*sustainability*

MDPI

*Systematic Review*

# Artificial Intelligence for Predicting Student Dropout from Digital Traces in Higher Education: A Systematic Review of Methods, Explainability, and Fairness Gaps for Equitable and Sustainable Retention

**Marco Aravena [1,*], Alvaro Jeria [2], Ricardo Soto [3,*], Broderick Crawford [3] and Rodrigo Olivares [1]**

[1] Escuela de Ingeniería Informática, Universidad de Valparaíso, Valparaíso 2362905, Chile; rodrigo.olivares@uv.cl
[2] Escuela de Ingeniería Industrial, Universidad de Valparaíso, Valparaíso 2362905, Chile; alvaro.jeria@uv.cl
[3] Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile; broderick.crawford@pucv.cl
* Correspondence: marco.aravena@uv.cl (M.A.); ricardo.soto@pucv.cl (R.S.)

**Abstract**

Student dropout limits the equal access to tertiary education sought by Sustainable Development Goal 4, and universities increasingly use machine learning models trained on student digital traces to identify students at risk. Following PRISMA 2020, this systematic review examined how such models are built, explained, and reported. Peer-reviewed empirical studies published in 2015–2025 were retrieved from Scopus and Web of Science, excluding online-only education. After re-assessment with written rules and independent double coding, 19 reports were classified by prediction target, horizon, data source, model family, explainability, and validation, and audited for reporting completeness with 16 items adapted from TRIPOD+AI and for signals of analytical quality. The main finding concerns the target: although framed as dropout prediction, 15 reports predicted course failure, so the evidence concerns academic risk more than departure. Learning management system records appeared in 16, ten compared machine and deep learning in parallel, and 11 used explainability, mainly to describe global risk factors. Ethical approval was documented in 8 reports, none reported performance across protected subgroups, and 16 lacked temporal or external validation. These gaps concern what is reported rather than proving biased models, but they prevent institutions from verifying whether early-warning systems serve all students equally.

**Keywords:** artificial intelligence; higher education; machine learning; dropout prediction; digital traces; explainable AI; algorithmic fairness; educational equity; SDG 4; systematic review

Check for updates

Received: 5 October 2026
Revised: 4 November 2026
Accepted: 5 November 2026
Published: 9 November 2026

**Correction Statement:** This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Student dropout remains a persistent challenge for higher education systems because it affects students, institutions, and societies simultaneously. For students, leaving university can reduce opportunities for social mobility and professional development; for institutions, it represents a loss of resources and raises questions about educational effectiveness; and for society, it limits the development of human capital [1–3]. The scale of the problem is substantial across regions. In the United States, 39% of full-time university students do not graduate within eight years, and 23.3% abandon their studies during the first year [4,5].

In Latin America, dropout rates exceed 50% in Colombia and approach 40% in Ecuador, with a concentration in the early semesters [4,6]. The Organization for Economic Co-operation and Development (OECD) reports a mean dropout rate of 31% among its member countries, reaching 46% in cases such as New Zealand [7]. Similar concerns have been documented in Europe, including abandonment rates of 27.5% in Andalusia and 12.2% in the transition from the first to the second year in Italy [4].

The expansion of higher education has increased the urgency of identifying at-risk students before dropout becomes irreversible. In Latin America and the Caribbean, enrollment rose from 17% in 1991 to 40% in 2010 [6,8], widening access but also requiring institutions to design timely support strategies for increasingly diverse student populations. Early identification is therefore central to retention policy, since preventive intervention depends on detecting academic disengagement, performance decline, or other risk signals while there is still time to act [4,9].

Dropout also bears on the 2030 Agenda for Sustainable Development. Its fourth goal (SDG 4) includes two targets that concern universities directly. Target 4.3 commits countries to ensure equal access for all women and men to affordable and quality tertiary education, and target 4.5 to eliminate disparities in access to education for vulnerable groups [10]. Enrollment alone satisfies neither target, because a student who leaves before graduation does not obtain the qualification that access was meant to provide. Retention is therefore part of what equal access means in practice. Dropout is also a question of economic and institutional sustainability, since every withdrawal represents public funding, family resources, and student time spent without the corresponding degree [2,3]. From this perspective, an early-warning system contributes to sustainable higher education only if it identifies risk reliably across all student groups, including those that leave the fewest traces in institutional systems.

Artificial Intelligence (AI), Machine Learning (ML), and Educational Data Mining (EDM) have become increasingly relevant in this context, as they enable the analysis of large volumes of educational data and the detection of patterns that would be difficult to identify manually [6,11]. These techniques have been applied to predict at-risk students using academic performance, demographic variables, and interactions with digital educational platforms [4,12]. Their capacity to model complex and non-linear relationships has made them especially useful for early-warning systems, with hybrid models reporting particularly high levels of predictive accuracy [4,5]. Research on ML-based retention prediction has expanded markedly since 2010, although developments in this area can be traced back to earlier work from the 1990s [12].

A key input for these predictive systems is the student digital trace. In this study, digital traces are understood as structured records generated automatically and continuously through students' interactions with institutional digital systems, including Learning Management Systems (LMS), Student Information Systems (SIS), web portals, and mobile applications [11]. These records encompass timestamped information on the frequency, duration, sequence, and nature of student actions in digital environments—logins, clicks, time on task, resource views, assignment submissions, quiz scores, forum participation, video interactions, enrollment records, and academic transcripts [12,13]. Unlike survey-based or self-reported information, digital traces are produced as a by-product of ordinary academic activity and therefore provide longitudinal evidence of engagement, participation, and academic progression. Among available sources, LMS records capture interaction with learning resources and assessments, while SIS records store demographic information, enrollment history, and historical academic performance [13]. Other emerging sources, such as Wi-Fi logs, library records, and institutional mobile applications, may broaden the behavioral scope of the analysis beyond the virtual classroom [12,13].

The models applied to these traces vary in complexity, ranging from classical ML algorithms—decision trees, logistic regression, Random Forest, support vector machines, and ensemble methods—to Deep Learning (DL) architectures such as artificial neural networks and recurrent models for temporal data [6,13,14]. Classical ML is often preferred because it combines predictive performance with relative interpretability and feasibility of institutional deployment, whereas DL may offer advantages with temporal sequences or high-dimensional behavioral data at the cost of larger datasets and greater computational resources [13,14].

Several systematic reviews have examined AI, EDM, and dropout prediction in higher education from different angles: predictive models combined with conversational agents for early intervention [4], the explanatory factors of dropout addressed through AI [6], comparisons of data mining and ML methods for student retention [15], AI applications targeting student performance, engagement, and self-efficacy [16], and ANN-based approaches to EDM [14]. Three recent reviews come closest to the present one. The first catalogs the factors, models, and explanatory techniques used for university dropout prediction between 2012 and 2024 [17]. The second covers ten years of machine learning for at-risk prediction in online education [18], and the third examines practical and explainable student performance prediction in higher education [19]. These reviews provide important evidence on variables, algorithms, prediction performance, and implementation challenges. To position the contribution of the present review, Table 1 compares it with related systematic reviews by focus, prediction target, treatment of digital traces and explainability, and main remaining gap.

**Table 1.** Comparison of prior reviews and the present study. All reviews address AI models. Keys: DT = student digital traces or data sources as an analytical category; XAI = explainability; Yes = central dimension; Part. = partially addressed; No = not treated as an analytical category.

| Review | Focus | Target | DT | XAI | Main Gap |
|---|---|---|---|---|---|
| Arévalo-Cordovilla and Peña [4] | Prediction and chatbots for early intervention | Student success | Part. | Part. | Traces not a structured category |
| Attiya and Shams [15] | Data mining techniques | Retention | Part. | No | No trace provenance or XAI |
| Liu et al. [19] | Practical and explainable AI | Performance | Part. | Yes | Performance, not dropout |
| Luo et al. [16] | AI for performance, engagement, self-efficacy | Performance | Part. | No | Not dropout-specific; no XAI |
| Okewu et al. [14] | Neural networks in EDM | Performance | Part. | No | Neural networks only |
| Parra-Sánchez et al. [6] | Explanatory factors via AI | Dropout | Part. | No | Factors; no traces or XAI |
| Pelima et al. [13] | ML with academic performance data | Graduation | Yes | No | Graduation, not dropout |
| Quimiz-Moreira et al. [17] | Factors, models, XAI, simulation | Dropout | Part. | Yes | No trace provenance, reporting, or fairness |
| Rodrigues et al. [20] | AI algorithms and features | Dropout | Part. | Part. | XAI only as a limitation |
| Shi et al. [18] | ML for at-risk students, 2014–2024 | At risk (online only) | Part. | Part. | Online education only |
| Present study | AI prediction from student digital traces | Dropout (T1–T3) | Yes | Yes | Integrates traces, models, XAI, and reporting completeness |

Existing reviews tend to address the field from one of four angles: dropout factors, algorithmic performance, EDM applications, or specific model families. Across these contributions, three limitations recur. First, data sources are typically discussed as variables or datasets rather than as digital traces with distinct behavioral and institutional meanings. Second, explainability is mentioned as a limitation of black-box models rather than treated as an analytical category for evaluating institutional actionability. Third, the gap between predictive accuracy and intervention is rarely examined: studies report metrics such as accuracy, precision, recall, F1-score, or area under the receiver operating characteristic curve (AUC), but few examine whether the predictions inform retention actions or improved outcomes [4,6]. These gaps are consequential because predictive models may inherit algorithmic bias from historical data and may operate as black boxes whose risk scores cannot be translated into pedagogical action without additional interpretation. Explainable Artificial Intelligence (XAI) techniques such as Shapley Additive exPlanations (SHAP) [21], Local Interpretable Model-agnostic Explanations (LIME) [22], and feature importance analysis address the interpretability dimension, while ethical considerations remain central to responsible deployment. These include privacy, informed consent, fairness, and accountability in the use of student data [6,13,14].

