight is crucial for academic and student support staff, as it enables them to identify root causes and respond accordingly.

Table 2: Practical constraints in reviewed studies-data source classification

Study	Features	Data source	Year	Instances	Type
Zhang et al. (2025)	❶ ❹ ❸ ❺	SPD24	2024	97,000	Public
Wong et al. (2025)	❶ ❹ ❸	Institutional database	2025	1,591	Private
Tong and Li (2025)	❶ ❹ ❸	XuetangX	2024	59,581	Public
Mosia (2025)	❹ ❺	Institutional database	n.d.	517	Private
Hoca and Dimililer (2025)	❶ ❹ ❺	Institutional registration system	2015–2020	20,974	Private
Adejo and Connolly (2018)	❶ ❸ ❺	Institutional database + survey	2016	141	Private
Nagy and Molontay (2024)	❶ ❹	Institutional database	2013–2017	8,508	Private
Alamuddin et al. (2019)	❶ ❹ ❺	Institutional LMS	2016	10,489	Private
Arnold and Pistilli (2012)	❶ ❹ ❸	Institutional LMS	2007–2009	7,170	Private
Vives et al. (2024)	❹ ❺	Institutional LMS	2020–2022	661	Private
Du et al. (2020)	❹ ❸	Institutional database	2014–2016	11,688	Private
Hoffait and Schyns (2017)	❶ ❹ ❻	Institutional database	2011–2014	6,845	Private
Huang et al. (2022)	❸	OULAD (Kuzilek et al., 2017)	2014	341	Public
Latif et al. (2023)	❸	DEEDS (Vahdat et al., 2015)	2015	115	Public
Maniyan et al. (2024)	❶ ❹	Institutional database	n.d.	648	Private
Masood et al. (2024)	❶ ❹ ❺ ❸	OULAD (Kuzilek et al., 2017)	2014	32,593	Public
Mustofa et al. (2025)	❶ ❹ ❻ ❺	Predict students' dropout and academic success	2021	4,424	Public
Olaya et al. (2020)	❶ ❹	Institutional database	2012–2016	3,362	Private
Ortigosa et al. (2019)	❶ ❹ ❸	Institutional LMS	2016–2017	11,000	Private
Pan et al. (2024)	Not defined	XuetangX KDDCup2015	2023 2023	23,839 72,395	Public Public
Romero and Liao (2025)	❶ ❹ ❻ ❺	Predict Students' Dropout and Academic Success	2008/2009 + 2018/2019	3,400	Public
Rabelo and Zárate (2025)	❶ ❹ ❻ ❺	Institutional database + survey	2018–2019	40,000	Private
Pek et al. (2023)	❶ ❹ ❺	Institutional database, questionnaire	ND	555	Private
Roy and Farid (2024)	❶ ❹ ❸	XAPI SSP (Cortez and Silva, 2008) HESP WOC2	2016 2008 2019 2020	480 1,044 145 486	Public Public Public Public
Skittou et al. (2024)	❶ ❹ ❻	Institutional database, LMS, national database	2020	125,354	Private
Sonnleitner et al. (2025)	❹ ❸	Institutional LMS, questionnaire	2020–2023	382	Private
Van Petegem et al. (2023)	❹ ❸	Institutional LMS	2016–2018	2080	Private
Wen and Juan (2023)	❶ ❹	OULAD (Kuzilek et al., 2017)	2014	32,593	Public
Zanellati et al. (2024)	❶ ❹	Institutional database, LMS	2018–2021	44,875	Private
Zhidkikh et al. (2024)	❹ ❸	Institutional LMS, questionnaire	2015–2021	2,615	Private
Delen et al. (2024)	❶ ❹ ❻ ❺	Institutional database, national databases	2009–2018	39,470	Private
Matz et al. (2023)	❶ ❹ ❸ ❺	Institutional database, Institutional App	2022	50,095	Private
Rebelo Marcolino et al. (2025)	❸	Institutional LMS	ND	567	Private

❶: Enrolment/institutional features; ❺: Social/emotional features (feedback, wearables etc.); ❷: Demographic/personal features; ❹: Academic/academic history features; ❸: Online learning/LMS behavior; ❻: Financial features

A further example is provided by Delen et al. (2024), who used SHAP to interpret student trends at both global and local levels. Globally, this supports policymakers in introducing institution-level preventative measures (Ortigosa et al., 2019), while locally it helps identify individual student circumstances. Supplying support staff with information about which factors contribute most to a student's dropout risk enables more tailored and effective interventions, moving beyond one-size-fits-all approaches.

3.4. Computational efficiency

Performance evaluation metrics for student dropout prediction models vary across studies, with accuracy, recall, F-score (Fahd et al., 2022; Xiao and Hu, 2023), and precision (Fahd et al., 2022) being the most commonly reported. However, measures relating to model training time, inference speed, and scalability are rarely considered, reflecting a clear disconnect between research and practical implementation. This issue was also identified by Ortigosa et al. (2019). Their study highlighted several challenges encountered when attempting to implement a previously developed model, including costly maintenance of complex algorithms, limited adaptability of outputs and user interfaces, insufficient explainability, and high computational demand. They concluded that to ensure real-world feasibility, the model needed to be simplified to a C5.0 decision tree, accepting a minor reduction in sensitivity and specificity in exchange for avoiding the high computational cost, finetuning complexity, and lack of interpretability associated with the original RF algorithm. As shown in Table 3, our review identified only five studies that explicitly reported computational measures such as training time (Romero and Liao, 2025; Roy and Farid, 2024; Vives et al., 2024; Zhang et al., 2025) and inference time (Pek et al., 2023). Consideration of model scalability was similarly limited. Evaluating performance on datasets of varying size is essential for understanding real-world behavior, where data volume typically increases over time (Zhang et al., 2025). For example, Zhang et al. (2025) reported an increase in training time from 15 to 110 seconds for their Graph Neural Network Transformer-

InceptionNet (GNN-TINet) model when scaling from 20,000 to 150,000 cases, compared with an increase from 20 to 170 seconds for an SVM model over the same range. Matz et al. (2023) evaluated an Elastic Net and an RF model across four university datasets that differed significantly in size, observing declining performance with larger datasets, particularly for the RF model. In another approach, Pan et al. (2024) assessed predictive accuracy using two datasets of different scales, prioritizing early-stage accuracy with minimal available information while still accounting for projected growth in student numbers. Despite the clear importance of scalability for real-world viability, only 26% of the reviewed studies explicitly evaluated the scalability of their models.

Table 3: Dataset diversity and institutional coverage - model practicality criteria

Study	Model	Interpretability	Computational demand	Scalability	Actionability
Adejo and Connolly (2018)	DT, ANN, SVM, ensemble	✗	✗	✗	✗
Alamuddin et al. (2019)	SLR	≈	✗	✓	✗
Arnold and Pistilli (2012)	Not defined	✓	✗	≈	✓
Delen et al. (2024)	Deep MLP	✓	✗	✗	✓
Du et al. (2020)	LVAEPre (DNN in LVAE framework), DT	✓	✗	✗	≈
Hoca and Dimililer (2025)	SVM, RF, CatBoost, KNN, MLP, NB, LR, CART	✓	≈	✗	✓
Hoffait and Schyns (2017)	RF, LR, ANN	✓	≈	✗	≈
Huang et al. (2022)	LSTM, LR, NB, RF, KNN, SVM, ID3, CART, MLP	✗	✗	✗	✗
Latif et al. (2023)	RF, FDT, BN, SVM, NB, SLR, boosting + bagging ensembles	✗	✗	≈	✗
Maniyan et al. (2024)	Hybrid of K-Means, Apriori, QUEST, NB, CHAID, CART, C5.0, SVM	≈	✗	✗	≈
Masood et al. (2024)	HDL model with ECNN and ResNetV2, MLP, DFFNN	✗	≈	≈	✗
Matz et al. (2023)	Elastic Net, RF	≈	✗	✓	✗
Mosia (2025)	LR	✓	≈	✗	≈
Mustofa et al. (2025)	HLRNN model	✓	✗	≈	✗
Nagy and Molontay (2024)	CatBoost, NGBoost, EBM, LR, XGBoost, GBC, LDA, Ada Boost, LightGBM, QDA	✓	✗	✗	✓
Olaya et al. (2020)	RF, XGBoost, XLearner, RLearner, KL, ED, Chi, CTS	≈	✗	≈	≈
Ortigosa et al. (2019)	C5.0	✓	≈	✓	✓
Pan et al. (2024)	MODDQN	✗	✗	✓	✗
Pek et al. (2023)	NB, RF, DT, KNN, SVM, AdaBoost, LR, ensemble	≈	✓	✗	✓
Rabelo and Zárate (2025)	CART, LR, FFMLP, ensemble	✗	✗	✗	≈
Rebelo Marcolino et al. (2025)	NSGA-II with RF, XGBoost, CatBoost, KNN, LR, NB	✓	≈	≈	≈
Romero and Liao (2025)	LASSO, RF, ANN, XGBoost, GAM	✓	✓	✗	≈
Roy and Farid (2024)	DT, KNN, LR, NB, SVM	✗	✓	✓	✗
Skittou et al. (2024)	SVM, RF, SGD, KNN	≈	✗	✗	✓
Sonnleitner et al. (2025)	Trial Exam Grade, LR, SLR, XGBoost, RF, SVM	≈	✗	✗	✗
Tong and Li (2025)	Ensemble of KNN, NB, RF, GBDT, XGBoost, MLP, and LR	✓	✗	✗	≈
Van Petegem et al. (2023)	SGD, LR, SVM, RF	✓	✗	✓	≈
Vives et al. (2024)	LSTM, DNN, DT, RF, LR, SVM, KNN	✗	✓	✗	✗
Wen and Juan (2023)	DNN and FNN	✗	✗	✗	✗
Wong et al. (2025)	Multiple regression	≈	✗	✗	✗
Zanellati et al. (2024)	RF, FTT	✓	✗	✗	✓
Zhang et al. (2025)	GNN-TINet	≈	✓	✓	✗
Zhidkikh et al. (2024)	SGD, LR, SVM, RF	≈	≈	✓	✗

