WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Interpreting Early Student Dropout Prediction Claims: A Critical Review of Task Definition, Validation, and Explainability

Prakher Katiyar [1], Dr. Gaurav Srivastava [2], Dr. Gaurav Srivastava [3]

[1, 2] Student, [3]Guide

Bansal Institute of Engineering & Technology, APJ Abdul Kalam Technical University AKTU

ABSTRACT

Predicting student dropout is considered an effective method for identifying learners who could benefit from timely academic support. However, the scores generated by a predictive model are only practically useful when the prediction task, the information available at the time, the target group, the purpose of validation, and the decision-making context are clearly defined. This research paper presents a critical review of selected open-access literature, encompassing both general overviews and methodological studies. This paper discusses topics such as estimating student risk, 'data leakage' in machine learning, reporting the accuracy of predictions, and model interpretability. Given that existing literature already provides comprehensive reviews on identifying 'high-risk' students in higher education and addressing the issue of dropouts—with a time-consistent framework for longitudinal dropout prediction also having been recently proposed—the objective of this research paper is not to introduce an entirely new analytical framework for 'data leakage'. Instead, this research paper provides a guide for reviewers, focusing on distinguishing between various types of issues—such as data leakage across feature or time dimensions; mismatches between prediction targets and risk sets; overlaps caused by duplicate entities; temporal validation mismatches; inadequate performance reporting; and the over-interpretation of feature attributions. This paper offers four practical contributions: a comparative analysis of existing reviews, a taxonomy of evaluation-related challenges, a comparison of validation design approaches, and an audit checklist for authors and reviewers. The analysis emphasizes that random cross-validation is not inherently incorrect; its appropriateness depends on the specific objectives of its application and the nature of the dependency structures present within the data. The paper further explains that when addressing issues such as class imbalance, model calibration, the potential for intervention, or threshold-based decision-making, focusing solely on "discrimination" is insufficient. While SHAP and similar attribution tools can describe the behavior of a fitted model, they do not establish causal relationships. Nor do they prove that a specific intervention has actually improved student outcomes. This work is a targeted critical review rather than an exhaustive systematic review; it does not involve comprehensive database searches, dual-reviewer screening, formal quality assessment, or estimates of the prevalence of the phenomena discussed. Therefore, the list provided in this paper should be viewed as a practical reference—rather than an empirically tested standard—that still awaits external validation.

Keywords: student dropout prediction; educational data mining; data leakage; temporal validation; risk set; model calibration; explainable artificial intelligence

Introduction

Higher education institutions are now using academic records, administrative variables, and activity data from learning platforms to identify students who might benefit from advising, tutoring, financial aid, or other forms of support. While predictive analytics can help institutions prioritize support interventions, high accuracy—or a large area under the Receiver Operating Characteristic (ROC) curve (AUC)—does not, by itself, demonstrate a model's true practical utility. Models should generate forecasts of future outcomes using only the information available at the time the prediction is made. Furthermore, when evaluating these models, it is essential to consider the target group and the specific context in which they will be applied (Alyahyan & Düştegör, 2020; Duro et al., 2026; Shi et al., 2025).

In academic datasets, data pertaining to a single learner may appear repeatedly across different weeks, course modules, or semesters. This data structure can lead to overly optimistic evaluation results. For instance, feature variables might inadvertently include information that was not available at the time of prediction; target outcomes may encompass various scenarios—such as dropping out, becoming inactive, failing a course, or leaving the institution; learners who have already experienced a specific event might still be classified as being "at risk" of that same event in the future; or data from the same learner could appear in both the training and testing sets. Furthermore, the "generalizability" estimated through random data splitting may differ significantly from the results obtained when testing on data from subsequent semesters. Although these issues are interconnected, they fall into distinct methodological categories and require different approaches for investigation. Research on human-machine learning across various scientific fields indicates that "data leakage" can undermine the reproducibility of research and lead to an overly optimistic assessment of model performance (Kapoor & Narayanan, 2023).