The primary studies show the same fragmentation along the three dimensions. Several compare classifiers on LMS records without an explanation mechanism [23,24], whereas others add explainability techniques to their models [25,26]. A third group broadens the data beyond the LMS, either by combining it with other institutional records [27,28] or with questionnaires administered for the study [29], or by relying on engagement captured through an institutional mobile application [30]. Each study advances one of these dimensions, but their contributions have rarely been examined together, and the present review takes on that role.

The present review addresses these gaps through an integrated framework that examines the type of digital trace used to construct the student risk profile, the AI model family employed to transform those traces into predictions, and the use of XAI techniques to make predictions interpretable. The review additionally identifies the methodological gaps and limitations reported in the literature, with particular attention to the conditions under which early-warning systems can be considered transferable, equitable, and institutionally useful. The novelty of the review lies in bringing these dimensions together rather than treating them in isolation, and in shifting the focus from predictive accuracy alone toward the methodological conditions that determine whether dropout prediction can support real institutional action.

This framework is one defensible way of organizing the evidence, not the only one. Generic frameworks for learning analytics describe the design of an analytics service through dimensions such as stakeholders, objectives, data, instruments, and internal and external constraints [31], and surveys of EDM and learning analytics classify studies by method and educational task [11]. For the prediction of student success, Alyahyan and Düştegör [32] proposed a six-stage process, from data collection to the evaluation of results, that guides how an individual study is conducted. These frameworks describe how to design or conduct an analytics study. The present framework serves a different purpose: it compares published studies along the dimensions that determine whether a prediction can be acted on, namely where the data come from, what is predicted and when, how predictions are explained, and how completely the methods are reported. For the last dimension, the review draws on TRIPOD+AI, which was developed for clinical prediction models [33]. Its predecessor has been used to audit the reporting of machine learning prediction models, with studies adhering to a median of 38.7% of the items [34], and field-agnostic recommendations for machine-learning-based science, such as REFORMS, address similar concerns about data

leakage, validation, and reproducibility outside medicine [35]. We found no adaptation of TRIPOD+AI to educational prediction; the adaptation used here is therefore exploratory, and its correspondence with the original items is documented so that it can be scrutinized (Supplementary Table S4).

The specific research questions guiding this study are the following:

RQ1: What institutional data sources are used to construct student digital traces for dropout prediction in higher education?

RQ2: Which AI model families are used to predict university dropout from student digital traces?

RQ3: What explainability mechanisms are used to make dropout prediction models interpretable for educational decision-making?

RQ4: What methodological gaps and limitations are reported in AI-based dropout prediction studies using student digital traces?

Studies in this area do not always predict dropout itself: many predict course failure or academic risk and present it as an early warning of dropout. The review therefore records the prediction target of each study in a three-level hierarchy (Section 2.1) and analyses the relation between the outcome that is predicted and the dropout outcome that is invoked; the resulting pattern is one of the main findings of the review (Section 4.1). RQ4 is addressed through an audit of how completely the studies report their methods, using items adapted from TRIPOD+AI, complemented by signals of analytical quality; both describe the published evidence rather than the underlying practices. The fourth question carries the sustainability argument of the review, which uses SDG target 4.5 as a normative frame rather than as an analytical category. Fairness evaluation, ethical approval, and the handling of missing data determine whether a dropout prediction model serves all students or only those best represented in institutional records. They therefore bear on whether the model advances SDG target 4.5 or works against it.

The remainder of the paper is organized as follows. Section 2 describes the review protocol, namely the eligibility criteria, the Boolean search strategy, the analytical categories, and the data extraction process. It also reports the search and screening procedure that produced the final corpus through the PRISMA flow. Section 3 presents a bibliometric overview, the prediction targets, horizons, and validation schemes, the classification of the reports along the analytical categories, the performance metrics they report, the sensitivity analyses, and the confidence in the findings. Section 4 first discusses the mismatch between the dropout framing of the reports and the outcome they predict, then the cross-cutting patterns, reports the audit of reporting completeness against TRIPOD+AI and the signals of analytical quality, examines the ethical issues raised by predicting dropout from digital traces and the implications for sustainable and equitable higher education, and states the limitations of the review. Section 5 presents the conclusions and the directions they open for future research.

## 2. Materials and Methods

This study is a systematic review with structured descriptive synthesis, reported in accordance with the PRISMA 2020 statement [36]; the completed PRISMA 2020 and PRISMA 2020 for Abstracts checklists are provided as Supplementary Tables S1 and S2. PRISMA 2020 was selected because it provides explicit guidance for transparent reporting, search reproducibility, study selection, and synthesis when meta-analysis is not feasible. These conditions apply directly to a corpus of studies on AI-based dropout prediction that differ substantially in data sources, prediction targets, models, and evaluation metrics. This section describes the eligibility criteria, the search strategy and its execution, and the analyt-

ical framework used to classify and extract data from the included studies. The execution of this protocol and the resulting PRISMA flow are reported in Section 2.5.

Registration and protocol. The review was registered retrospectively in OSF Registries on 18 May 2026 (DOI: https://doi.org/10.17605/OSF.IO/MTQH2) [37], after the database searches (14 April 2026), screening, and the first round of data extraction had been completed; the registration explicitly discloses this foreknowledge. No separate review protocol was prepared beyond the registration form, which documents the research questions, eligibility criteria, information sources, coding variables, and analysis plan. Because registration was retrospective, it cannot rule out post hoc decisions, and this is treated as a formal limitation of the review (Section 4.7). The review deviates from the registration in six respects. First, the title was revised to reflect the sustainability framing of the review. Second, the registered question on recurrent configurations across trace type, model family, explainability, and reporting quality is addressed through the integrated mapping in Section 3.3 and the cross-cutting analysis in Section 4.2 rather than as a separate research question. Third, fairness evidence is reported at the measurement and mitigation levels; the acknowledgement level of the registered three-level scale is recorded as "discusses the risk of bias only". Fourth, the registered configuration profile of responsible-deployment readiness was not operationalized as a separate index. Fifth, the registered criterion on "institutional student data or digital traces" was operationalized as the presence of interaction-generated digital traces (Section 2.1). Sixth, in response to peer review, the operational definition of dropout prediction was made explicit through a hierarchy of prediction targets (Section 2.1); this specification does not broaden the object of study, because the search strings required a dropout, attrition, or retention term, and studies of academic risk not framed as dropout are outside the scope by design. Making the targets explicit also revealed that most reports framed as dropout prediction model course failure, a pattern reported as a finding (Section 4.1). In addition, all 21 previously included reports were re-assessed against written decision rules with independent double coding (Section 2.3); two reports were excluded as a result (Section 2.5).

Synthesis approach. A statistical meta-analysis was not feasible because the reviewed studies did not share a common outcome metric or a sufficiently homogeneous methodological structure from which comparable effect sizes could be derived. Alternatives to a descriptive synthesis were considered. Realist synthesis explains how and why interventions work in particular contexts [38], but the primary studies report predictive performance rather than the contexts and mechanisms through which predictions are used, so the information it requires is not available in this corpus. Framework synthesis organizes extracted evidence under an a priori conceptual framework [39]; the present review follows this logic in part, because the analytical categories in Section 2.3 were defined before extraction and applied to every study. The synthesis was therefore conducted as a structured synthesis without meta-analysis, following the SWiM reporting items where applicable [40]: studies were grouped by the analytical categories, and counts and proportions were tabulated for each category and reporting item. Such counts describe how often a design choice or reporting practice occurs; they treat each report as one unit, ignore sample size and precision, and cannot support claims about the magnitude of effects or the superiority of any model family. They are interpreted in this descriptive sense throughout.

Reporting completeness and analytical quality. Two complementary assessments were applied to every included study. The first audits reporting completeness with the 16 Methods items of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis with Artificial Intelligence (TRIPOD+AI) statement [33], adapted to this domain. TRIPOD+AI is a reporting guideline: an item coded as not reported means that the information is absent from the report, not that the corre-

sponding analytical step, ethical approval, or evaluation was not performed. The audit is therefore not an assessment of risk of bias, and its results are interpreted as evidence about transparency. The second assessment records seven signals of analytical quality (Q1–Q7, Section 2.3) that capture design problems detectable from the report, such as information leakage or the absence of temporal or external validation. These signals are descriptive flags, not a validated risk-of-bias score. Established risk-of-bias tools were considered but fit this corpus less directly. PROBAST and its update PROBAST+AI assess risk of bias in prediction model studies [41,42], but their signaling questions are framed around clinical participants, predictors, and outcomes. QUADAS-2 targets diagnostic accuracy studies [43], and the Mixed Methods Appraisal Tool appraises qualitative, quantitative, and mixed methods studies in general [44], so neither addresses the analytical decisions specific to machine learning prediction. The results of both assessments are reported in Section 4.4.

The review question is stated in PICOS terms. The population comprises students enrolled in campus-based or blended higher education programs at university level, at any stage of their academic trajectory. The intervention is the construction of a predictive model of dropout, attrition, or retention, or of course failure or academic risk explicitly framed as an early warning of dropout, using artificial intelligence, machine learning, or deep learning trained on interaction-generated digital traces. The comparators are the alternative model families, trace sources, and baseline or comparator algorithms evaluated within each primary study; no comparison against a non-algorithmic reference standard was required for inclusion. The outcomes of interest are the reported predictive performance of the models, the explainability mechanisms used to render their outputs interpretable, and the reporting practices and analytical quality signals audited in this review. Eligible study designs are empirical, peer-reviewed journal articles and full conference papers reporting the development, evaluation, or institutional application of such models; reviews, meta-analyses, purely theoretical contributions, and studies restricted to online-only education were excluded.