Interpretability: XAI, white-box model, or surrogate model (✓), feature impact or statistical analysis (≈), little to no consideration (✗); Computational demand: Training time measured (✓), adjustments to lower complexity (≈), little to no consideration (✗); Scalability: Multiple datasets of varied sizes or a growing dataset (✓), considerations for scalability made (≈), little to no consideration (✗); Actionability: LAD/visualization or alerts (✓), feature impact or deducted rules (≈), little to no consideration (✗)

4. Proposed novel framework for learning analytics implementation

Drawing on the gaps identified in our review, we developed a six-layered framework that proposes a practicality-oriented approach to reducing the divide between research and practice. The framework introduces the core functions required for practical predictive models and organizes them into four inner layers and two outer layers within the research process. Its design is informed by the evaluation tables used to structure our literature analysis, specifically Tables 1, 2, and 3. These tables collectively highlight limited consideration of key criteria such as generalizability, interpretability, computational demand, scalability, and actionability. As we summarized in Table 4, each layer addresses specific shortcomings in these areas. The order of the layers reflects how they should be considered, enabling this section to function as a guide for evaluating a model's practicality before experimentation begins. We recommend following this order while recognizing that the inner layers influence one another, as illustrated in Fig. 4. These inner layers represent the core decisions that shape the experiment setup and should be defined at the outset, guided by the research questions and aims. Surrounding this core are the two outer layers,

which are not directly selected by the researcher but instead serve as mechanisms that drive model evolution. These layers represent iterative reflection on the decisions made in the inner layers, first through the lens of this study's overarching conceptual purpose, and subsequently through ethical and risk-related considerations. Fig. 4 visualizes the interconnected nature of the layers and illustrates how outer-layer awareness encompasses the entire process, presenting the framework as a full-cycle approach.

4.1. Explanation of each layer

4.1.1. Data and input layer

The first layer of our framework addresses best practices for dataset selection in practical LA research, guided by the gaps identified in the literature review. Public availability of datasets is a central driver of reproducibility. Institution-specific datasets can constrain models by binding them to a particular context, a unique feature set, or a fixed sample size. In addition, the private nature of these datasets restricts external validation and hinders replication in settings beyond the original institution. Accordingly, this layer encourages the use of publicly available datasets or, where feasible, the publication of anonymized private datasets. As an alternative, researchers may consider excluding highly specialized variables from private datasets. While analyzing model performance using unique features can be insightful, such features impede reproducibility because comparable variables are unlikely to be available in other datasets. Restricting the input to widely accessible features enables benefaction research, in which existing datasets with similar attribute sets can be used to replicate or extend findings when the original data cannot be released. Where studies involve multiple datasets, these considerations should be applied consistently across all of them. By introducing practicality considerations at such an early stage, the framework seeks to reduce barriers to model adoption and to strengthen the reproducibility of research in this field.

4.1.2. Algorithm selection layer

The second layer emphasizes the role of algorithm selection in relation to the intended outcome. Our framework proposes a novel outcome-oriented approach to model selection that foregrounds interpretability and computational feasibility early in the research process. Researchers should consider the trade-offs between interpretability, computational cost, and predictive performance when choosing algorithms. Training complex or deep ML models can be computationally intensive, and the tools required to interpret these models are often costly (Hoca and Dimililer, 2025). As a result, substantial feature reduction may be necessary to maintain feasibility (Roy and Farid, 2024).

Conversely, simpler models may struggle with the large, complex datasets expected of typical established universities (Masood et al., 2024). For this reason, researchers introducing a novel algorithm should consider including a diverse set of comparison models. This selection will also be shaped by the decisions made in Layer 1. For large and complex datasets, it is beneficial to evaluate both white-box and black-box models to assess interpretability and computational demand, alongside standard performance metrics. When using a private dataset, opting for a white-box model can provide greater insight into the modelling process. If the research aims of the study are to evaluate a specific aspect of the modelling process, such as a data balancing method or an interpretability tool, integrating it into multiple algorithms enables more comprehensive evaluation. In other cases, such as replication studies with proposed enhancements, comparing the original with the refined version can fulfil the objectives of this layer. Through this outcome-based model selection logic, the framework embeds a continuous awareness of practical considerations throughout LA research.

4.1.3. Practicality and deployment layer

This layer advocates for the integration of practicality and deployment measures into model evaluation. While achieving high accuracy and minimizing false positives is essential, particularly for models that may directly impact students, additional considerations are necessary to build trust in the system, promote adoption and ensure efficient operation (Nagy and Molontay, 2024; Ortigosa et al., 2019). This layer introduces a novel performance-deployment trade-off perspective, emphasizing that model evaluation should extend beyond predictive accuracy. Researchers are encouraged to measure training and inference time, and to explicitly assess scalability, generalizability, interpretability and actionability. Importantly, these criteria should be considered alongside accuracy rather than after it.

Models with comparable accuracy but lower computational cost and greater transparency should be prioritized over more accurate but resource-intensive alternatives. A lack of interpretability or transparency is often the source of ethical concerns in ML models that inform decision-making or label student data. Generalizability and scalability should ideally form part of practicality assessments, although these may not always be feasible to test directly. The choices made in the first two layers (dataset and input selection, and algorithm selection) can determine whether evaluation of these criteria is possible. Where certain aspects cannot be assessed, the next layer can provide compensatory balance.