Recent research literature also emphasizes the need for careful clarification of one's position when introducing methodological contributions. Shi et al. (2025) conducted a comprehensive analysis of 161 studies on "risk prediction" in online education; Duro et al. (2026) reviewed 61 studies that utilized machine learning and deep learning methods to predict "dropouts" (students leaving their studies prematurely) in higher education; and López-Muñoz et al. (2026) carried out a systematic review, which ultimately included 75 studies that passed a quality assessment. Fauszt (2026) specifically explored several key elements essential for "longitudinal" dropout prediction. These include cut-off points for prediction, temporal constraints on predictor information, the continuity of at-risk groups, and a hierarchical framework for validation strategies. Therefore, this research paper does not claim that these methodological elements are entirely novel. Rather, it brings together aspects related to task definition, validation design, interpretation of metrics, and explanatory discussion. Its objective is to develop a concise evaluation framework for examining research on 'early warning systems' for student dropout.

This review aims to address the following three research questions: (1) What minimum information is required to understand and evaluate the findings of studies focused on the early prediction of student dropout? (2) How can issues such as "feature-time leakage" (the inappropriate inclusion of temporal information in the data), mismatches between target outcomes and the identified at-risk group, learner overlap, and the generalizability of outcomes across different time periods be identified and evaluated? (3) What evidence is necessary to demonstrate that a model's predictive performance and its explanations are sufficiently reliable and relevant to serve as a valid basis for decision-making regarding student support?

Review approach and scope

- **Review design**

This article adopts a critical, descriptive approach, utilizing a purposive selection of available academic literature. Through targeted searches and reviews on publishers' websites, the authors identified open-access review articles concerning the prediction of student risk and dropout. Additionally, they included literature addressing specific methodological topics, such as data leakage, validation based on repeated observations, evaluation metrics for imbalanced datasets, model calibration, and feature attribution. The selected literature is directly relevant to the research questions posed in this review and can be located and verified via publishers' or repository records. These sources encompass broad reviews in the field of educational prediction, recently proposed frameworks that account for data leakage, and methodological references spanning various disciplines.

The literature selection process described here cannot be classified as a 'systematic review.' This study did not employ a pre-registered protocol, nor did it involve an exhaustive database search, independent screening by two reviewers, the use of formal tools to assess 'risk of bias,' or the maintenance of a comprehensive record of excluded studies. Furthermore, the search did not include subscription-based databases. Additionally, this article does not provide a PRISMA flow diagram, screening statistics, or pooled estimates, nor does it specify the number of studies that contained particular methodological shortcomings. Since PRISMA 2020 is a reporting guideline for systematic reviews, it cannot serve as a substitute for the actual execution and documentation of a systematic search process (Page et al., 2021). Consequently, the findings presented in this article focus on interpretive and methodological aspects, rather than providing quantitative estimates regarding the prevalence of various types of shortcomings.

- **Scope and definitions**

This study focuses on predicting student 'dropouts'—or related 'high-risk' outcomes—within the context of higher education and online learning. Terms such as 'dropout,' 'withdrawal,' 'failure to complete,' 'academic failure,' 'retention,' and 'lack of engagement' are viewed as outcomes that are both interconnected and distinct from one another. Existing literature is referenced solely to define the scope of this study or to provide evidence in support of specific arguments. Furthermore, the datasets mentioned (e.g., the "Open University Learning Analytics Dataset") are cited only to illustrate why data from long-term activities raise questions regarding 'significant time points'; this should not be interpreted as an implication that any specific model or dataset is inherently flawed (Kuzilek et al., 2017).

What existing literature establishes—and what remains to be interpreted

Existing reviews cover various aspects, such as model families, data sources, risk factors, interpretability, and deployment. Due to significant diversity in prediction tasks, datasets, definitions of events, and validation protocols, the findings from these reviews cannot be consolidated into a single, unified ranking of algorithms. For instance, a model evaluated based on a static data snapshot taken at the end of a term cannot be compared to a model evaluated using continuous student records collected over several weeks. This research paper highlights a practical distinction: on one hand, documenting which models were used, and on the other, examining whether the reported scores actually measure the prediction tasks as defined in the respective studies.

Table 1. Selected open-access reviews and methodological sources relevant to the critical synthesis

Source	Scope or contribution reported by the source	Critical implication for interpreting student-risk models
Alyahyan and Düştegör (2020)	Literature review of approaches to predicting academic success in higher education.	Prediction quality should be read alongside the outcome, predictors, dataset, and decision context; algorithm lists alone do not establish generalizability.