### *2.1. Eligibility Criteria*

Inclusion and exclusion criteria were defined according to four dimensions: publication type, language, temporal window, and object of study. The review included empirical, peer-reviewed publications indexed in Scopus or Web of Science, comprising both journal articles and full conference papers (i.e., complete research papers presented at peer-reviewed international conferences and subsequently indexed). This dual inclusion was adopted because conference proceedings indexed in the selected databases provide original empirical evidence subject to peer review and represent a significant proportion of the methodological output in fast-moving computational fields such as AI-based educational analytics. Conference abstracts, posters, editorials, systematic reviews, meta-analyses, theoretical papers, book chapters, and gray literature were excluded to maintain homogeneity in the corpus and to ensure that all included studies report substantive empirical methodology. Only publications in English were included, given the predominance of English in international scientific communication and the resulting comparability between studies.

The search period spans 2015–2025, corresponding to the phase of rapid expansion of AI, ML, DL, and learning analytics applications in higher education. This temporal window was selected to capture recent methodological developments in AI-based dropout prediction while excluding older studies that predate the widespread institutional adoption of digital learning platforms and data-driven early-warning systems. The publication

window refers to the year of publication; the searches were last executed on 14 April 2026, and no record indexed after that date was incorporated.

The object of study was limited to empirical research that predicts university dropout with AI, ML, or DL models, including studies that predict course failure or academic risk explicitly framed as an early warning of dropout; studies framed within educational data mining (EDM) or learning analytics were included when they applied such models. The prediction target was classified in a three-level hierarchy: (T1) dropout, attrition, or retention at programme or institution level; (T2) withdrawal from or non-completion of a course or module; and (T3) course failure or academic risk, eligible only when the authors explicitly linked it to dropout, retention, or attrition as an early warning. Targets limited to grades or academic performance without that link, and other targets such as cluster labels or satisfaction, were out of scope. Studies were included when the predictive model used interaction-generated digital traces, that is, records produced automatically through students' use of institutional digital platforms (e.g., LMS activity logs, mobile application or institutional portal usage), either alone or combined with SIS, administrative, or academic records. Studies were excluded when they relied exclusively on static administrative or academic records, registry data, or surveys without interaction-generated traces; when they focused exclusively on Massive Open Online Courses (MOOCs), online courses, distance learning, e-learning, self-paced education, or non-university educational contexts; when they did not address dropout, retention, or an eligible early-risk target; when they did not use an AI-based predictive model; when they lacked empirical evidence; or when they did not provide sufficient methodological information to identify the data source, model type, prediction target, or evaluation procedure.

Decision rules for borderline cases. Before the re-assessment of the corpus, the two reviewers agreed and froze seven written rules (29 October 2026; available in the OSF repository). Reports whose model results appeared only in figures were eligible but flagged (P1). Emergency remote teaching in a campus-based university (P2) and datasets mixing face-to-face and online sections without separating them (P3) were eligible but flagged. A T3 target was accepted when the link to dropout appeared anywhere in the report (P4); when it appeared only in the abstract, introduction, or conclusions, the report was flagged as having weak T3 framing. A continuous grade with an author-defined risk threshold was coded as T3 when it met P4 (P5). Reports analysing the same dataset were retained when they applied different techniques or reported different results, and the shared dataset was recorded (P6). Finally, decision rules for the TRIPOD+AI items were fixed in advance (P7; Section 4.4). Flagged reports were retained in the main synthesis and removed in a sensitivity analysis (Section 2.4).

### *2.2. Search Strategy and Execution*

The search was conducted in Scopus and Web of Science (WoS) for two complementary reasons. WoS offers selective indexing of high-impact and multidisciplinary journals, while Scopus provides broad coverage of scientific and technical literature. Combined, these databases cover research on AI, EDM, learning analytics, and dropout prediction in higher education, including the main IEEE and ACM proceedings series. Conference papers in venues indexed by neither database remain outside the search. The "articles and conference papers/proceedings papers" filter was applied in both databases to ensure consistency with the eligibility criteria; conference papers therefore passed the document-type filter and are represented in the included corpus.

The search algorithm was structured around five conceptual scopes adapted from the Population–Intervention–Comparison–Outcome framework: Problem (the dropout phenomenon), Population (university-level higher education), Tools (ML/DL techniques), Phenomenon of interest (prediction and early-warning systems), and Exclusion (online-only educational modalities outside the scope of the review). The exclusion scope was added because studies focused exclusively on MOOCs, distance learning, e-learning, or self-paced courses operate under data conditions and pedagogical assumptions that differ substantially from those of institutional higher education, and were therefore outside the analytical frame of this review. Because the `AND NOT` operator acts on titles, abstracts, and keywords, it may also have removed campus-based or blended studies that mention an online component in these fields. This possibility could not be quantified, because the excluded records were not retrieved, and it is treated as a limitation on coverage (Section 4.7). Table 2 reports the keywords selected for each scope, and these were combined with Boolean operators to form the search algorithm (SA), defined as:

$$\begin{aligned} SA = {} & (\textsc{Problem}) \text{ AND } (\textsc{Population}) \text{ AND} \\ & (\textsc{Tools}) \text{ AND } (\textsc{Phenomenon of Interest}) \\ & \text{AND NOT } (\textsc{Exclusion}). \end{aligned}$$

The corresponding database-specific strings, applied to title, abstract, and keyword fields (`TITLE-ABS-KEY` in Scopus; `TS=` in WoS), are reported in Table 3. Both strings were executed on 14 April 2026 in both databases, with the same filters applied in each, and produced the initial pool of records subsequently processed through the PRISMA flow described in Section 2.5.

As a terminological check, the Scopus search was also run on the same date with an expanded Problem block that added the terms "student success", "student persistence", "academic risk", and "retention analytics". This check was exploratory: the additional records it returned were inspected against the eligibility criteria, but they were not exported or logged in the screening platform, so their number cannot be reported and they are not represented in the PRISMA flow diagram. No additional eligible study was identified. Because the check retained the exclusion block, it does not bear on records that the block may have removed.

**Table 2.** Search terms by conceptual scope. Each term is listed once per scope. An asterisk (*) denotes truncation, so that all word endings are retrieved (e.g., forecast* retrieves forecast, forecasts, and forecasting).

| Scope | Search Terms (Combined with OR Within Each Scope) |
|---|---|
| Problem | "student dropout" OR "student drop-out" OR "student attrition" OR "academic desertion" OR "university dropout" OR "university drop-out" OR "higher education dropout" OR "higher education drop-out" OR "student retention" |
| Population | "higher education" OR university OR universities OR "tertiary education" |
| Tools | "artificial intelligence" OR AI OR "machine learning" OR ML OR "deep learning" |
| Phenomenon of interest | prediction OR predicting OR forecast* OR "early warning system*" OR "prediction algorithm*" |
| Exclusion (AND NOT) | "massive open online course*" OR MOOC* OR "distance learning" OR "online learning" OR "e-learning" OR "online university education" OR "online course*" OR "self-paced education" |

**Table 3.** Boolean search strings by database. An asterisk (*) denotes truncation, so that all word endings are retrieved.

| Database | Search String |
| --- | --- |
| Scopus | TITLE-ABS-KEY ( ``student dropout'' OR ``student drop-out'' OR ``student attrition'' OR ``academic desertion'' OR ``university dropout'' OR ``university drop-out'' OR ``higher education dropout'' OR ``higher education drop-out'' OR ``student retention'' ) AND TITLE-ABS-KEY (``higher education'' OR university OR universities OR ``tertiary education'') AND TITLE-ABS-KEY (``artificial intelligence'' OR ``AI'' OR ``machine learning'' OR ``ML'' OR ``deep learning'') AND TITLE-ABS-KEY ( prediction OR predicting OR forecast* OR ``early warning system*'' OR ``prediction algorithm*'' ) AND NOT TITLE-ABS-KEY ( ``massive open online course*'' OR ``MOOC*'' OR ``distance learning'' OR ``online learning'' OR ``e-learning'' OR ``online university education'' OR ``online course*'' OR ``self-paced education'' ) |
| WoS | TS=(``student dropout'' OR ``student drop-out'' OR ``student attrition'' OR ``academic desertion'' OR ``university dropout'' OR ``university drop-out'' OR ``higher education dropout'' OR ``higher education drop-out'' OR ``student retention'') AND TS=(``higher education'' OR university OR universities OR ``tertiary education'') AND TS=(``artificial intelligence'' OR ``AI'' OR ``machine learning'' OR ``ML'' OR ``deep learning'') AND TS=(prediction OR predicting OR forecast* OR ``early warning system*'' OR ``prediction algorithm*'') NOT TS=(``massive open online course*'' OR ``MOOC*'' OR ``distance learning'' OR ``online learning'' OR ``e-learning'' OR ``online university education'' OR ``online course*'' OR ``self-paced education'') |

### *2.3. Analytical Categories and Data Extraction*

The included studies were classified along three core analytical categories, which capture the main components of an AI-based dropout prediction system: the source of the student data used as input, the model family that transforms those data into risk estimates, and the use of explainability mechanisms. The framework also includes four descriptive variables that condition the interpretation of the core categories: the prediction target, the prediction horizon, the validation scheme, and the evaluation of fairness. The categories are dimensions of the same study rather than mutually exclusive classes, so a study can, for example, be hybrid in its data source and mixed in its model family. Table 4 lists every variable with its admissible values; the full codebook, with the decision rule for each variable, is available in the OSF repository.