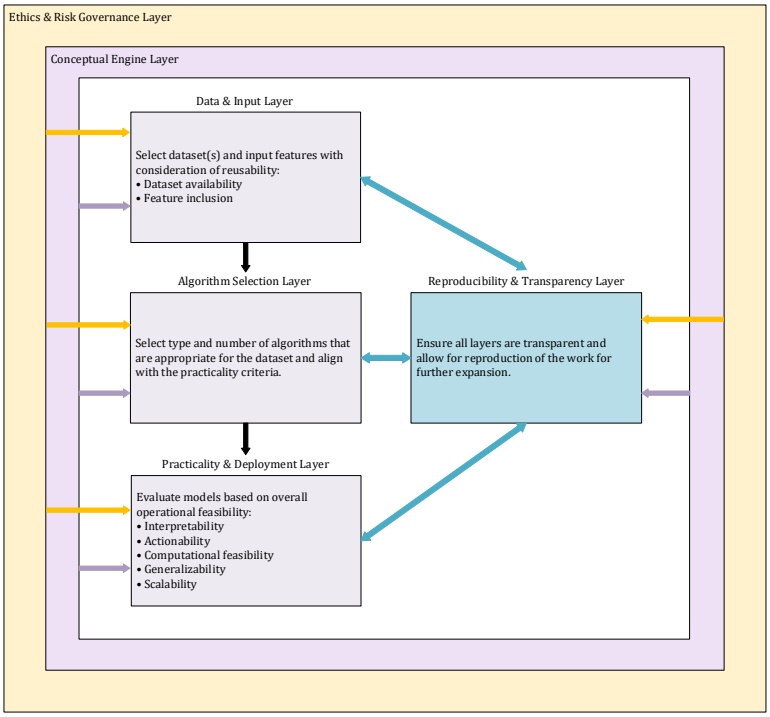

Fig. 4: Visualization of full-cycle application of our proposed framework

Table 4: Classification of models and educational tasks - framework layers table

Layer	Informed by	Function	Novelty contribution	Example
1. Data and input layer	Table 2: Practical constraints in reviewed studies - data source classification	Defines types of datasets and input features selected.	Emphasizes use of public-only datasets for reusability and minimal administrative dependency.	Use Open University Learning Analytics Dataset (OULAD) with variables like age, gender, and course access logs.
2. Algorithm selection layer	Table 3: Dataset diversity and institutional coverage - model practicality criteria	Justifies selection of ML models based on accuracy, interpretability, and dropout sensitivity.	Introduces outcome-based model selection logic.	Select logistic regression when needing transparency in classification for at-risk students.
3. Practicality and deployment layer	Table 3: Dataset diversity and institutional coverage - model practicality criteria	Evaluates operational feasibility, including latency, explainability, and actionability.	Establishes a performance vs. deployment trade-off metric.	Choose decision tree model if it provides near-equal accuracy with faster inference time than an ensemble method.
4. Reproducibility and transparency layer	Table 1: Overview of reviewed learning analytics studies reproducibility assessment	Assesses reproducibility across code, methods, and data access.	Includes a compliance filter to assess model replicability.	Rate a study high if it provides both source code and public dataset in supplementary materials.
5. Conceptual engine layer	Table 5: Study reproducibility indicators - conceptual engine framework	Visualizes flow from input → process → output, aligning with real-world operations.	Supports a full-cycle model representation for LA deployment.	Diagram showing: Demographic + LMS Logs → ML Processing → Dropout Alert for Staff.
6. Ethics and risk governance layer	Table 6: Ethical risk classification	Identifies ethical risks such as bias, surveillance, and opacity.	Provides a risk-aware logic for safe model deployment.	Flag a model using postcode as a variable due to potential bias, requiring ethics review.

4.1.4. Reproducibility and transparency layer

Reproduction studies offer significant benefits for validating and further developing LA models. In our proposed framework, we present reproduction studies as a practical solution to resource limitations by enabling researchers to build on existing models. This allows for focused exploration on overlooked areas. For example, where generalizability or scalability could not be assessed in the original study, reproduction studies can evaluate model robustness using additional datasets. Ensuring reproducibility is especially valuable when a novel algorithm is introduced. Studies that provide access to datasets and source code should therefore be regarded highly. As we demonstrate in Table 1, only a small proportion of reviewed studies offer sufficient detail for full reproduction, which substantially restricts progress in the field. Fig. 4 illustrates how this layer interacts with the other inner layers, positioning it as a balancing mechanism. For instance, where one study emphasizes training-time optimization, a reproduction study may focus on evaluating appropriate interpretability tools.

Reproducible studies also enhance transparency, enabling deeper examination of ethical concerns and potential biases within models (Zhidkikh et al., 2024). This transparency is essential for supporting the outer layers of the framework. At the same time, the two-way connection between this layer and preceding layers reflects their interdependence. The extent to which a study is reproducible depends on the decisions made in the first three layers, while those decisions should themselves be informed by reproducibility requirements. Accordingly, we describe this layer as both a compliance filter and a

balancing tool, ensuring that models are replicable for validation and future development.

4.1.5. Conceptual engine layer

Following the four practical inner layers, the first outer layer of the framework is the conceptual engine. As outlined in Table 5, this layer involves revisiting the decisions made within the inner layers through the lens of the overarching framework concept, namely, achieving long-term operational feasibility. The layer serves two primary functions: to prompt researchers to re-evaluate the design of their current study, and to support the examination of existing models for reproduction or benefaction. The term "engine" signifies the ongoing momentum it introduces into the decision-making processes of the inner layers. In its first function, the conceptual engine encourages researchers to critically review their methodological choices to ensure stronger alignment with operational feasibility. This reflective process is valuable even when the current model is not expected to reach full deployment. For example, if interpretability issues become apparent at this stage, researchers may decide to incorporate additional algorithms in Layer 2 to facilitate more robust interpretability assessment. This illustrates how the conceptual engine guides strategic adjustments before experimentation begins. The second function pertains to reproduction studies, where the conceptual engine supports mapping an existing model onto the inner layers to identify opportunities for extension or refinement. In this sense, the engine represents a continual revisiting of the study's conceptual foundation, with the aim of enhancing the criteria that underpin model practicality in a collaborative and iterative cycle. This layer also reinforces the need to balance the first four layers, ensuring that reproducibility and transparency enable other researchers to test proposed enhancements. Widespread application of this approach facilitates improved comparability between models and supports more nuanced investigation. Fig. 4 provides a visual representation of this closed research loop, illustrating how the conceptual engine surrounds the first four layers to create a process of continuous improvement. Ultimately, this layer supports more effective innovation by acknowledging resource constraints, prompting peer review, and encouraging refinement built upon robust foundational models.

Table 5: Study reproducibility indicators examined through the conceptual engine framework

Input	Process	Output	Relevance to framework
Public dataset and source code of proposed model	Reproduce proposed model; Evaluate and compare to original	Validated model ready for expansion	Layers 1, 2, and 4
Reproduced model	Assess model variations to optimize performance and practicality trade-off	Model optimized for practical implementation	Layers 3 and 6
Public datasets varied in scale and input features	Evaluate generalizability and scalability of model	Widely applicable model	Layer 3
Finalized model	Record source code and detailed methodology	Reproducibility information available in an online repository	Layer 4

4.1.6. Ethics and risk governance layer

The final layer of the framework centers on risk awareness and mitigation to support ethical research and implementation planning. At this stage, researchers are encouraged to re-examine their methodological decisions through an ethical lens and make any necessary adjustments. This is particularly important when seeking to narrow the gap between theory and practice, as awareness of deployment-related risks must begin during the research phase (Ortigosa et al., 2019). As summarized in Table 6, we identified seven major ethical risk areas across relevant research literature and proposed corresponding mitigation strategies. When reviewing a study through this outer layer, researchers should identify any risks stemming from earlier decisions and evaluate them using the FATE framework as explored by Memarian and Doleck (2023). The term refers to the principles of fairness, accountability, transparency, and ethics, against which each layer's decisions should be assessed. Beginning with the Data and Input Layer, dataset selection and preparation directly affect model fairness. Careful consideration and pre-processing are essential, which may include excluding sensitive features where these pose risks of discrimination (Memarian and Doleck, 2023). However, removing too many features can introduce selection bias or diminish model performance, requiring a balanced approach. Privacy and consent of data are recurring concerns when working with student records. At the research stage, privacy risks can be mitigated by using public datasets or by fully anonymizing private data before modelling. This allows researchers to analyze and develop models in a risk-reduced environment.

The Algorithm Selection Layer most directly influences accountability and transparency, although fairness considerations also apply. Accountability concerns revolve around who is responsible for decisions or outcomes driven by the model (Memarian and Doleck, 2023). These risks can reduce user trust and hinder adoption. While institutional policy largely determines accountability, researchers can mitigate associated risks by improving interpretability, thereby enhancing transparency of the model's decision-making process. Memarian and Doleck (2023) argued that end-users, who are typically non-technical staff, may lack understanding of how the model works, potentially undermining perceived fairness and trust. Fairness considerations extend to future deployment decisions. For example, naming

conventions for the classification outputs can lead to unjust consequences. Labels such as "lost causes" (Olaya et al., 2020) may stigmatize students and influence how staff perceive them.

The Practicality and Deployment Layer offers opportunities to strengthen FATE characteristics. Enhancing interpretability and actionability increases transparency, which in turn supports fairness and accountability. Encouraging holistic model evaluation can also extend beyond the five practicality criteria, for example, by incorporating fairness analysis to identify biases involving protected features. Interpretability is particularly critical for FATE, especially when addressing ethical risks associated with misclassification. Interpretable models or interpretability tools can reveal why a student was flagged at risk and enable staff to examine the context before making contact, helping to prevent inappropriate or harmful interventions.

The Reproducibility and Transparency Layer further supports transparency but also requires rigorous data privacy protocols when private datasets are used. Trade-offs between the earlier layers can significantly affect FATE characteristics. For example, choices around datasets and algorithms have direct implications for the reproducibility and transparency of a study. Ethics is often the most challenging dimension to define, as it depends on institutional values and individual perspectives (Memarian and Doleck, 2023). West et al. (2016) similarly highlighted the fluidity of ethical boundaries, noting that what is considered ethical is shaped by the underlying values of researchers and institutions. The overarching aim of this framework is to promote predictive learning analytics research that genuinely improves student support and enables effective interventions for those most at risk. However, researchers must weigh the potential costs and long-term implications of their decisions. For example, Table 6 identifies self-report bias as a risk when incorporating survey-based data to improve predictive accuracy. While this may be intended to benefit students, it raises the question of whether it is appropriate to use direct personal input to fuel ML systems, or to request such data at scale. These decisions must be evaluated from a real-world perspective, considering potential long-term impacts that may cause harm (West et al., 2016). Although ethical reassessment is essential at the point of implementation, reviewing each layer's decisions through FATE principles during the research stage can significantly strengthen the development of ethical and responsible ML systems.