Source	Scope or contribution reported by the source	Critical implication for interpreting student-risk models
Shi et al. (2025)	Systematic review of 161 studies published from 2014 to 2024 on at-risk student prediction in online education.	Online-learning evidence is valuable but should not automatically be generalized to every campus, course, or institution.
Duro et al. (2026)	Review of 61 studies on machine-learning and deep-learning approaches to higher-education dropout prediction.	Model family, data handling, interpretability, and deployment constraints need to be discussed together.
Lopez-Muñoz et al. (2026)	Systematic review that identified 301 records and retained 75 quality-assessed studies, covering methods and risk factors.	Risk-factor synthesis can inform hypotheses, but predictive association is not proof of causality or intervention benefit.
Fauszt (2026)	Time-consistent framework and empirical demonstration addressing cutoffs, pre-cutoff information, risk sets, entity separation, and cohort-time validation.	This is direct conceptual overlap with leakage-aware longitudinal evaluation; the present article does not claim those concepts as novel.
Kapoor and Narayanan (2023)	Cross-disciplinary analysis of machine-learning leakage and reproducibility problems	Leakage is a methodological validity issue, not merely a coding nuisance; it warrants explicit reporting.
Kuzilek et al. (2017)	Open University Learning Analytics Dataset description, including student demographics, assessment results, and dated learning-platform activity.	Longitudinal event records require careful definition of when an activity summary becomes observable.

The purpose of this comparison is to highlight the complementarity between different perspectives, rather than attempting to construct an exhaustive picture that encompasses every single aspect of the field. Review studies reveal common patterns across various methods and risk factors, while methodology-focused studies clarify how the design of predictive tasks alters the meaning of the resulting scores. Fauszt's (2026) study is highly relevant to the core issues addressed in this article, and contributors should explicitly point out this overlap. The contribution of this paper lies in providing a refined set of guidelines for interpretation, which systematically integrates task definition, model validation, indicator selection, and the limitations of interpretive findings—it does not aim to propose a new formal framework, nor is it an empirical replication study.

Defining the prediction task before assessing the model

- **Prediction cutoff and horizon**

If a system claims to be able to predict student withdrawals "in advance," it must specify a clear timeframe. A research paper should explicitly state when the prediction is made, what outcomes (or time periods) are being predicted, and what subsequent actions should be taken. Predicting final outcomes based on data available at the end of a course is fundamentally different from predicting withdrawals using records available as early as the fourth week. If multiple time points are evaluated, there must be clear criteria regarding the availability of data (features) for each point, as well as the specific student population to which these criteria apply. A model that performs well in retrospective classification—using historical data—will not necessarily function effectively as an "early warning system"; the research title and conclusions must align with the specific task being evaluated (Fauszt, 2026).

- **Feature availability and leakage**

"Feature-time leakage" occurs when predictor variables incorporate information that should not be available at the time a prediction is made. This issue can be concealed behind seemingly innocuous feature names; for instance, a feature labeled "weekly engagement" might represent the number of clicks recorded by the end of a specific week, or it could be a cumulative value calculated retrospectively using data from an entire academic term. When the objective is to generate early predictions about future outcomes, features such as "end-of-course total," "final grade," "course withdrawal codes," or "post-event status" must be scrutinized with extreme care. A "feature dictionary" should clearly document the meaning, source, observation timestamp (or time frame), and aggregation rules for every variable. Furthermore, it should specify which features are available at various prediction cutoff points.

Data leakage can also arise during the data preprocessing stage. Processes such as imputation (filling in missing values), data scaling, feature selection, resampling, and hyperparameter tuning must be performed exclusively using the "training data"—without incorporating any information from the "test set" (which is held aside). Research on various forms of data leakage indicates that such errors can occur at any stage of the modeling pipeline. Often, these issues are difficult to detect by merely examining the final evaluation metrics (Kapoor & Narayanan, 2023). Instead of reporting only the final predictive model, researchers should document the entire modeling process.