Category 1—Type of digital trace. This category records the data sources used to construct the student digital trace, since the type of source determines which dimensions of student behavior can be observed. Studies were classified as LMS-based when all predictors came from learning management platforms; as based on another interaction platform when the interaction records came from a platform other than the LMS (such as an institutional mobile application, a programming environment, or an educational simulator); as hybrid when either type of interaction record was combined with institutional records (SIS, enrollment, transcripts, or demographic data); and as including external data when non-institutional sources were added (public neighbourhood statistics, external social networks, or questionnaires administered for the study). When an LMS was combined with another interaction platform, the study was coded as LMS-based. The type of interaction record was also coded as behavioural (clicks, accesses, time), assessment-based (grades and graded submissions recorded on the platform), or both.

Category 2—AI model family. This category records the type of model used for dropout prediction, since the model family determines how relationships between variables are represented, how complex the detected patterns can be, and how interpretable the resulting predictions may be. Studies were classified as classical ML when they used algorithms such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, Naive

Bayes, K-Nearest Neighbors, XGBoost, or other ensemble methods; as DL when they used neural network architectures such as Artificial Neural Networks, Multilayer Perceptrons, Convolutional Neural Networks, Recurrent Neural Networks, or Long Short-Term Memory networks; and as mixed when they combined ML and DL techniques. The family was coded from all models evaluated in the study, including baselines, so a study that compared a neural network with tree-based classifiers was coded as mixed. Because the mixed category can conceal different designs, its subtype was coded as a parallel comparison of ML and DL models, a pipeline in which ML feeds a DL model, a DL representation feeding an ML classifier, or an ensemble across families.

Category 3—Explainability mechanism. This category records whether the study included an explicit mechanism for interpreting the model or its predictions, given that dropout prediction models are intended to support educational decision-making and risk classifications must be understandable for institutional or pedagogical use. Studies were classified as XAI-equipped when they reported at least one technique applied to a trained model or its predictions (such as SHAP, LIME, feature importance analysis, rule extraction, or displayed coefficients), and as non-XAI when they reported predictive results without an explicit explanation mechanism; correlations computed before modelling and descriptive statistics did not count. For XAI-equipped studies, three further variables were coded so that the use of explainability, and not only its presence, could be described: its function (global understanding of risk factors, feature selection, model validation or debugging, individual explanations for users, or several of these), its scope (global, local, or both), and whether the explanations were communicated to end users.

Descriptive variables. The prediction target was coded with the T1–T3 hierarchy of Section 2.1, together with the unit to which the predicted event refers (programme, semester or year, or course). The prediction horizon was coded as before or at the start of the period, early (up to one third of the period), intermediate, late (more than two thirds), retrospective (data from the whole period), or multiple time cuts. The validation scheme was coded as random hold-out, *k*-fold cross-validation, nested cross-validation, temporal or cohort-based, external, or combined; a combined scheme required two schemes supporting the main results. Fairness was coded at the highest level reached: none, discussion of the risk of bias only, performance by subgroup, a fairness metric, or mitigation. Only demographic, socioeconomic, or other protected subgroups counted as fairness subgroups; behavioural subgroups, such as engagement terciles, were recorded but did not count.

Signals of analytical quality. Seven signals were coded as present, absent, or not determinable: (Q1) predictors that contain the outcome or posterior information, such as grade components of the final grade; (Q2) resampling, feature selection, or normalisation applied to the whole dataset before the train–test split; (Q3) repeated observations of the same student on both sides of the split, as with sliding windows or log records split without grouping by student; (Q4) no temporal or external validation; (Q5) tuning or selection with the same data on which performance is reported, when only the selected result is reported; (Q6) few minority-class events (fewer than 50 in total under cross-validation, or fewer than 20 in a held-out test set); and (Q7) internal contradictions in data or results. A signal was coded as present when the evidence in the report showed it, even if the authors did not state it explicitly.

**Table 4.** Analytical framework: core categories, descriptive variables, and admissible values. The full codebook, with the decision rule for each variable, is deposited in the OSF repository.

| Variable | Admissible Values |
|---|---|
| C1. Type of digital trace (source) | LMS; other interaction platform; hybrid (LMS + institutional records); hybrid (other platform + institutional records); includes external, non-institutional data |
| Type of interaction record | Behavioural (clicks, accesses, time); assessment-based (grades and graded submissions on the platform); both |
| C2. AI model family | Classical ML; DL; mixed (ML + DL), coded from all models evaluated |
| Subtype of mixed | Parallel comparison of ML and DL; ML → DL pipeline; DL representation → ML classifier; ensemble across families |
| C3. Explainability (XAI) | Yes (technique applied to a trained model or its predictions); no |
| Function of XAI | Global understanding of risk factors; feature selection; model validation or debugging; individual explanations for users; several |
| Scope and communication | Global, local, or both; explanations communicated to end users (yes/no) |
| Prediction target | T1 programme dropout; T2 course withdrawal or non-completion; T3 failure or academic risk framed as an early warning of dropout |
| Unit of the predicted event | Programme or institution; semester or year; course or module |
| Prediction horizon | Before or at the start of the period; early ($\leq$1/3); intermediate; late (>2/3); retrospective (whole period); multiple time cuts; not reported |
| Validation scheme | Random hold-out; *k*-fold; nested; temporal or cohort-based; external; combined |
| Fairness evaluation | None; discussion of risk of bias only; performance by protected subgroup; fairness metric; mitigation |
| TRIPOD+AI Methods items | 16 items coded 1 (reported), 0 (not reported), or NA (not applicable); score = items reported/items applicable (Section 4.4) |
| Quality signals Q1–Q7 | Present; absent; not determinable |

Data extraction, reliability, and disagreement resolution. The first round of extraction, completed before the re-assessment, was coded by consensus and its pre-consensus records were not retained. For the re-assessment, the coding was repeated from the full texts for all 21 previously included reports, including the two that were finally excluded. Two reviewers (M.A. and A.J.) worked from a written codebook and a common template. Each coded independently in a separate spreadsheet, without access to the other's coding, and recorded a supporting quotation and its location for every eligibility criterion, characteristic, TRIPOD+AI item, and quality signal. Evidence dossiers with verbatim passages organised by variable were prepared to help locate information in each report, and each reviewer prefilled the spreadsheet with a different generative AI tool (M.A.: Claude Opus 5.5, Anthropic; A.J.: Z.AI GLM-5.3, Zhipu AI; both accessed on 30 October 2026) and verified every code against the full text. All final codes were decided by the reviewers. The two pre-consensus spreadsheets were frozen and fingerprinted (SHA-256) before comparison.

Agreement was computed on the pre-consensus spreadsheets with Cohen's kappa [45] and Krippendorff's alpha [46] for each categorical variable (Supplementary Table S3). Agreement on the inclusion decision, coded as include, include with a flag, merge with another report, or exclude, was fair ($\kappa = 0.36$, $\alpha = 0.33$; 52% observed agreement), and most disagreements concerned the boundary between performance targets and early-warning targets ($\kappa = 0.73$ for the target type). Agreement was substantial for the TRIPOD+AI items pooled ($\kappa = 0.80$; 91% over 336 item ratings), ranged from 0.47 to 1.00 for the categorical characteristics, and was low for several quality signals (0.06 to 0.62). Because both spread-

sheets were AI-assisted, these coefficients describe concordance between two independently verified codings rather than between unaided coders.

Disagreements were identified by automated cell-by-cell comparison (145 in total). Each was resolved in a joint session by returning to the full text: the two codes and their supporting evidence were compared against the codebook rule, and the consensus value was recorded with a written justification and the quotation that supported it. A third author (R.O.) was designated as arbiter for disagreements that persisted after discussion; none required arbitration. When a disagreement revealed a gap in a rule, the clarification was recorded with its date and applied uniformly to all reports; eight clarifications were recorded (for the trace source, record type, fairness, three quality signals, and two TRIPOD+AI items). Two worked examples of borderline decisions are given in Supplementary Material S1. In the first, the supervised model predicted membership in engagement clusters rather than dropout itself; one reviewer coded the target as out of scope and the other as an early-warning indicator. Applying rule P4 and the target hierarchy, consensus classified it as T3 with a weak-framing flag, because the authors validated the clusters against observed dropout and used them to identify at-risk students [23]. In the second, the declared aim was pass/fail classification and dropout prevention appeared only in the abstract; consensus retained the report as T3 with the same flag [47]. Both reports are removed in the sensitivity analysis restricted to unflagged reports.

Conflicts of interest of the reviewers. No author of this review is an author of any included study, which was verified against the author lists of all included reports. The authors declare no financial or professional relationship with the authors of the included studies or with the developers of the platforms and tools they examined.

### *2.4. Synthesis Methods, Reporting Bias, and Certainty Assessment*

Synthesis. All 19 included reports were eligible for every synthesis. Extracted data were tabulated by report and grouped along the analytical categories; counts and proportions were computed for each category, descriptive variable, TRIPOD+AI item, and quality signal, and the joint distribution of the core categories was mapped report by report (Section 3.3). Reported performance metrics were tabulated as published, without conversion or pooling, because outcome definitions, metrics, and validation schemes differed across studies (Section 3.4). Heterogeneity was explored descriptively by comparing outcome definitions, data sources, prediction horizons, and validation procedures.

Sensitivity analyses. Five sensitivity analyses recomputed every count: (S1) excluding the eight flagged reports ($n = 11$); (S2) excluding the two reports from the same research group that share a methodological approach [48,49] ($n = 17$); (S3) retaining one report per dataset ($n = 18$); and (S4) restricting the corpus to programme dropout and course withdrawal targets (T1–T2; $n = 4$); and (S5) excluding conference papers ($n = 11$). The mixed model family was disaggregated by subtype in the main analysis. A sensitivity analysis including online-only education was not feasible, because such studies were excluded by the search itself and would require a new search; this is discussed as a limitation (Section 4.7).