In summary, the six-layered framework provides a new approach for planning research into predictive modelling within LA. Although each layer stands independently and relates to specific aims and research questions of a study, the first four layers both influence and depend on one another. Recognizing the practical constraints of limited resources, the framework encourages researchers to seek an appropriate balance across these layers rather than striving for complete alignment with all criteria. The fifth layer (Conceptual Engine) drives continuous model improvement by prompting ongoing reflection on the research approach in relation to the framework's overarching goal of operational feasibility. This layer also underpins reproduction studies, enabling robust base models to be mapped against the inner layers to identify and implement targeted enhancements. Finally, the Ethics and Risk Governance Layer adds a further level of review, ensuring that decisions across the research design are examined through the principles of fairness, accountability, transparency, and ethics (FATE).

Table 6: Ethical risk classification

Risk	Description	Related studies	Mitigation strategies
Self-report bias	Survey responses self-reported by students may be biased.	Adejo and Connolly (2018), Rabelo and Zárate (2025), Sonnleitner et al. (2025), and Zhidkikh et al. (2024)	Avoid inclusion of opinion-based self-reporting. Combine self-report data with trace data.
Selection bias	Sample may not be representative of the real population.	Delen et al. (2024), Hoca and Dimililer (2025), Maniyan et al. (2024), Matz et al. (2023), Olaya et al. (2020), Rabelo and Zárate (2025), Rebelo Marcolino et al. (2025), and Roy and Farid (2024)	Avoid exclusion of students in dataset. Avoid including features that are only partially available.
Temporal bias	Behaviors may change throughout the term, and predictions need to be adjusted.	Adejo and Connolly (2018), Delen et al. (2024), Du et al. (2020), Hoca and Dimililer (2025), Huang et al. (2022), Latif et al. (2023), Mustofa et al. (2025), Nagy and Molontay (2024), Rabelo and Zárate (2025), Romero and Liao (2025), Tong and Li (2025), Wen and Juan (2023), Wong et al. (2025), and Zhidkikh et al. (2024)	Develop models that can work with early data and re-assess risk classification throughout the study period.
Lack of interpretability	Cannot check whether the model works correctly. Cannot check for fair and unbiased decision-making. Cannot provide reason for categorization.	Adejo and Connolly (2018), Huang et al. (2022), Latif et al. (2023), Masood et al. (2024), Nagy and Molontay (2024), Pan et al. (2024), Pek et al. (2023), Rabelo and Zárate (2025), Roy and Farid (2024), Van Petegem et al. (2023), Vives et al. (2024), and Wen and Juan (2023)	Use white-box models or introduce interpretability tools to investigate the model's decision-making process.
Privacy and consent	Sensitive private data is used.	Adejo and Connolly (2018), Delen et al. (2024), Du et al. (2020), Hoca and Dimililer (2025), Hoffait and Schyns (2017), Maniyan et al. (2024), Mosia (2025), Nagy and Molontay (2024), Olaya et al. (2020), Ortigosa et al. (2019), Pan et al. (2024), Pek et al. (2023), Rabelo and Zárate (2025), Rebelo Marcolino et al. (2025), Skittou et al. (2024), Sonnleitner et al. (2025), Van Petegem et al. (2023), Vives et al. (2024), Wong et al. (2025), Zanellati et al. (2024), and Zhidkikh et al. (2024)	Ensure collection of informed consent from all affected students, establishment of secure storage and cybersecurity measures, and monitor access carefully.
Stigmatization of classes	Classification labels may lead to stigma against students in that category.	Hoffait and Schyns (2017), Latif et al. (2023), and Olaya et al. (2020)	Select wording carefully and set policies that regulate treatment of students based on their risk classification.
Wrongful classification	Students may be classified incorrectly and receive incorrect treatment.	Adejo and Connolly (2018), Huang et al. (2022), Latif et al. (2023), Masood et al. (2024), Nagy and Molontay (2024), Pan et al. (2024), Pek et al. (2023), Rabelo and Zárate (2025), Roy and Farid (2024), Van Petegem et al. (2023), Vives et al. (2024), and Wen and Juan (2023)	Interpretability of models can allow staff to investigate reasons for classification and adjust intervention accordingly.

5. Mapping literature to implementation framework goals

Following the development of the layered framework, we re-assessed the reviewed studies using the practicality criteria to determine their alignment with the goals of each layer. We first evaluate alignment with each layer individually, then provide an overarching analysis based on these findings. As outlined in Section 4, the first four layers encourage early consideration of deployment feasibility within the research design, while Layer 5 introduces a full-cycle perspective and Layer 6 focuses on ethics and risk awareness. For each layer, we developed a set of recommendations derived from the framework goals. Consistent with the evaluation tables, studies were classified as having met the goal, partially met it, or not met it.

5.1. Data and input layer

We consider studies to fully align with the first layer when they meet two or more of the following conditions: (1) use of public datasets or publication of the dataset, (2) inclusion of multiple datasets, or (3) investigation of different feature combinations. Studies meeting only one of these conditions, or (4) whose models rely solely on widely available features, are classified as partially aligned. Studies that do not meet any of these criteria are classified as not aligned. The implementation-focused studies such as Arnold and Pistilli (2012) and Alamuddin et al. (2019) were excluded from assessment, because they rely on live institutional data that cannot be made public and is inherently specialized. Although Ortigosa et al. (2019) involved an implemented model, making it inherently specialized, it was only included in the assessment due to their focus on model development rather than reporting outcomes of deployment.

5.1.1. No alignment

Delen et al. (2024), Du et al. (2020), Olaya et al. (2020), Ortigosa et al. (2019), Pek et al. (2023), Rabelo and Zárate (2025), Rebelo Marcolino et al. (2025), Skittou et al. (2024), Sonnleitner et al. (2025), Van Petegem et al. (2023), and Zhidkikh et al. (2024) did not align with the Data and Input Layer. These studies relied on a single private dataset, often containing specialized or uncommon variables such as detailed discussion forum activity, family academic history, or self-assessment data. While Delen et al. (2024) and Skittou et al. (2024) made use of national databases, neither explored alternative feature combinations, and therefore, they do not meet any of the alignment criteria.

A pattern emerges within this group: studies that fail this layer tend to show weak or no alignment with other layers, with more than half failing two or more layers. In contrast, most show strong alignment with Layer 3 (Practicality and Deployment). This may reflect the use of institution-specific datasets, suggesting an implicit aim of supporting local operational deployment. This is further reinforced by their weak alignment with Layer 4, indicating that models are designed for specialized rather than generalizable use.

5.1.2. Partial alignment

Mosia (2025) used a small private dataset containing widely available features and employed a simple regression model, meeting criterion (4). Likewise, Hoca and Dimililer (2025), Hoffait and Schyns (2017), Nagy and Molontay (2024), Vives et al. (2024), and Zanellati et al. (2024) applied popular ML algorithms to common features, but their reliance on private datasets prevents full alignment.

Zhang et al. (2025) utilized a large public dataset (1) with a broad range of input features, including common demographic and academic attributes. However, inclusion of variables such as mental health status, teacher feedback, school environment ratings, and emotion-recognition wearables data, without evaluating feature subsets that exclude these less accessible attributes, precluded full alignment. Latif et al. (2023), Masood et al. (2024), and Tong and Li (2025) all used large public datasets (1) with some common features. Although data from massive open online courses (MOOCs) typically exceeds what standard learning management systems (LMS) collect, the availability of code and data supports reproducibility and aligns with our framework's objectives. Huang et al. (2022) similarly used a public dataset (1) but limited their analysis to online learning behavior data, supported by a literature review. Romero and Liao (2025) removed statistically insignificant features from their public dataset (1), resulting in a feature set largely composed of widely available attributes (4). Mustofa et al. (2025) included the full set of commonly available features (4) in their public dataset (1).

Adejo and Connolly (2018) used a small private dataset and incorporated self-assessment survey data. However, their detailed evaluation of the contribution of different feature types (3) demonstrates alignment with the goals of this layer. Wong et al. (2025), although they used private data, investigated feature types thoroughly using common variables (4) and evaluated two feature inclusion strategies (3), reflecting consideration for practical deployment. Maniyan et al. (2024) similarly used private data with common features (4) and evaluated feature types separately (3), showing methodological alignment with the framework.