- **Risk set, target, and censoring**

A "risk set" refers to the group of individuals who, at a given prediction cutoff point, are eligible to experience a specific future event. Predicting "failure to complete one's studies" among all enrolled students is distinct from predicting "subsequent dropout" among students who are actively pursuing their studies at a specific point in time. In a prospective analysis, students for whom the event has already occurred prior to the cutoff point are no longer eligible to experience that event subsequently. When conducting research in this field, the outcome metric must be clearly defined—for instance, does it refer to "withdrawing from a course," "leaving the institution," "academic failure," "failure to complete one's studies," or a proxy indicator (such as "prolonged inactivity")? Furthermore, the definition should specify the date of the event and the prediction time horizon. It should also be clarified how instances such as transfers, changes in coursework, delays in academic progress, and "censored observations" are handled.

This is not merely a subtle difference in terminology. Varying definitions of the target alter the population under consideration, the labels assigned, and the types of decisions that the model can effectively support. A model designed to predict the "administrative status" of students at the end of a term should not be conflated with a model designed to predict "future dropout" among students who are currently enrolled. Therefore, when evaluating the outcomes of such activities, they should not be considered as estimates based on a single metric, nor should they be directly compared with one another.

Table 2. Evaluation threats, diagnostic questions, and corrective checks

Threat	Diagnostic question	Recommended check
Feature-time leakage	Could any predictor value or aggregate include information recorded after the prediction cutoff?	Report cutoff and feature windows; audit source timestamps and aggregation logic; exclude post-cutoff fields.
Preprocessing leakage	Were imputation, scaling, feature selection, resampling, or tuning influenced by test data?	Fit all data-dependent transformations within training folds; reserve a final untouched test set.
Target/risk-set mismatch	Were learners included after the target event had already occurred, or were distinct outcomes conflated?	Define event, eligibility, horizon, censoring, transfers, and non-completion rules before modelling.
Entity overlap	Can records from the same student occur in training and test partitions?	Report partition unit and overlap check; group records when the deployment question requires unseen learners.
Temporal mismatch	Does evaluation test the same time direction and cohort shift expected in deployment?	Use chronology-aware evaluation where future-term performance is the target; report drift and cohort changes.
Metric mismatch	Does the reported metric reflect event prevalence and the cost/capacity of acting on predictions?	Report multiple complementary measures, uncertainty, class prevalence, calibration, and threshold or capacity assumptions.
Explanation overreach	Is attribution being interpreted as a cause or as evidence that an intervention works?	State the explanation estimand and robustness checks; separate prediction explanation from causal and intervention claims.

Validation design must match the generalization target

There is no single method for splitting data into training and test sets that is appropriate for every situation. Random cross-validation is suitable when observations are independent of one another, the deployment environment mirrors the sampling environment, and the temporal order of data points

is irrelevant to the conclusions being drawn. However, this approach can yield misleading results if the recurring records contain information pertinent to the learning process, if the primary objective is to make predictions about the future, or if subsequent batches of data differ from earlier ones. Most importantly, you should clearly state what is being estimated and explain why that estimate is relevant to your intended application (Fauszt, 2026; Kapoor & Narayanan, 2023).

Table 3. Validation designs and the questions they can answer

Design	Question it can help answer	Main caveat
Random row-wise split / ordinary CV	How well does the model predict randomly held-out rows from a similar data-generating process?	May leak learner identity across partitions with repeated records; does not itself test later-cohort performance.
Learner-grouped split	How well does the model generalize to learners not represented in training?	If rows from different periods are otherwise mixed, it does not prove future-term generalization.
Chronological split	How well does a model trained on earlier records perform on later records?	Repeated learners may still cross partitions; time drift may be entangled with changes in cohort composition.
Grouped and chronological split	How well does the model perform on later periods for learners held out from model development?	Often more deployment-relevant, but may reduce sample size and make results noisier.
External institution/course validation	Does performance transfer to a different institution, course, or educational context?	Differences in grading, course design, support systems, and data collection require careful interpretation.

Group-based validation and chronological validation are designed to address different types of problems. Grouping helps control for the overlap of entities (samples), while chronological validation assesses whether training data from earlier time periods can predict outcomes in subsequent periods. A robust evaluation involving "future groups" often requires a combination of both approaches. External validation provides stronger evidence of a model's generalizability, provided that the definitions of outcomes and the temporal reference points for characteristics used by the external institution are consistent with those in the original study. Authors should disclose the following information: the entity used for data partitioning, the chronological arrangement, the handling of overlapping entities, the scope of preprocessing, the parameter-tuning process, and whether the final test set influenced any modeling decisions.