Reporting bias. Statistical methods for detecting publication bias, such as funnel plots or Egger's test, were not applicable, because the included studies share no common effect measure and the corpus is small and heterogeneous. Reporting bias was instead assessed qualitatively along four dimensions: (i) publication bias favoring high-performing models; (ii) selective outcome reporting, operationalized as reliance on threshold-dependent metrics such as accuracy without a threshold-independent measure such as AUC, and as near-ceiling results obtained under a single held-out split; (iii) language and database coverage bias arising from the restriction to English-language records indexed in Scopus and WoS; and

(iv) non-reporting in the TRIPOD+AI audit, where an unreported practice cannot be distinguished from an absent one.

Certainty of the evidence. GRADE [50] was not applied in its standard form, because it is designed to rate confidence in pooled effect estimates, whereas this review synthesizes the prevalence of data sources, model families, and reporting practices. Confidence in each main finding was instead rated as high, moderate, low, or very low using domains adapted from GRADE: methodological limitations of the contributing studies, inconsistency across studies, indirectness with respect to the review question, imprecision related to the number of contributing studies, and reporting bias. Each finding started at high confidence, because it consists of a direct count over the included corpus, and was downgraded by one level for each serious concern. Ratings were assigned by M.A. and A.J. and agreed by consensus.

### *2.5. Search and Screening Process*

The execution of the search and screening protocol defined in Section 2 proceeded through seven sequential phases: initial search in Scopus and WoS, application of language/period/document-type filters, export to the Rayyan screening platform, duplicate removal, screening by title and abstract, retrieval of full texts, and final eligibility verification through full-text reading, followed by the re-assessment of the included reports described above. Table 5 reports the number of records retrieved and excluded at each phase; Figure 1 presents the corresponding PRISMA 2020 flow diagram.

**Table 5.** Execution of the search and screening algorithm by phase.

| Phase | Action | Records |
| --- | --- | --- |
| Initial search | Execution of the search algorithm in Scopus and WoSon 14 April 2026. | WoS: 159; Scopus: 305; total: 464 |
| Filter application | Filters by language (English), period (2015–2025), and document type (journal article or conference paper). | Removed: 51; remaining: 413 |
| Export and management | Export of records in RIS format and import into Rayyan. | 413 records imported |
| Duplicate removal | Automated identification and manual confirmation of duplicates in Rayyan. | Duplicates removed: 116; remaining: 297 |
| Title/abstract screening | Independent, blinded review of titles and abstracts by two reviewers against the eligibility criteria. | Excluded: 270; remaining: 27 |
| Full-text retrieval | Search for and access to full-text articles (2 behind a paywall). | Not retrieved: 2; available for full reading: 25 |
| Final eligibility | Full-text reading by two reviewers and final verification of eligibility. | Excluded: 4; retained: 21 |
| Re-assessment | Re-assessment of the 21 retained reports with written decision rules and independent double coding. | Excluded: 2; included: 19 reports (18 distinct datasets) |

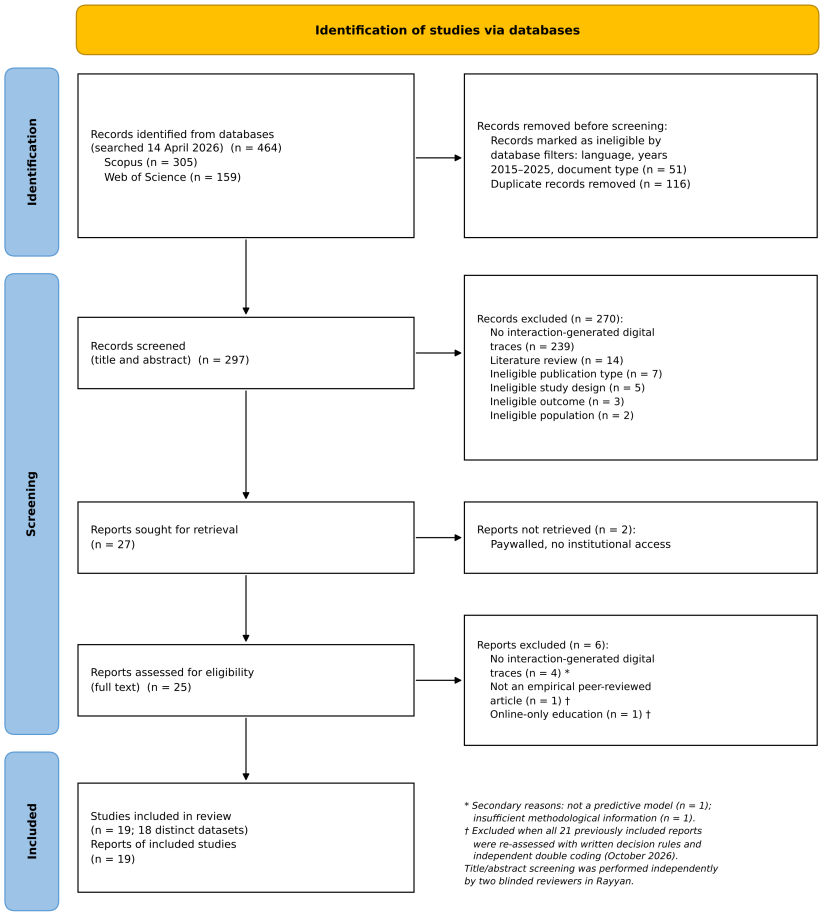



**Figure 1.** PRISMA 2020 flow diagram of the search and screening process (searches executed on 14 April 2026). Two reports included in the first round were excluded when all 21 were re-assessed with written decision rules and independent double coding (Section 2.5). Two included reports analyse the same public dataset, so the 19 reports correspond to 18 distinct datasets. The exploratory terminological check described in Section 2.2 is not shown because its records were not retained.

The initial search, executed on 14 April 2026, retrieved 464 records (305 from Scopus and 159 from WoS). Filters by language (English), publication period (2015–2025), and document type (journal article or conference paper) reduced the pool to 413 records, which were exported in RIS format and imported into Rayyan [51]. Automated duplicate detection followed by manual confirmation removed 116 duplicates, leaving 297 unique records.

Title and abstract screening was performed independently by two reviewers (M.A. and A.J.) in Rayyan with blinding enabled, between 20 and 25 April 2026. Agreement on the initial, pre-discussion decisions was substantial [52]: the reviewers disagreed on 17 of the 297 records (observed agreement 94.3%; Cohen's $\kappa = 0.72$, 95% bootstrap confidence interval 0.58–0.83, with "maybe" decisions treated as provisional inclusions). All conflicts were resolved by discussion between the two reviewers, without recourse to a third reviewer. Screening discarded 270 records: 239 did not use interaction-generated digital traces, 14 were literature reviews, 7 were of an ineligible publication type, 5 had an ineligible study design, 3 addressed an ineligible outcome, and 2 an ineligible population. The remaining 27 reports were sought for retrieval, and full texts were assessed by the same two reviewers following the same procedure.

Two reports could not be retrieved because they were behind a paywall not covered by the institutional subscriptions available to the authors [53,54]. Of the 25 reports assessed in full, four were excluded in the first round because they relied exclusively on static

academic, administrative, registry, or survey data and used no interaction-generated digital traces [55–58]. In addition, one of them did not develop a predictive model but an offline reinforcement-learning policy for allocating student aid [58], and another did not report sufficient information to identify the model architecture or the validation procedure [57].

The re-assessment of the 21 reports retained in the first round excluded two more. One is classified by its publisher as a conceptual paper rather than an empirical research article [59]; it analyses the same dataset as an included report [48]. The other analyses three exclusively online higher education institutions [60]. Of the 19 included reports, eight were retained with a flag: three for mixed or emergency remote modality [24,61,62], one because its results are reported only in figures [63], two for weak T3 framing [23,47], one for weak evidence of interaction traces [64], and one because its dataset includes some secondary-school records that are not analysed separately [65]. Two included reports analyse the same public dataset with different techniques [49,62], so the corpus comprises 19 reports on 18 distinct datasets. The eligibility decision for each report, with the criterion codes and supporting quotations, is provided in the OSF repository. The final corpus forms the basis of the analyses reported in Section 3.

## 3. Results

This section reports the characteristics of the 19 included reports and the synthesized findings. The bibliometric overview is followed by the prediction targets, horizons, and validation schemes, the classification along the three core analytical categories, the performance metrics reported across the corpus, the sensitivity analyses, and the assessment of reporting biases and confidence in the findings. The results of the reporting-completeness audit and of the analytical quality signals are reported in Section 4.4.

### *3.1. Bibliometric Overview*

The final corpus comprises 19 peer-reviewed reports, published between 2015 and 2025, that predict dropout, directly or through course failure or academic risk framed as an early warning of dropout, in campus-based or blended higher education with AI models trained on interaction-generated digital traces. The temporal and geographic distribution of these reports shows a sharp acceleration of output in the most recent years and a set of producing countries that is broad in number but concentrated in practice.

Figure 2 shows the temporal distribution of the included reports. The earliest dates from 2015 [63], followed by a gap until 2020. One report was published in each of 2015, 2020, and 2021; three in 2023; five in 2024; and eight in 2025, including in-press works indexed before the search date. Thirteen of the 19 reports (68%) were published in 2024 or 2025, and 16 appeared from 2023 onward. No report was published between 2016 and 2019 or in 2022. With annual counts this small, Figure 2 describes recent growth but cannot support fitting a trend model.