Among the 16 partially aligned studies, many exhibit weak or no alignment with other layers, with nearly one-third neglecting two layers. However, three fully align with Layer 2 (Algorithm Selection), suggesting potential resource constraints where authors who prioritize testing multiple algorithms may limit the complexity or variety of datasets they include.

5.1.3. Full alignment

Pan et al. (2024) evaluated their algorithm using two large public datasets (1, 2) with common MOOC-related features. Comparing two fully public datasets aligns well with the proposed framework's emphasis on continuous improvement and reproducibility. Matz et al. (2023) and Roy and Farid (2024) selected four datasets of varying sizes and feature compositions (1, 2, 3), demonstrating early consideration of practicality. Matz et al. (2023) further tested whether models trained on one institution's dataset could predict dropout in another, extending their analysis across contexts. Wen and Juan (2023), although they used a single dataset, met criterion (1) by selecting a publicly available dataset and criterion (3) through a detailed evaluation of feature-type combinations. Their separation of feature types into distinct datasets moves beyond feature contribution analysis and towards diagnosing deployment challenges for each data type.

These fully aligned studies also show strong performance in other layers as well, particularly Layer 4 (Reproducibility and Transparency). This is likely due to the emphasis placed on public datasets, which inherently support reproducibility. However, half of these studies do not align with Layer 2, reinforcing the earlier observation that extensive dataset-focused experimentation may limit the resources available for testing multiple algorithms or conducting extended performance evaluations.

5.2. Algorithm selection layer

For the second layer, studies are expected to (1) include multiple algorithms, (2) compare both white-box and black-box models, and (3) assess model performance beyond accuracy measures. These criteria support thorough model comparison, outcome-based selection logic, and a practical implementation perspective. Owing to the significance of each criterion, studies must meet at least two criteria to be classified as partially aligned, and one or none results in a no-alignment classification.

5.2.1. No alignment

Delen et al. (2024) applied only a single deep learning model and evaluated it using accuracy, precision, recall, and F-score, resulting in no alignment with this layer.

Mosia (2025) and Wong et al. (2025) also evaluated a single white-box model. Although their focus suggests some practical awareness, the absence of comparative modelling and additional practicality measures means they do not meet the alignment criteria.

While Masood et al. (2024), Olaya et al. (2020), Pan et al. (2024), Wen and Juan (2023), and Zanellati et al. (2024) compared multiple models (1), they relied exclusively on black-box algorithms and accuracy-based performance measures, which do not satisfy the combined criteria for this layer.

Misalignment in this layer appears strongly associated with misalignment in other layers, with all but one of the studies also lacking alignment in one or two additional layers. Pan et al. (2024) is the exception, achieving full alignment with Layers 1 and 4, likely due to the interconnected nature of those two layers.

5.2.2. Partial alignment

Adejo and Connolly (2018), Du et al. (2020), Latif et al. (2023), Rabelo and Zárate (2025), and Tong and Li (2025) compared a range of models, including white-box, black-box, and an ensemble (1, 2). Although their evaluation focused primarily on accuracy and error metrics, they addressed two of the required criteria and therefore partially aligned with the goals in this layer.

Similarly, Hoca and Dimililer (2025), Hoffait and Schyns (2017), Huang et al. (2022), Maniyan et al. (2024), Matz et al. (2023), Mustofa et al. (2025), Nagy and Molontay (2024), Ortigosa et al. (2019), Rebelo Marcolino et al. (2025), Skittou et al. (2024), Sonnleitner et al. (2025), Van Petegem et al. (2023), and Zhidkikh et al. (2024) compared white-box and black-box models (1, 2) but relied predominantly on accuracy-based metrics, placing them in this category.

In a different context, Arnold and Pistilli (2012) focused on real-world implementation using a single model, but evaluated it using operational outcomes such as retention rates and user feedback (3). According to the framework criteria, these studies also partially align.

Across the partially aligned, overall alignment with other layers is low. Most demonstrate only partial or no alignment across the remaining layers. Two studies achieve full alignment with the Practicality and Deployment Layer, presumably because their interpretability-focused approach introduced additional practicality considerations.

5.2.3. Full alignment

Zhang et al. (2025) fully align with this layer by comparing their proposed model with eight additional ML models (1), including a white-box DT (2). They also assessed training time across dataset scales and incorporated a predictive consistency score in addition to accuracy measures (3), demonstrating clear attention to deployment feasibility. Similarly, Pek et al. (2023), Romero and Liao (2025), Roy and Farid (2024), and Vives et al. (2024) compared both white-box and black-box models (1, 2) and incorporated computational cost measures in their evaluations (3). Roy and Farid (2024) further evaluated the feature reduction factor, ensuring their feature selection method did not oversimplify the dataset.

This categorization supports earlier observations, showing strong alignment with the Algorithm Selection Layer, which correlates with strong alignment in the Practicality and Deployment Layer. Two partially aligned studies showed a similar pattern, reinforcing the idea that robust algorithm comparison encourages attention to practicality measures. Each fully aligned study also achieved partial alignment with the Reproducibility and Transparency Later, suggesting that research focused on algorithm evaluation tends to provide more detailed methodologies or pseudocode, increasing transparency.

5.3. Practicality and deployment layer

For the third layer, we evaluated whether each study aligned with the five practicality criteria outlined in the evaluation tables: Interpretability (1), actionability (2), computational feasibility (3), scalability (4), and generalizability (5). Each study received a full, partial, or no consideration rating for each criterion. Given the resource limitations typical of empirical studies, we consider a study fully aligned with this layer if it fully addresses at least three criteria or does not fail any criterion outright. Studies that fully address fewer than three criteria and fail; at least one is classified as partially aligned. Studies that do not fully address any criterion and fail one or more are classified as not aligned. As noted in the descriptions of Tables 1 and 3, studies received partial alignment where theoretical considerations or practical adjustments were made without explicit methodological inclusion.

5.3.1. No alignment

Adejo and Connolly (2018) and Huang et al. (2022) demonstrated no explicit consideration of the practicality criteria. Adejo and Connolly (2018) focused primarily on feature contribution and model-type comparisons, while Huang et al. (2022) examined sequential versus aggregated datasets. Latif et al. (2023), Rabelo and Zárate (2025), Sonnleitner et al. (2025), Wen and Juan (2023), and Wong et al. (2025) each fully addressed only one criterion. Maniyan et al. (2024), Masood et al. (2024), and Olaya et al. (2020) showed some implicit consideration of various criteria, but none explicitly.

As with previous layers, these studies also show weak alignment across the framework. Only one study aligns fully with another layer. Most show particularly poor alignment with Layer 2 (Algorithm Selection), followed by Layer 4 (Reproducibility and Transparency), likely because limited attention to model development is also reflected in reduced attention to practicality.

5.3.2. Partial alignment

Among the implementation-focused studies, we classified Alamuddin et al. (2019) as partially aligned. Both involved multiple universities (4) and, although not explicitly stated, would have considered the remaining criteria during their operational evaluations. For example, the authors note that several institutions discontinued participation due to lack of actionability, indicating shortcomings in interpretability, actionability, and generalizability. Arnold and Pistilli (2012) emphasized interpretability (1) and actionability (2) through their traffic-light system and staff dashboard, and the sustained use of their system suggests consideration for scalability (4). However, their focus on a model tailored to Purdue University limits generalizability, and computational feasibility considerations are not reported.

Du et al. (2020), Skittou et al. (2024), and Tong and Li (2025) each fully addressed one practicality criterion and partially addressed another. Vives et al. (2024) included the training time measure (3) but omitted other practicality measures. Hoffait and Schyns (2017) and Mosia (2025) also neglected only one criterion fully, but explicitly addressed only interpretability (1). Delen et al. (2024) and Zanellati et al. (2024) included detailed feature contribution analysis and XAI tools, demonstrating attention to actionability (2). Pan et al. (2024) evaluated scalability and generalizability (4, 5) through diverse datasets. Studies with slightly greater consideration include Matz et al. (2023), Mustofa et al. (2025), Nagy and Molontay (2024), Pek et al. (2023), Romero and Liao (2025), Van Petegem et al. (2023), and Zhidkikh et al. (2024), each neglecting two criteria. Hoca and Dimililer (2025) and Zhang et al. (2025) showed strong alignment, meeting two criteria fully and two partially, but they still fail to consider one criterion as required under our proposed framework. As such, they are categorized as partially aligned.

Studies in this category show varied alignment across the other layers. Some show multiple misalignments, while others fully align with one or more layers. This may stem from the multi-criteria approach used for categorization: Several studies in this group address interpretability and actionability, which may explain the stronger alignment with the Algorithm Selection Layer, while studies meeting Layer 1 criteria tend to address scalability.