Performance reporting: beyond a headline AUC

Different evaluation metrics are designed to answer different questions. 'Discrimination' assesses whether individuals who experience a particular event (e.g., students at risk of dropping out) are assigned higher risk scores compared to those who do not; 'calibration' indicates how well predicted probabilities align with the observed frequency of outcomes; and 'threshold-based metrics' describe the results of applying specific decision rules. In the context of 'early-warning' systems, educational institutions may have limited support staff or a restricted capacity to reach out to students. Therefore, the primary practical concern is often determining how many of the students identified as 'high-risk' are truly at risk, rather than simply looking at the overall AUC value.

When an event—such as dropping out—occurs relatively infrequently, 'precision-recall curves' and 'positive predictive value' (PPV) provide more useful information than 'ROC curves' alone. These metrics help determine the proportion of individuals classified as "positive" (i.e., at risk) who are, in fact, at risk (Saito & Rehmsmeier, 2015). The choice of metric depends on the frequency of the event and the specific context in which it is being used; consequently, no single "universal" metric should be considered the definitive standard in this field. At a minimum, research should report the following: the prevalence of the outcome, discrimination metrics, sensitivity/recall, precision/positive predictive value, evaluations based on specific thresholds, and—where possible—an assessment of uncertainty. If a model provides probabilities, or if its outputs are used to prioritize interventions, then 'calibration' must be explicitly evaluated. Even if the ranking is correct, poor calibration can render risk classification and subsequent decision-making unreliable (Van Calster et al., 2019).

Another critical issue concerns the selection of metrics following multiple experimental iterations. If the same 'test set' is used for feature selection, hyperparameter tuning, model comparison, and final performance reporting, the resulting scores no longer constitute valid "independent estimates." Researchers should keep the model development process separate from the final evaluation process and clearly explain how the final model and decision thresholds were determined. Unless different studies are largely comparable in terms of outcome definitions, data-splitting strategies, and evaluation

metrics, any observed differences in their results should not be interpreted directly as an indication of the relative superiority or inferiority of the algorithms.

Explainability: interpretation, stability, and actionability are different claims

Feature-attribution methods like SHAP aim to measure the contribution of individual features to a specific prediction or to an entire dataset, given a specific model and background distribution (Lundberg and Lee, 2017). Such explanations are helpful in clarifying the behavior of the model, identifying unexpected dependencies on certain variables, and error analysis. However, these methods cannot by themselves determine the reasons behind student dropout. They cannot prove that changing any specific characteristic will necessarily change student outcomes, nor can they guarantee that interventions based on these explanations will improve educational outcomes.

When drawing explanatory conclusions, it is important to be clear about what is being explained: is it an individual prediction, the overall behavior of the model, patterns within a specific subgroup, or a proposed intervention? Furthermore, “robustness” is a multidimensional concept. Consistency across different random seeds is not the same as consistency across data folds, cohorts, subgroups, model architectures, or interpretation methods. Even when the classification ranking of attributes remains stable, the underlying relationships may be confusing or not actionable. In contrast, an unstable ranking indicates that interpretations at the attribute level are highly sensitive to choices made during the modeling process. Research in this area should clearly state which dimension of explanation stability is being tested and avoid confusing "explainability" with "causality" or "objectivity".

In institutional use, the quality of an explanation should be evaluated in the context of the specific decisions it supports, while also considering fairness, confidentiality, human monitoring, and the potential risk of stigmatizing students. Risk scores should only serve as a reference to provide support, and not as definitive, qualitative judgments about a learner. If it is claimed that an intervention helps improve student retention, specific evidence related to that intervention should be provided – ideally through a prospective evaluation design. Predictive performance metrics or explanatory visualizations alone cannot serve as a substitute for such evidence.

Reviewer-oriented audit checklist

The checklist provided below summarizes the issues discussed above. It is intended as a practical aid to facilitate review, rather than as a formal reporting standard. Reviewers may mark each point as “reported,” “partially reported,” or “not reported”; if the relevant actions or outcomes cannot be understood due to a lack of detail, reviewers may request further clarification.