The geographic distribution of author affiliations, presented in Figure 3 and summarized in Table 6, includes 21 countries across the 19 reports. A country is counted once per report when at least one author lists an affiliation there, including secondary affiliations, and Table A1 in Appendix A maps each report to its countries. Australia [48,49,66] and the United States [30,61,62] each appear in three reports, and Canada, Croatia, Ireland, Russia, and Spain in two each. The remaining 14 countries appear in one report each, and five of them (Pakistan, South Korea, Mexico, the Dominican Republic, and Angola) appear only through a multinational study whose eight authors list affiliations in eight countries [62]. The count of 21 countries therefore overstates the geographic breadth of the research activity: the corpus spans 16 countries when that study is set aside, 13 of the

19 reports involve at least one of the seven countries with two or more reports, and no report is led from Latin America.

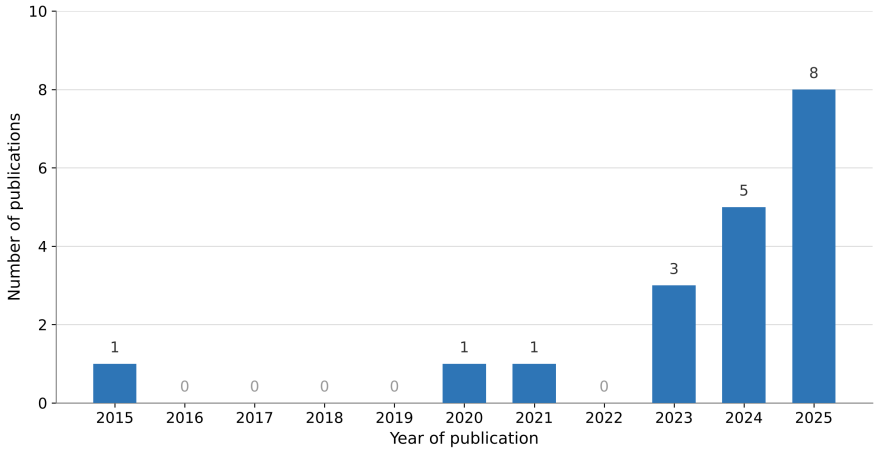



**Figure 2.** Number of publications included in the review, grouped by year of publication ($n$ = 19, 2015–2025).

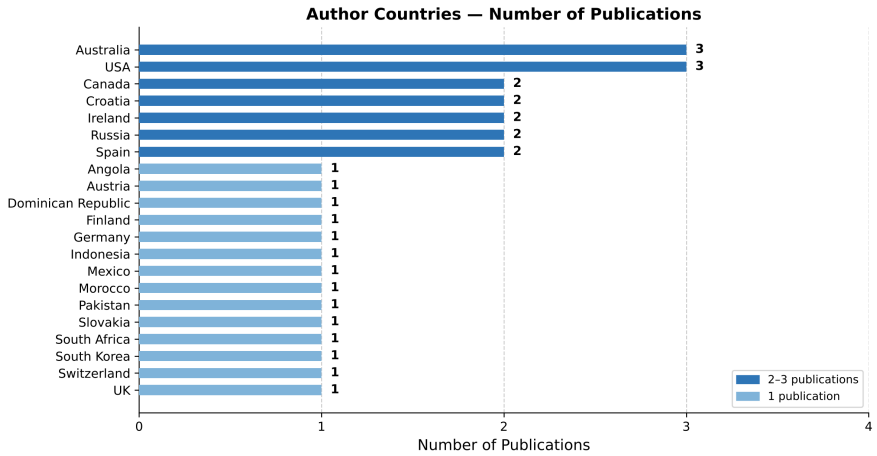



**Figure 3.** Author-country distribution across the included publications. A country is counted once per publication regardless of the number of co-authors from that country ($n$ = 19 publications, 21 unique countries).

**Table 6.** Number of publications per author country; 19 included reports.

| Country | Publications | Country | Publications |
|---|---|---|---|
| Australia | 3 | Finland | 1 |
| United States | 3 | Indonesia | 1 |
| Spain | 2 | United Kingdom | 1 |
| Canada | 2 | South Africa | 1 |
| Croatia | 2 | Austria | 1 |
| Ireland | 2 | Pakistan | 1 |
| Russia | 2 | South Korea | 1 |
| Morocco | 1 | Germany | 1 |
| Slovakia | 1 | Switzerland | 1 |
| Mexico | 1 | Dominican Republic | 1 |
| Angola | 1 | | |
| | | Total (unique countries): 21 | |

Taken together, these patterns describe a recent and geographically concentrated literature, so the findings reported below reflect current practice in a limited set of research communities rather than a global or historical picture.

### *3.2. Prediction Targets, Horizons, and Validation*

Table 7 summarizes the distribution of the 19 reports across all coded variables, and Section 3.3 maps each report to its values. Fifteen reports (79%) predict course failure or academic risk explicitly framed as an early warning of dropout (T3); three predict dropout at programme level (T1) [25,30,64], and one predicts withdrawal from a course (T2) [24]. Thus, although every report was retrieved by a dropout-related search and presents its model as a contribution to dropout prediction or early warning, the outcome modelled in most of them is course failure or academic risk rather than observed departure from the programme or institution. We refer to this pattern as target substitution. It is stable when the flagged reports are removed (9 of 11; Section 3.5), and its consequences are discussed in Section 4.1. T3 targets were defined in different ways: as a final grade below a pass threshold (for example, below 40%, 50%, or 70%), as at least one failed course in the semester, as membership in high-risk clusters, or as a continuous grade with an author-defined risk band [29,65].

The prediction horizon is reported as multiple time cuts within the period in nine reports, so that performance can be traced from the first weeks onward [27,63,66,67]. Six reports use data from the whole period, which describes the association between engagement and outcomes but does not show how early a warning could be issued; all six have a T3 target. One report predicts from the first semester alone [30], one from the midpoint of a 12-week course [68], one from week 12 of a 15-week course [24], and one does not report a horizon relative to the course length [26].

The validation scheme supporting the main results is a random hold-out split in seven reports and *k*-fold cross-validation in seven. Three combine two schemes, and only two use a temporal or cohort-based design in which the model is tested on students not yet observed when it was trained [25,64]; both predict programme dropout (T1), so no T3 model is tested on a later cohort. One additional report tests transfer of models across universities, with a marked loss of discrimination [30].

**Table 7.** Distribution of the 19 included reports across the coded variables. Percentages are relative to the 19 reports.

| Variable | Value | *n* | % |
|---|---|---|---|
| Prediction target | T3 failure or academic risk (early warning) | 15 | 79 |
| | T1 programme dropout | 3 | 16 |
| | T2 course withdrawal | 1 | 5 |
| Prediction horizon | Multiple cuts | 9 | 47 |
| | Retrospective | 6 | 32 |
| | Early | 1 | 5 |
| | Intermediate | 1 | 5 |
| | Late | 1 | 5 |
| | Not reported | 1 | 5 |
| C1. Trace source | Hybrid: LMS + institutional records | 7 | 37 |
| | LMS | 6 | 32 |
| | Includes external data | 3 | 16 |
| | Other interaction platform | 2 | 11 |
| | Hybrid: other platform + institutional records | 1 | 5 |
| Record type | Behavioural | 9 | 47 |
| | Both | 8 | 42 |
| | Assessment | 2 | 11 |

**Table 7.** *Cont.*

| Variable | Value | *n* | % |
|---|---|---|---|
| C2. Model family | Mixed (ML + DL) | 10 | 53 |
| | Classical ML | 9 | 47 |
| | Deep learning only | 0 | 0 |
| Mixed subtype | Parallel comparison ML vs DL | 10 | 53 |
| | ML → DL pipeline | 0 | 0 |
| | DL representation → ML | 0 | 0 |
| C3. Explainability | XAI reported | 11 | 58 |
| | No XAI | 8 | 42 |
| Function of XAI | Global understanding of risk factors | 6 | 32 |
| | Feature selection | 3 | 16 |
| | Several functions | 2 | 11 |
| | Model validation or debugging | 0 | 0 |
| | Individual explanations for users | 0 | 0 |
| Scope of XAI | Global | 10 | 53 |
| | Global and local | 1 | 5 |
| | Local | 0 | 0 |
| Validation scheme | Random hold-out | 7 | 37 |
| | *k*-fold CV | 7 | 37 |
| | Combined | 3 | 16 |
| | Temporal/cohort | 2 | 11 |
| Fairness evaluation | None | 14 | 74 |
| | Discusses risk of bias only | 5 | 26 |
| | Performance by protected subgroup | 0 | 0 |

### *3.3. Classification Along Analytical Categories*

This subsection reports the distribution of the 19 included reports across the three core analytical categories defined in Section 2.3. Tables 7 and 8 present every category in a single format, and Section 3.4 gives the best-performing model of each report.