5.3.3. Full alignment

Rebelo Marcolino et al. (2025) fully aligned with this layer, addressing all practicality criteria at least partially and adopting an outcome-oriented approach. Roy and Farid (2024) did not explicitly assess interpretability, but they evaluated training time, scalability, and generalizability (3, 4, 5). Moreover, by relying primarily on white-box models alongside their feature selection algorithm, interpretability is incorporated implicitly. Ortigosa et al. (2019) align closely with the framework, failing to assess only generalizability. Their thorough consideration of computational feasibility and real-

world deployment positions them strongly with this layer.

No clear pattern emerges among these fully aligned studies, reflecting the diverse ways in which practical considerations can be embedded within the model development.

5.4. Reproducibility and transparency layer

In contrast to the previous layers, the criteria for reproducibility and transparency were defined more explicitly. As outlined in Table 1, studies were assessed according to three indicators: (1) availability of the full dataset, (2) availability of code or detailed pseudocode, and (3) provision of a detailed methodology. Full alignment requires all three criteria. Partial alignment requires a detailed methodological description and at least one other criterion fully or partially fulfilled. Studies that only provide a methodological description or less are classified as not aligned.

5.4.1. No alignment

Adejo and Connolly (2018), Alamuddin et al. (2019), Arnold and Pistilli (2012), Du et al. (2020), Hoca and Dimililer (2025), Maniyan et al. (2024), Nagy and Molontay (2024), Ortigosa et al. (2019), Wong et al. (2025), and Zanellati et al. (2024) did not provide sufficient information to enable reproduction of their methods. Notably, these studies also lacked detailed methodological descriptions, meaning that reproduction would be limited to replicating only the conceptual approach. Although Mosia (2025) included a detailed mathematical model description, the broader methodology lacked sufficient detail to enable replication, resulting in a no-alignment classification.

These studies also show generally weak alignment with the inner layers, particularly Layers 1 (Dataset and Input) and 2 (Algorithm Selection). This is likely due to dataset and algorithm-selection decisions being major determinants of reproducibility, and these studies offering limited transparency in both areas.

5.4.2. Partial alignment

Delen et al. (2024), Olaya et al. (2020), Rabelo and Zárate (2025), Van Petegem et al. (2023), and Zhidkikh et al. (2024) were classified as partially reproducible due to the level of methodological detail provided (3). Rebelo Marcolino et al. (2025) also met this category, supported by the availability of extended model information in an earlier conference paper. Hoffait and Schyns (2017) provided less methodological detail, but included pseudocode (2), details for the final code and hyperparameter settings. Vives et al. (2024) similarly provided pseudocode (2), missing only dataset availability to enable full reproduction. Skittou et al. (2024) and Sonnleitner et al. (2025) indicated that their dataset or code was available upon request (1, 2). These studies were therefore classified as partially reproducible, acknowledging that such access is not guaranteed. Although Pek et al. (2023) provided extensive methodological information (3), including hyperparameters and code descriptions, their dataset and code were not openly available, preventing full alignment. Huang et al. (2022), Latif et al. (2023), Masood et al. (2024), Mustofa et al. (2025), Romero and Liao (2025), Roy and Farid (2024), and Wen and Juan (2023) also provided limited methodological (3), despite using publicly available datasets (1), and were therefore classified as partially aligned. Tong and Li (2025) provided both their original and modified datasets (1), pseudocode (2) and adequate methodological details (3), placing them in the partial alignment category. No clear pattern emerges within this category, likely because it includes the majority of studies. Although Layers 1 (Data and Input) and 2 (Algorithm Selection) directly influence reproducibility, they do not guarantee alignment with this layer due to the additional requirement of comprehensive methodological reporting.

5.4.3. Full alignment

Only two studies were found to be fully reproducible in our review. Matz et al. (2023) and Pan et al. (2024) provided complete datasets and code (1, 2), alongside detailed descriptions of their experimental procedures (3). Each also showed full alignment with the Layer 1 (Data and Input), consistent with the requirement that datasets must be publicly available to achieve full reproducibility. Although the small number limits broader conclusions, one of these studies showed no alignment with Layer 2 (Algorithm Selection). This observation supports earlier indications that a stronger data-selection focus may coincide with reduced algorithmic experimentation.

5.5. Conceptual engine layer

For the fifth layer we assessed whether studies demonstrated an ongoing, collaborative improvement cycle. As defined previously, this requires re-evaluating earlier methodological decisions with awareness of the framework's core aims of operational feasibility and reproducibility, ensuring that future research teams can reproduce and extend the proposed model. From our analysis across the previous layers, Roy and Farid (2024) were the only study to meet the framework's expectations in three layers, and partial alignment for the Reproducibility and Transparency Layer. Their study illustrates how comprehensive methodological choices can support the conceptual engine, even when full reproducibility is not achieved. The role of Layer 4 as a balancing mechanism is central to enabling the conceptual engine function.

Given that most studies must prioritize only a limited set of factors due to resource constraints, reproducibility becomes the mechanism through which the engine "moves": by enabling other researchers to build upon neglected areas, validate findings and iteratively enhance the base model. However, reproducibility represents the most significant weakness across the reviewed literature, with only two studies achieving full alignment. This is reflected in Table 7, which summarizes alignment across Layers 1 to 4 and forms the basis for assessing the conceptual engine. Studies such as Matz et al. (2023) and Pan et al. (2024) displayed gaps in earlier considerations, yet their strong reproducibility allows for validation and expansion. This highlights a critical shortcoming in current LA research practice: insufficient reproducibility impedes the collaborative progression that this layer aims to drive. The conceptual engine layer illustrates how an iterative approach, enabled by reproducible studies, can support the advancement of resource-constrained research.

5.6. Ethics and risk governance layer

Studies on student dropout prediction encounter challenges related to ethics and risk governance. As outlined in Section 4, ethical boundaries in LA research can be difficult to define, prompting our development of guidelines to support structured reflection. To evaluate alignment in this sixth layer, we categorized the reviewed studies into ethical risk areas and summarized them in Table 6. When assessing these findings, we also examined how alignment with previous layers mitigates specific risks. For example, studies that avoid incorporating survey or self-reported data reduce the likelihood of self-report bias, thereby enhancing sustainability. The framework also mitigates risk more broadly by promoting transparency, interpretability, and evaluation of methodological decisions against the FATE principles. This enables researchers and end-users to identify potential biases and detect wrongful classification prior to contacting students. Privacy and consent emerged as major concerns, particularly for studies relying on private institutional datasets. All studies not aligned with Layer 1 (Dataset and Input) displayed risks in this area. Our framework, therefore, recommends the use of public datasets where possible or, alternatively, full anonymization of a private dataset before use. Temporal bias and lack of interpretability were the next most common risk categories. Studies showing temporal bias tended to have low overall alignment, although no consistent pattern was evident. Lack of interpretability aligned with classifications in the Practicality and Deployment Layer, with most such studies receiving partial or low alignment there as well. Interestingly, many of these studies showed strong alignment with the Layers 1 and 2, suggesting that their final selected models were frequently black-box algorithms, which inherently reduce interpretability. Stigmatization, selection bias, and self-report bias were identified less frequently but were most evident in studies oriented towards deployment. Issues such as the fair treatment of students are managed at an institutional level through policy and procedure and are therefore less amendable to methodological correction. Similarly, excluding data or relying on partially available student records is not feasible in operational settings. These risk areas did not exhibit clear patterns when assessed against alignment with the inner layers.