Table 4. Minimum audit checklist for an early student-risk prediction study

Domain	Minimum information to report	Warning sign requiring clarification
Prediction task	Prediction cutoff, forecast horizon, target event, intended decision, and deployment setting.	The paper calls the task “early” but does not state when the prediction is issued.
Feature timing	Feature provenance, timestamp/aggregation window, availability rule, and preprocessing boundary.	Full-period aggregates or post-event information appear in an early-prediction model.
Risk set and outcome	Eligibility at each cutoff, event definition/date, censoring and handling of transfers/course changes.	Already-withdrawn learners remain at risk for a future event, or distinct outcomes are mixed without justification.
Data splitting	Split unit, train-test entity overlap, chronology, tuning strategy, and final test-set separation.	Repeated records cross partitions without considering the deployment target.
Generalization target	Whether evaluation targets new rows, new learners, future terms, new courses, or new institutions.	A random split is interpreted as proof of future or external validity.
Performance and uncertainty	Class prevalence, discrimination, precision/recall, threshold performance, uncertainty, and calibration where relevant.	One headline score is reported without prevalence or decision context.

Domain	Minimum information to report	Warning sign requiring clarification
Explainability	Attribution method, claim being made, assessed stability dimension, limitations, and causal caveat.	Feature importance is described as a cause or as evidence of intervention effectiveness.
Reproducibility and governance	Data provenance, code or sufficient pipeline detail, missing-data decisions, privacy, fairness, human oversight.	A reader cannot reconstruct the evaluation or understand how a score would be used responsibly.

This checklist has been developed with the intent of promoting careful consideration. If an item is marked as "not reported," it should not be automatically assumed that a "data leak" has occurred or that the study is invalid; rather, it simply means that, based on the content of the research paper, it cannot be determined whether the necessary criteria for validity have been met. Similarly, the use of random data splitting is not necessarily always incorrect. It is the reviewer's role to determine whether the chosen research design is appropriate for the prediction task and for the claims made in the paper regarding generalization.