**Table 8.** Integrated mapping of the 19 included reports, ordered by year. Target: T1 programme dropout, T2 course withdrawal, T3 failure or academic risk framed as an early warning (unit of the predicted event in parentheses). Flags (reports retained but removed in sensitivity analysis S1): $^{M}$ mixed or emergency remote modality; $^{G}$ results only in figures; $^{W}$ weak T3 framing; $^{I}$ weak evidence of interaction traces; $^{P}$ dataset includes some secondary-school records.

| Report | Year | Target | Horizon | Trace Source | Model Family | XAI (Function) | Validation |
|---|---|---|---|---|---|---|---|
| Corrigan et al. [63] $^{G}$ | 2015 | T3 (course) | Multiple cuts | Hybrid: LMS + institutional records | Classical ML | No | *k*-fold CV |
| Oreshin et al. [64] $^{I}$ | 2020 | T1 (programme) | Multiple cuts | Includes external data | Classical ML | Yes (global risk factors) | Temporal/cohort |
| Domladovac [47] $^{W}$ | 2021 | T3 (course) | Retrospective | Hybrid: LMS + institutional records | Mixed (ML + DL) | No | *k*-fold CV |
| Fahd and Miah [48] | 2023 | T3 (course) | Retrospective | Other interaction platform | Mixed (ML + DL) | No | *k*-fold CV |
| Matz et al. [30] | 2023 | T1 (programme) | Early | Hybrid: other platform + institutional records | Classical ML | Yes (global risk factors) | Combined |
| Poellhuber et al. [24] $^{M}$ | 2023 | T2 (course) | Late | LMS | Classical ML | No | Combined |

**Table 8.** *Cont.*

| Report | Year | Target | Horizon | Trace Source | Model Family | XAI (Function) | Validation |
|---|---|---|---|---|---|---|---|
| Kustitskaya et al. [27] | 2024 | T3 (semester) | Multiple cuts | Hybrid: LMS + institutional records | Classical ML | Yes (feature selection) | Random hold-out |
| Pecuchova and Drlik [23] $^{W}$ | 2024 | T3 (course) | Multiple cuts | LMS | Classical ML | No | Combined |
| Perkash et al. [62] $^{M}$ | 2024 | T3 (course) | Retrospective | Hybrid: LMS + institutional records | Mixed (ML + DL) | Yes (feature selection) | Random hold-out |
| Rico-Juan et al. [29] | 2024 | T3 (course) | Multiple cuts | Includes external data | Mixed (ML + DL) | Yes (several) | *k*-fold CV |
| Vaarma and Li [25] | 2024 | T1 (programme) | Multiple cuts | Hybrid: LMS + institutional records | Mixed (ML + DL) | Yes (global risk factors) | Temporal/cohort |
| Al-Gahmi [61] $^{M}$ | 2025 | T3 (course) | Multiple cuts | LMS | Mixed (ML + DL) | Yes (global risk factors) | Random hold-out |
| Bettahi et al. [26] | 2025 | T3 (course) | Not reported | LMS | Mixed (ML + DL) | Yes (several) | Random hold-out |
| Fahd and Miah [49] | 2025 | T3 (course) | Retrospective | LMS | Mixed (ML + DL) | Yes (feature selection) | *k*-fold CV |
| Madahana and Ekoru [28] | 2025 | T3 (course) | Retrospective | Hybrid: LMS + institutional records | Mixed (ML + DL) | No | *k*-fold CV |
| Mutka et al. [67] | 2025 | T3 (course) | Multiple cuts | Other interaction platform | Classical ML | No | *k*-fold CV |
| Nnamoko et al. [68] | 2025 | T3 (course) | Intermediate | Includes external data | Classical ML | No | Random hold-out |
| Pratiwi et al. [65] $^{P}$ | 2025 | T3 (course) | Retrospective | LMS | Mixed (ML + DL) | Yes (global risk factors) | Random hold-out |
| Schnedlitz et al. [66] | 2025 | T3 (course) | Multiple cuts | Hybrid: LMS + institutional records | Classical ML | Yes (global risk factors) | Random hold-out |

#### 3.3.1. Type of Digital Trace

LMS records are the most common source of interaction traces: they are used in 16 of the 19 reports, either alone (6) or combined with institutional records (7) or with external data (3). Institutional records added to the traces include demographic data, enrollment history, and prior academic records [25,27,28]. External data comprise neighbourhood statistics [68], public social-network profiles [64], and personality questionnaires administered for the study [29]. Three reports draw their interaction traces from platforms other than an LMS: an institutional mobile application [30], a programming environment [67], and an educational simulator [48]. The traces are behavioural (clicks, accesses, time on task) in nine reports and combine behavioural and assessment records in eight; two rely mainly on grades recorded on the platform [61,64].

#### 3.3.2. AI Model Family

When every model evaluated in a report is taken into account, including baselines, ten reports compare classical ML with DL models and nine use classical ML only. No report relies on DL alone: the report that proposes a neural ordinary differential equation model also evaluates tree-based and logistic baselines [26]. The mixed category is homogeneous in design: all ten reports run ML and DL models in parallel and compare their performance, and none chains the two families in a pipeline or uses a DL representation as input to an ML classifier. The best-performing or selected model is a tree-based ensemble (Random

Forest, gradient boosting, XGBoost, or CatBoost) in eight reports and a neural network in three (Section 3.4).

#### 3.3.3. Explainability Mechanism

Eleven of the 19 reports (58%) apply an explainability technique to a trained model or its predictions, most often feature importance from tree ensembles, permutation importance, or SHAP. Its function is mainly to describe which variables drive risk at the level of the whole model (six reports) or to select variables (three); two use it for several purposes [26,29]. Ten of the eleven apply it only globally, and one combines global and local explanations [29]. Explanations of individual predictions for students, advisers, or instructors, which are the form most relevant for acting on an early warning, are illustrated in two reports but not evaluated with users. RQ3 can therefore be answered with respect to the presence, function, and scope of explainability, but not with respect to its effect on decisions, which no report examines.

### *3.4. Performance Metrics Reported Across the Corpus*

The reports use a heterogeneous but overlapping set of evaluation metrics: classification accuracy, precision, recall, $F_1$, and the area under the receiver operating characteristic curve (AUC) dominate, complemented by regression metrics in the two reports that predict a continuous grade [29,65]. Table 9 summarizes the metrics, the best reported result, and the validation scheme of each report. The table is descriptive, and its values should not be compared across rows, because targets, metrics, horizons, and validation schemes differ.

**Table 9.** Performance metrics and representative results for the 19 included reports, ordered by year. The table is descriptive: best results are reported in the metric(s) adopted by each study and should not be interpreted as directly comparable across rows. Acc. = accuracy; Prec. = precision; Rec. = recall; FNR = false negative rate; AP = average precision; TPR/TNR = true positive/negative rate; CV = cross-validation.

| Report | Year | Target | Metric(s) | Best Reported Result | Validation |
|---|---|---|---|---|---|
| Corrigan et al. [63] | 2015 | T3 | ROC AUC by week | AUC > 0.60 from week 4 (one module; values shown only in a figure) | *k*-fold CV |
| Oreshin et al. [64] | 2020 | T1 | AUC, Acc., Prec., Rec., $F_1$, nRMSE | AUC = 0.97; Acc. = 0.91; Rec. = 0.83 (CatBoost, static dropout model) | Temporal/cohort |
| Domladovac [47] | 2021 | T3 | FNR, Rec., Acc. | FNR = 0.13; Acc. = 0.86 (gradient boosted trees) | *k*-fold CV |
| Fahd and Miah [48] | 2023 | T3 | Acc. | Acc. = 98.7% (MLP after SMOTE); 85.7% (RF, before augmentation) | *k*-fold CV |
| Matz et al. [30] | 2023 | T1 | AUC, $F_1$, TPR, TNR | AUC = 0.88 (RF, one university); mean AUC 0.73; cross-university transfer AUC 0.63 | Combined |
| Poellhuber et al. [24] | 2023 | T2 | Acc., Prec., Rec., $F_1$, AUC | Mean Acc. = 0.95; mean Rec. = 0.27 | Combined |
| Kustitskaya et al. [27] | 2024 | T3 | Acc., Prec., Rec., weighted $F_1$ | Weighted $F_1$ = 0.875–0.98 from week 2 to semester end (stacked ensemble) | Random hold-out |

**Table 9.** *Cont.*

| Report | Year | Target | Metric(s) | Best Reported Result | Validation |
|---|---|---|---|---|---|
| Pecuchova and Drlik [23] | 2024 | T3 | Prec., Rec., $F_1$, AUC; clustering indices | $F_1 = 0.71$–$0.97$ for high-risk clusters; AUC = 0.93–0.99 with full-semester data (cluster-membership target) | Combined |
| Perkash et al. [62] | 2024 | T3 | Acc., Prec., Rec., $F_1$ | Acc. = 99.1%; $F_1 = 0.99$ (decision tree after SMOTE) | Random hold-out |
| Rico-Juan et al. [29] | 2024 | T3 | RMSE (regression) | RMSE = 1.42 on a 0–10 grade scale (ARD) | *k*-fold CV |
| Vaarma and Li [25] | 2024 | T1 | AUC, AP, $F_1$, Prec., Rec. | AUC = 0.853; Prec. = 0.81; Rec. = 0.47 (CatBoost, mean of 60 monthly models) | Temporal/cohort |
| Al-Gahmi [61] | 2025 | T3 | Acc., Prec., Rec., $F_1$ | $F_1 = 0.83$ (MLP); Acc. = 0.94 (LR) | Random hold-out |
| Bettahi et al. [26] | 2025 | T3 | Macro $F_1$, class-wise $F_1$, AUC | Macro $F_1 = 0.87$; AUC = 0.93 (CR-NODE) | Random hold-out |
| Fahd and Miah [49] | 2025 | T3 | Acc., Prec., $F_1$ | Acc. = 99.8% (LSTM after SMOTE); 81.7% without augmentation | *k*-fold CV |
| Madahana and Ekoru [28] | 2025 | T3 | Acc. | Acc. = 98.5% (attributed to DT in the results table and to RF in the abstract) | *k*-fold CV |
| Mutka et al. [67] | 2025 | T3 | Class-wise $F_1$, Acc. | $F_1$ (fail) = 0.93 and $F_1$ (pass) up to 0.99 by week 5; combined model $F_1 = 1.00$ from week 5–6 (XGBoost) | *k*-fold CV |
| Nnamoko et al. [68] | 2025 | T3 | Macro $F_1$, Acc. | Macro $F_1 = 0.81$; Acc. = 0.91; 7 of 12 failing students detected (SVM, week 6) | Random hold-out |
| Pratiwi et al. [65] | 2025 | T3 | $R^2$, RMSE, MAE (regression) | $R^2 = 0.81$; RMSE = 7.12 (RF) | Random hold-out |
| Schnedlitz et al. [66] | 2025 | T3 | Acc., $F_1$, Prec., Rec. | Week 2: Acc. = 0.80, $F_1 = 0.86$ (DT); full course: Acc. = 0.93 (linear SVM) | Random hold-out |

Classification accuracy, defined as the proportion of correctly labeled instances over the total, is the most frequently reported measure [28,48,49,61,62,66], with reported values spanning a wide range: from 0.80 at week 2 of a programming course [66] to near-perfect figures exceeding 98% when the minority class is oversampled [48,49,62]. Several authors caution against relying solely on accuracy when the underlying class distribution is imbalanced, a condition common in dropout datasets where non-dropout students typically form the majority class [25,26,68]. In such contexts, a trivial classifier that always predicts the majority class can attain a deceptively high accuracy, motivating the adoption of complementary metrics.