Table 7: Evaluation of literature based on a layered framework

Study	Layer 1	Layer 2	Layer 3	Layer 4
Adejo and Connolly (2018)	≈	≈	✗	✗
Alamuddin et al. (2019)	✗	≈	≈	✗
Arnold and Pistilli (2012)	✗	≈	≈	✗
Delen et al. (2024)	✗	✗	≈	≈
Du et al. (2020)	✗	≈	≈	✗
Hoca and Dimililer (2025)	≈	≈	≈	✗
Hoffait and Schyns (2017)	≈	≈	≈	≈
Huang et al. (2022)	≈	≈	✗	≈
Latif et al. (2023)	≈	≈	✗	≈
Maniyan et al. (2024)	≈	≈	✗	✗
Masood et al. (2024)	≈	✗	✗	≈
Matz et al. (2023)	✓	≈	≈	✓
Mosia (2025)	≈	✗	≈	✗
Mustofa et al. (2025)	≈	≈	≈	≈
Nagy and Molontay (2024)	≈	≈	≈	✗
Olaya et al. (2020)	✗	✗	✗	≈
Ortigosa et al. (2019)	✗	≈	✓	✗
Pan et al. (2024)	✓	✗	≈	✓
Pek et al. (2023)	✗	✓	≈	≈
Rabelo and Zárate (2025)	✗	≈	✗	≈
Rebelo Marcolino et al. (2025)	✗	≈	✓	≈
Romero and Liao (2025)	≈	✓	≈	≈
Roy and Farid (2024)	✓	✓	✓	≈
Skittou et al. (2024)	✗	≈	≈	≈
Sonnleitner et al. (2025)	✗	≈	✗	≈
Tong and Li (2025)	≈	≈	≈	≈
Van Petegem et al. (2023)	✗	≈	≈	≈
Vives et al. (2024)	≈	✓	≈	≈
Wen and Juan (2023)	✓	✗	✗	≈
Wong et al. (2025)	≈	✗	✗	✗
Zanellati et al. (2024)	≈	✗	≈	✗
Zhang et al. (2025)	≈	✓	≈	≈
Zhidkikh et al. (2024)	✗	≈	≈	≈

6. Future research directions for practical learning analytics

For researchers investigating student dropout or performance prediction, we recommend drawing on prior work to conduct benefit studies. In cases where a specific adjustment or tool is under examination, we find benefaction research to offer particular value by conserving resources and reducing methodological heterogeneity across studies. Where a novel model is proposed, our proposed six-layered framework presented in this paper underscores the importance of transparency and reproducibility, as these enable more rigorous and targeted investigation. This includes prioritizing the use of public datasets or, where feasible, publishing anonymized versions of private datasets. We further recommend the routine inclusion of computational performance metrics, such as training time and inference time, as standard performance evaluation metrics.

Computational feasibility should be assessed, especially when complex models or additional tools, such as interpretability techniques, are incorporated, and when scalability is a concern. As interpretability is a core concept within the framework, it should be integrated from the outset and considered throughout the research process. Similarly, actionability should be evaluated at each stage of model design, although detailed analysis may be more achievable in benefaction studies due to resource limitations. Finally, we encourage future research to incorporate considerations of generalizability and scalability during dataset selection. Using diverse, publicly available datasets can reduce early-stage model specialization and support more robust real-world applicability.

7. Conclusion and implications for practice

This review paper introduced a novel framework for guiding research on predictive models for student dropout. The framework promotes a full-cycle approach that prioritizes practicality over prediction accuracy alone. The findings of the literature review establish clear criteria to strengthen future research methodologies and help bridge the divide between theoretical studies and practical implementation. Through this, our proposed six-layered framework aims to support more comprehensive investigation of models during the research phase, ultimately contributing to the development of models that are viable in a real-world context. Our systematic evaluation of the literature using these criteria revealed substantial heterogeneity in current research practices and a persistent gap between model development and model deployment. By adopting the recommended adjustments, future research can enhance transparency, encourage deeper investigation, and improve opportunities for advancement in the field. In doing so, researchers can help reduce implementation barriers and ultimately deliver greater benefit to institutional staff and students.

List of abbreviations

AdaBoost	Adaptive boosting
AFSA	Adaptive feature selection algorithm
ANN	Artificial neural networks
AUC	Area under the curve
BN	Bayesian network
CART	Classification and regression trees
CatBoost	Categorical boosting
CHAID	Chi-squared automatic interaction detection
CTS	Contextual treatment selection
DFFNN	Deep feedforward neural networks
DNN	Deep neural network
DT	Decision trees
EBM	Explainable boosting classifier
ECNN	Enhanced convolution neural networks
ED	Euclidean distance
EDM	Educational data mining
FATE	Fairness, accountability, transparency, and ethics
FDT	Fast decision trees
FFMLP	Feed-forward multilayer perceptron
FNN	Feedforward neural network
FS	Forward selection
FTT	Feature tokenizer transformer
GAM	Generalized additive model
GBC	Gradient boosting classifier
GNN-TINet	Graph neural network transformer-inceptionnet
HDL	Hybrid deep learning model
HLRNN	Hybrid logistic regression and neural network
ID3	Iterative dichotomizer 3
KL	Kullback-leibler divergence
KNN	K-nearest neighbor
LA	Learning analytics
LASSO	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
LightGBM	Light gradient boosting machine
LIME	Local interpretable model-agnostic explanations
LMS	Learning management system
LR	Logistic regression
LSTM	Long short-term memory
LVAE	Latent variational autoencoder
ML	Machine learning
MLP	Multilayer perceptron
MODDQN	Deep q-network in multiple-objective Markov decision process
MOOCs	Massive open online courses
NB	Naïve bayes
NGBoost	Natural gradient boosting
NSGA-II	Non-dominated sorting genetic algorithm
OULAD	Open university learning analytics dataset
QDA	Quadratic discriminant analysis
QUEST	Quick, unbiased, and efficient statistical tree
ResNetV2	Residual network v2
RF	Random forest
RQ	Research question
SGD	Stochastic gradient descent
SHAP	Shapley additive explanations
SJR	SCImago journal and country rank
SLR	Simple linear regression
SVM	Support vector machine
t-SNE	t-distributed stochastic neighboring embedding
XAI	Explainable artificial intelligence
XGBoost	Extreme gradient boosting

Acknowledgment

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under grant no. (GPIP: 282- 830-2024). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Adejo OW and Connolly T (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1): 61-75. https://doi.org/10.1108/JARHE-09-2017-0113

Alamuddin R, Rossman D, and Kurzweil M (2019). Interim findings report from the MAAPS advising experiment. ITHAKA S+ R Report, Ithaka S+R, New York, USA. https://doi.org/10.18665/sr.311567

Arnold KE and Pistilli MD (2012). Course signals at Purdue: Using learning analytics to increase student success. In the Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Association for Computing Machinery, Vancouver, Canada: 267–270. https://doi.org/10.1145/2330601.2330666 **PMid:22496114**

Barbeiro L, Gomes A, Correia FB, and Bernardino J (2024). A review of educational data mining trends. Procedia Computer Science, 237: 88-95. https://doi.org/10.1016/j.procs.2024.05.083

Collberg C and Proebsting TA (2016). Repeatability in computer systems research. Communications of the ACM, 59(3): 62-69. https://doi.org/10.1145/2812803

Cortez P and Silva A (2008). Student performance. UCI Machine Learning Repository, Irvine, USA.

Delen D, Davazdahemami B, and Rasouli Dezfouli E (2024). Predicting and mitigating freshmen student attrition: A local-explainable machine learning framework. Information Systems Frontiers, 26: 641-662. https://doi.org/10.1007/s10796-023-10397-3 **PMid:37361887 PMCid:PMC10097523**

Du X, Yang J, and Hung JL (2020). An integrated framework based on latent variational autoencoder for providing early warning of at-risk students. IEEE Access, 8: 10110-10122. https://doi.org/10.1109/ACCESS.2020.2964845

Fahd K, Venkatraman S, Miah SJ, and Ahmed K (2022). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. Education and Information Technologies, 27: 3743-3775. https://doi.org/10.1007/s10639-021-10741-7

Gundersen OE (2021). The fundamental principles of reproducibility. Philosophical Transactions of the Royal Society A, 379(2197): 20200210. https://doi.org/10.1098/rsta.2020.0210 **PMid:33775150**

Hoca S and Dimililer N (2025). A machine learning framework for student retention policy development: A case study. Applied Sciences, 15(6): 2989. https://doi.org/10.3390/app15062989

Hoffait AS and Schyns M (2017). Early detection of university students with potential difficulties. Decision Support Systems, 101: 1-11. https://doi.org/10.1016/j.dss.2017.05.003

Huang H, Yuan S, He T, and Hou R (2022). Use of behavior dynamics to improve early detection of at-risk students in online courses. Mobile Networks and Applications, 27: 441-452. https://doi.org/10.1007/s11036-021-01844-z

Kuzilek J, Hlosta M, and Zdrahal Z (2017). Open university learning analytics dataset. Scientific Data, 4: 170171. https://doi.org/10.1038/sdata.2017.171 **PMid:29182599 PMCid:PMC5704676**

Latif G, Abdelhamid SE, Fawagreh KS, Brahim GB, and Alghazo R (2023). Machine learning in higher education: Students' performance assessment considering online activity logs. IEEE Access, 11: 69586-69600. https://doi.org/10.1109/ACCESS.2023.3287972

Loyola-González O (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access, 7: 154096-154113. https://doi.org/10.1109/ACCESS.2019.2949286

Maniyan S, Ghousi R, and Haeri A (2024). Data mining-based decision support system for educational decision makers: Extracting rules to enhance academic efficiency. Computers and Education: Artificial Intelligence, 6: 100242. https://doi.org/10.1016/j.caeai.2024.100242