Conceptual workflow for reviewing a prediction claim

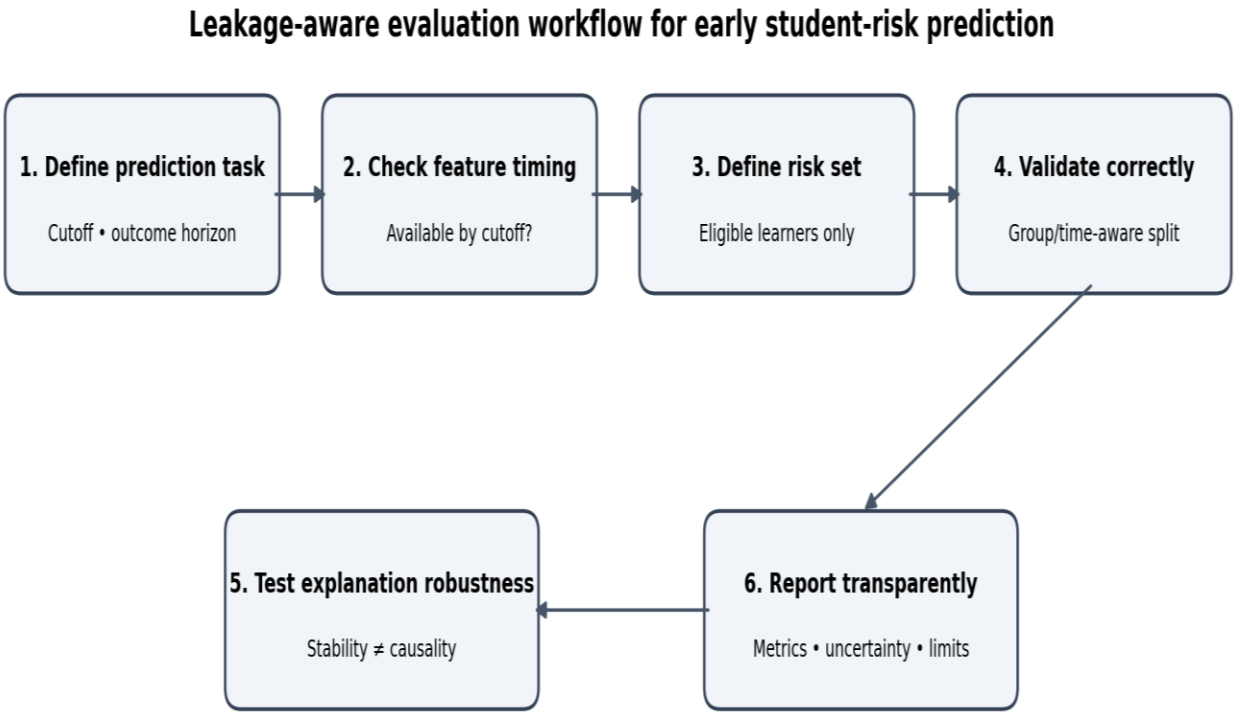


Figure 1. Conceptual workflow for reviewing early student-risk prediction claims.

Note: This workflow is a combination of the explanations and suggestions provided in the evaluation report; it is neither an empirically validated process, nor has it been evaluated as an intervention or a predictive model.

Implications and future research

- **Implications for authors and reviewers**

Before selecting an algorithm, authors should clearly define the prediction task and design a validation scheme aligned with their objectives. To enhance the quality of a research report, it should include the following: a table detailing feature availability, clear criteria for defining the risk set, a description or diagram of the data-splitting strategy, the incidence rate of the outcome, and a clear statement regarding the type of generalization being tested. If the study pertains to future academic terms, a realistic time-series validation approach should be employed rather than relying solely on random cross-validation. In scenarios involving learners who are part of the same course but have not yet been "seen" by the model, it may be necessary to perform data splitting at the entity (e.g., student) level. In any case, the validation scheme must be consistent with the stated objectives of the study.

Reviewers should distinguish between "a lack of documentation" and "flaws in methodology." For instance, if a paper fails to specify timestamps for features, a reviewer should not immediately assume that "future information" has been used; instead, they should ask the authors to clarify how the features were constructed and to which specific point in time they correspond. Conversely, if a variable that only emerges after a student has dropped out is used to predict the occurrence of that dropout, it is not merely a documentation oversight but a serious problem. Making such distinctions helps ensure that the review process remains both thoroughly critical and impartial.

- **An empirical agenda**

Future research should not be limited to merely presenting a new checklist; rather, it should evaluate whether the use of such a checklist improves the consistency of reviews and enhances the validity of the models. This can be tested through multi-rater studies, which assess the level of agreement among multiple evaluators who independently use the checklist to code published research. Benchmark analyses can be conducted using the same publicly available longitudinal datasets, employing various partitioning strategies—such as row-based, learner-based, or chronological partitioning, or a combination of learner-based and chronological grouping—with all feature windows pre-defined. Such studies should report key metrics—including AUC, precision-recall, calibration, and uncertainty—as well as variations in the number of overlapping learners across different partitions. These studies should follow a pre-established analysis plan and clearly distinguish between "synthetic demonstrations" and "results derived from actual student records."

A second research priority is to link predictive indicators with support for actionable decision-making. Studies should clarify an institution's capacity for intervention and assess whether risk-based rankings are effective in optimizing the allocation of "voluntary, non-punitive" support resources. Where possible, the evaluation of intervention effectiveness should be kept separate from the assessment of the models' predictive accuracy. A third priority is to report differences across various groups and cohorts, while ensuring strict adherence to privacy standards and avoiding unsubstantiated claims regarding the specific reasons why individuals might drop out or leave a study.

- **Future Scope**

Future research should investigate the practical value of "evaluation methods that account for data leakage" through reproducible studies based on students' long-term records. Researchers should pre-specify the prediction threshold (cutoff) and the outcome horizon, while also documenting the precise time points at which specific features become available. Where data structures and research questions permit, studies should compare multiple validation approaches—such as row-wise validation, student-grouped validation, institution-based validation, and chronological validation. Any observed differences in performance should be derived from empirical experiments and accompanied by uncertainty assessments; they should not be inferred solely from the arguments presented in this review.