Masood JAIS, Chakravarthy NK, Asirvatham D, Marjani M, Shafiq DA, and Nidamanuri S (2024). A hybrid deep learning model to predict high-risk students in virtual learning environments. IEEE Access, 12: 103687-103703. https://doi.org/10.1109/ACCESS.2024.3434644

Matz SC, Bukow CS, Peters H, Deacons C, Dinu A, and Stachl C (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. Scientific Reports, 13: 5705. https://doi.org/10.1038/s41598-023-32484-w **PMid:37029155 PMCid:PMC10082180**

Memarian B and Doleck T (2023). Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. Computers and Education: Artificial Intelligence, 5: 100152. https://doi.org/10.1016/j.caeai.2023.100152

Molla-Esparza C, Gómez-Núñez MI, and García-García FJ (2025). Applications of learning analytics in the study of academic performance in higher education: A pilot-tested meta-review protocol. International Journal of Educational Research Open, 8: 100433. https://doi.org/10.1016/j.ijedro.2024.100433

Mosia M (2025). A Bayesian state-space approach to dynamic hierarchical logistic regression for evolving student risk in educational analytics. Data, 10(2): 23. https://doi.org/10.3390/data10020023

Mustofa S, Emon YR, Mamun SB, Akhy SA, and Ahad MT (2025). A novel AI-driven model for student dropout risk analysis with explainable AI insights. Computers and Education: Artificial Intelligence, 8: 100352. https://doi.org/10.1016/j.caeai.2024.100352

Nabil A, Seyam M, and Abou-Elfetouh A (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. IEEE Access, 9: 140731-140746. https://doi.org/10.1109/ACCESS.2021.3119596

Nagy M and Molontay R (2024). Interpretable dropout prediction: Towards XAI-based personalized intervention. International Journal of Artificial Intelligence in Education, 34: 274-300. https://doi.org/10.1007/s40593-023-00331-8

Olaya D, Vásquez J, Maldonado S, Miranda J, and Verbeke W (2020). Uplift modeling for preventing student dropout in higher education. Decision Support Systems, 134: 113320. https://doi.org/10.1016/j.dss.2020.113320

Ortigosa A, Carro RM, Bravo-Agapito J, Lizcano D, Alcolea JJ, and Blanco O (2019). From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system. IEEE Transactions on Learning Technologies, 12(2): 264-277. https://doi.org/10.1109/TLT.2019.2911608

Pan F, Zhang H, Li X, Zhang M, and Ji Y (2024). Achieving optimal trade-off for student dropout prediction with multi-objective reinforcement learning. PeerJ Computer Science, 10: e2034. https://doi.org/10.7717/peerj-cs.2034 **PMid:38855215 PMCid:PMC11157558**

Pek RZ, Özyer ST, Elhage T, Özyer T, and Alhajj R (2023). The role of machine learning in identifying students at-risk and minimizing failure. IEEE Access, 11: 1224-1243. https://doi.org/10.1109/ACCESS.2022.3232984

Rabelo AM and Zárate LE (2025). A model for predicting dropout of higher education students. Data Science and Management, 8(1): 72-85. https://doi.org/10.1016/j.dsm.2024.07.001

Raghupathi W, Raghupathi V, and Ren J (2022). Reproducibility in computing research: An empirical study. IEEE Access, 10: 29207-29223. https://doi.org/10.1109/ACCESS.2022.3158675

Ramaswami G, Susnjak T, Mathrani A, and Umer R (2023). Use of predictive analytics within learning analytics dashboards: A review of case studies. Technology, Knowledge and Learning, 28: 959-980. https://doi.org/10.1007/s10758-022-09613-x

Rebelo Marcolino M, Reis Porto T, Thompsen Primo T, Targino R, Ramos V, Marques Queiroga E, Munoz R, and Cechinel C (2025). Student dropout prediction through machine learning optimization: Insights from moodle log data. Scientific Reports, 15: 9840. https://doi.org/10.1038/s41598-025-93918-1 **PMid:40119104 PMCid:PMC11928464**

Romero S and Liao X (2025). Statistical and machine learning models for predicting university dropout and scholarship impact. PLOS ONE, 20(6): e0325047. https://doi.org/10.1371/journal.pone.0325047 **PMid:40560971 PMCid:PMC12193850**

Roy K and Farid DM (2024). An adaptive feature selection algorithm for student performance prediction. IEEE Access, 12: 75577-75598. https://doi.org/10.1109/ACCESS.2024.3406252

Sailer M, Ninaus M, Huber SE, Bauer E, and Greiff S (2024). The end is the beginning is the end: The closed-loop learning analytics framework. Computers in Human Behavior, 158: 108305. https://doi.org/10.1016/j.chb.2024.108305

Sghir N, Adadi A, and Lahmer M (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). Education and Information Technologies, 28: 8299-8333. https://doi.org/10.1007/s10639-022-11536-0 **PMid:36571084 PMCid:PMC9765383**

Siemens G (2013). Learning analytics: The emergence of a discipline. American Behavioral Scientist, 57(10): 1380-1400. https://doi.org/10.1177/0002764213498851

Skittou M, Merrouchi M, and Gadi T (2024). Development of an early warning system to support educational planning process by identifying at-risk students. IEEE Access, 12: 2260-2271. https://doi.org/10.1109/ACCESS.2023.3348091

Sonnleitner B, Madou T, Deceuninck M, Theodosiou F, and Sagaert YR (2025). Evaluation of early student performance prediction given concept drift. Computers and Education: Artificial Intelligence, 8: 100369. https://doi.org/10.1016/j.caeai.2025.100369

Tinto V (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45(1): 89-125. https://doi.org/10.3102/00346543045001089

Tong T and Li Z (2025). Predicting learning achievement using ensemble learning with result explanation. PLOS ONE, 20(1): e0312124. https://doi.org/10.1371/journal.pone.0312124 **PMid:39745993 PMCid:PMC11694977**

Topali P, Ortega-Arranz A, Rodríguez-Triana MJ, Er E, Khalil M, and Akçapınar G (2025). Designing human-centered learning analytics and artificial intelligence in education solutions: A systematic literature review. Behaviour & Information Technology, 44(5): 1071-1098. https://doi.org/10.1080/0144929X.2024.2345295

Vahdat M, Oneto L, Anguita D, Funk M, and Rauterberg M (2015). Educational process mining (EPM): A learning analytics data set. UCI Machine Learning Repository, Irvine, USA.

Van Petegem C, Deconinck L, Mourisse D, Maertens R, Strijbol N, Dhoedt B, De Wever B, Dawyndt P, and Mesuere B (2023). Pass/fail prediction in programming courses. Journal of Educational Computing Research, 61(1): 68-95. https://doi.org/10.1177/07356331221085595

Vives L, Cabezas I, Vives JC, Reyes NG, Aquino J, Cóndor JB, and Altamirano SFS (2024). Prediction of students' academic performance in the programming fundamentals course using long short-term memory neural networks. IEEE Access, 12: 5882-5898. https://doi.org/10.1109/ACCESS.2024.3350169

Wen X and Juan H (2023). Early prediction of students' performance using a deep neural network based on online learning activity sequence. Applied Sciences, 13(15): 8933. https://doi.org/10.3390/app13158933

West D, Huijser H, and Heath D (2016). Putting an ethical lens on learning analytics. Educational Technology Research and Development, 64: 903-922. https://doi.org/10.1007/s11423-016-9464-3

Wong A, Lee WL, Chan MSL, Tan YE, Huang JMK, and Lee YH (2025). Digital learning resources and student success: Analyzing engagement and academic performance. Journal of Applied Learning & Teaching, 8(S2): 45-54. https://doi.org/10.37074/jalt.2025.8.S2.3

Xiao W and Hu J (2023). A state-of-the-art survey of predicting students' performance using artificial neural networks. Engineering Reports, 5(8): e12652. https://doi.org/10.1002/eng2.12652

Zanellati A, Zingaro SP, and Gabbrielli M (2024). Balancing performance and explainability in academic dropout prediction. IEEE Transactions on Learning Technologies, 17: 2086-2099. https://doi.org/10.1109/TLT.2024.3425959

Zhang X, Zhang Y, Chen AL, Yu M, and Zhang L (2025). Optimizing multi label student performance prediction with GNN-TINet: A contextual multidimensional deep learning framework. PLOS ONE, 20(1): e0314823. https://doi.org/10.1371/journal.pone.0314823 **PMid:39841673 PMCid:PMC11753673**

Zhidkikh D, Heilala V, Van Petegem C, Dawyndt P, Jarvinen M, Viitanen S, and Hämäläinen R (2024). Reproducing predictive learning analytics in CS1: Toward generalizable and explainable models for enhancing student retention. Journal of Learning Analytics, 11(1): 132-150. https://doi.org/10.18608/jla.2024.7979