Another avenue for research involves external and prospective validation across various institutions, programs, and student cohorts. Such studies can assess whether a model maintains robust calibration even when there are changes in curricula, student demographics, data collection practices, or institutional policies. In cases where repeated measurements are available, researchers should also examine "temporal drift" and establish mechanisms for monitoring and updating models before they are deployed for operational use.

Regarding explainability, research should move beyond the mere presentation of "feature-attribution diagrams" to evaluate the stability of these explanations, their variation across subgroups, and their practical utility within specific advisory or decision-making contexts. Assessments of fairness and privacy should be conducted within a robust governance framework; risk scores should serve as aids to professional judgment, not as a substitute for it. Furthermore, the audit checklist proposed in this article should be evaluated by multiple independent reviewers to determine whether it enhances the completeness of reporting and the consistency of evaluation results. These points represent research priorities suggested by this review, rather than established research findings.

Limitations

Rather than a rigorous systematic review, this article is a critical review based on a purposive selection of available literature. The literature covered is not exhaustive, and the selection process may have favored publications that are more widely indexed or easily accessible. This review does not incorporate dual screening, formal risk-of-bias assessments, or quantitative synthesis; consequently, it cannot provide a precise estimate of the prevalence of issues such as "data leakage," validation practices, or interpretability checks within this field. The checklist presented here has not been validated through expert consensus, inter-rater reliability testing, or empirical benchmarking. Many of the suggestions offered are methodological conclusions drawn from the general body of literature on machine learning validation; in practice, these may need to be adapted to the specific data structures and predictive objectives of individual organizations. These limitations are explicitly noted to ensure that readers do not interpret the findings presented here as more definitive or robust evidence than they actually are.

Conclusion

Predicting student dropout should not be framed merely as a classification problem based on performance metrics; rather, it should be viewed as a decision-making task that inherently involves a temporal dimension. A robust analysis requires several critical components: clearly defined cutoff points and prediction windows; the availability of feature data at specific time points; well-defined risk sets and event criteria; validation strategies aligned with generalization goals; and evaluation metrics that account for both class distribution and the requirements of practical decision-making. Issues such as "feature-time" leakage, mismatches between target definitions and risk sets, entity overlap, and temporal inconsistencies must be addressed individually,

as resolving one does not automatically resolve the others. Beyond a model's discriminative power, its calibration and threshold-based utility are equally important. Furthermore, while methods like SHAP can explain the behavior of a fitted model, they cannot clarify the underlying causal mechanisms or the impact of specific interventions. Although recent research has addressed many of these challenges, this paper does not propose a new framework; instead, it offers a refined, critical analysis of existing literature and provides an evaluation checklist for researchers. Before drawing broad conclusions, it is essential to conduct more systematic, replicable reviews and empirical evaluations using this checklist.

References:

1. Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. International Journal of Educational Technology in Higher Education, 17, Article 3. https://doi.org/10.1186/s41239-020-0177-7
2. Duro, B., Gomes, A., Correia, F. B., Borges, A. R., & Bernardino, J. (2026). Machine learning and deep learning for dropout prediction in higher education: A review. Computers, 15(3), Article 164. https://doi.org/10.3390/computers15030164
3. Fauszt, T. (2026). Time-consistent prediction in higher education: A framework for preventing data leakage in longitudinal models. Information, 17(6), Article 581. https://doi.org/10.3390/info17060581
4. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804
5. Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. Scientific Data, 4, Article 170171. https://doi.org/10.1038/sdata.2017.171
6. Lopez-Muñoz, G. L., Gonzalez Serrano, C., & Sanchez-Ferreira, C. (2026). Student dropout prediction in higher education: A systematic review of machine learning methods and risk factors. Journal of Information Technology Education: Research, 25, Article 20. https://doi.org/10.28945/5776
7. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d 76c43dfd28b67767-Abstract.html
8. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71
9. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432
10. Shi, H., Zhang, N., Caskurlu, S., & Na, H. (2025). Applications of machine learning for at-risk student prediction in online education: A 10-year systematic review of literature. Journal of Computer Assisted Learning, 41(4), Article e70058. https://doi.org/10.1111/jcal.70058
11. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, Article 230. https://doi.org/10.1186/s12916-019-1466-7