Precision and recall (also referred to as the true positive rate or sensitivity) are therefore widely reported alongside accuracy. Precision quantifies the fraction of predicted positive cases that are truly positive, whereas recall measures the fraction of actual positive cases that

the model successfully identifies. Both metrics are derived from the confusion matrix entries of true positives (TP), false positives (FP), and false negatives (FN), as defined in (1):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{1}$$

Reported precision ranges from about 0.60 for a high-risk cluster [23] to above 0.99 in oversampled settings [49], and reported recall from 0.27 [24] to 1.00 [62], reflecting the trade-off between the cost of false alarms and the cost of missed at-risk students.

The F1 score, the harmonic mean of precision and recall defined in (2), is the single most consistently adopted summary metric across the corpus [23–26,30,67,68]:

$$\text{F}_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

Its capacity to balance precision and recall makes it particularly suitable when class imbalance is present. Bettahi et al. [26] adopt the Macro F1 variant, which averages the per-class F1 score and thus assigns equal weight to the minority (failing) and majority (passing) classes; they argue that this is more informative than overall accuracy for imbalanced educational datasets, and report that their proposed CR-NODE model achieves a Macro F1 of 0.875, outperforming LSTM (0.812) and XGBoost (0.830) on the same test set. Mutka et al. [67] report F1 scores of up to 0.99 for pass/fail classification in introductory programming courses using XGBoost on behavioral time-series features.

The area under the receiver operating characteristic curve (AUC-ROC) is the second most commonly used metric, preferred by studies dealing with class-imbalanced data because it evaluates the model's discriminative ability across all possible classification thresholds without depending on a single decision boundary [25,26,30,64]. AUC values reported in the corpus range from 0.63 when models are transferred across universities [30] to 0.97 for a programme-dropout model evaluated with temporal validation [64]; these values come from different targets and designs and do not establish the superiority of any model family.

For regression-oriented formulations used in studies predicting continuous outcomes such as final exam scores, classification metrics are replaced by the root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination ($R^2$) [29,65]. Pratiwi et al. [65] report that Random Forest achieves the best regression performance (RMSE = 7.12, MAE = 5.46, $R^2 = 0.81$) among three models evaluated on a dataset of 10,000 student records.

The validation schemes behind these results are described in Section 3.2; most are random hold-out splits or $k$-fold cross-validation within a single cohort, which estimate performance on students similar to those used for training rather than on future cohorts. Several studies also apply class-imbalance correction techniques, most notably the Synthetic Minority Oversampling Technique (SMOTE) and its variants, before model training, reporting that these preprocessing steps substantially improve recall and F1 scores for the minority (at-risk) class [49,62,68]; when oversampling is applied before the data are split, synthetic cases can reach the test set and inflate these gains (Section 4.4).

What drives the variation in reported performance. Reported results range from a recall of 0.27 [24] to accuracies above 98% [28,48,49,62]. Four features of the reports are associated with this range more clearly than the model family. First, the target: course-level failure in small cohorts, often defined with grade components that also enter the predictors, yields the highest figures, whereas programme-level dropout predicted months in advance yields more moderate discrimination in two of the three T1 reports (AUC between 0.73 and 0.88) [25,30]. Second, the horizon: where results are reported by time cut, performance rises as the period advances and more of the outcome is observed [25,27,66]. Third, the evaluation

design: three of the four near-ceiling accuracies were obtained with oversampling applied before the data were split, and none of the four under temporal or external validation. The cohort-based evaluation of monthly models and the cross-university transfer of models give the most conservative estimates [25,30]. Fourth, the metric: accuracy alone conceals poor detection of the minority class, as shown by a report that combines a mean accuracy of 0.95 with a mean recall of 0.27 [24]. Dataset size and trace richness do not show a consistent relation with reported performance in this corpus. These associations are descriptive and based on few reports per configuration; they indicate why the headline figures cannot be compared, not which design yields the best predictions.

### *3.5. Sensitivity Analyses*

Table 10 reports the main counts under the five sensitivity analyses. None of the central findings changes direction. The majority of reports with an explainability mechanism no longer depends on a single research group (10 of 17 without it), no report evaluates fairness across protected subgroups in any analysis, and the absence of temporal or external validation remains the most frequent quality signal. Restricting the corpus to unflagged reports raises mean reporting completeness from 71.6% to 75.9% of the applicable items. The four reports with programme dropout or course withdrawal targets differ most from the rest: all four use temporal or combined validation, and their mean reporting completeness is the highest (79.7%), although four reports are too few for comparison. Restricting the corpus to journal articles (S5) leaves the direction of every central finding unchanged: 9 of the 11 reports model a T3 target, 7 report explainability, none evaluates fairness across protected subgroups, and mean reporting completeness is 74.2%.

**Table 10.** Sensitivity analyses. S1: without the eight flagged reports; S2: without the two reports from one research group; S3: one report per dataset; S4: programme dropout and course withdrawal targets only (T1–T2); S5: journal articles only. Counts are numbers of reports; TRIPOD+AI completeness is the mean percentage of applicable items reported.

| Indicator | Main ($n = 19$) | S1 ($n = 11$) | S2 ($n = 17$) | S3 ($n = 18$) | S4 ($n = 4$) | S5 ($n = 11$) |
|---|---|---|---|---|---|---|
| T3 target (failure or academic risk) | 15 | 9 | 13 | 14 | 0 | 9 |
| LMS-based traces (alone or combined) | 16 | 8 | 15 | 15 | 3 | 8 |
| Mixed ML + DL family | 10 | 6 | 8 | 9 | 1 | 6 |
| XAI reported | 11 (58%) | 7 (64%) | 10 (59%) | 10 (56%) | 3 | 7 (64%) |
| Fairness evaluated across protected subgroups | 0 | 0 | 0 | 0 | 0 | 0 |
| Temporal or cohort-based validation | 2 | 1 | 2 | 2 | 2 | 1 |
| Q4 no temporal or external validation | 16 | 9 | 14 | 15 | 1 | 9 |
| Q7 internal contradictions | 16 | 8 | 14 | 15 | 4 | 9 |
| TRIPOD+AI completeness (mean %) | 71.6 | 75.9 | 72.3 | 72.5 | 79.7 | 74.2 |

### *3.6. Reporting Biases and Confidence in the Findings*

The qualitative assessment of reporting bias identified signals along all four dimensions. Regarding publication bias favoring high-performing models, five reports present best results at or near the ceiling (accuracy or $F_1$ of at least 0.98) [28,48,49,62,67], one of them on a single held-out split [62], whereas three disclose a minority-class recall below 0.60 [24,25,68]. Regarding selective outcome reporting, AUC is listed among the metrics in 7 of the 17 reports that evaluate a classifier, and two report accuracy as their only quantitative metric [28,48]. Regarding coverage, the restriction to English-language records in Scopus and WoS is consistent with the absence of reports led from Latin America and the scarcity of contributions from other regions with high enrollment (Section 4.7). Finally, the TRIPOD+AI audit records what the reports describe, not what their authors did, so its

low rates for sample-size justification, ethical approval, and fairness evaluation are lower bounds on the underlying practices. Taken together, these signals suggest that the published literature may overstate the predictive performance attainable in institutional settings.

Table 11 summarizes the confidence ratings for the main findings. Findings that consist of direct counts of targets, data sources, and model families are rated with high confidence. These ratings refer to direct descriptive counts within this corpus, under the stated eligibility criteria and coding rules; they do not estimate the prevalence of these patterns across the wider literature on student dropout prediction. Findings on reporting practices are rated as moderate, because non-reporting cannot be distinguished from non-performance; the finding on the absence of temporal or external validation is rated as moderate because agreement on the coding of quality signals was limited; and comparative statements about the relative predictive performance of model families are rated as very low.

**Table 11.** Confidence in the main findings of the review, rated with domains adapted from GRADE. Ratings apply to this corpus under the stated eligibility criteria; they are not estimates for the wider literature. RoB = methodological limitations of the contributing reports; Inc. = inconsistency; Ind. = indirectness; Imp. = imprecision; Rep. = reporting bias. ✓ = no serious concern; ↓ = serious concern (downgraded one level).

| Finding | Reports | RoB | Inc. | Ind. | Imp. | Rep. | Confidence | Rationale |
|---|---|---|---|---|---|---|---|---|
| Reports framed as dropout prediction mostly model course failure (target substitution) | 19 | ✓ | ✓ | ✓ | ✓ | ✓ | High | Direct count (15/19; 9/11 without flagged reports); conditional on the T3 eligibility rule |
| LMS records are the most common source of interaction traces (RQ1) | 19 | ✓ | ✓ | ✓ | ✓ | ✓ | High | Direct count (16/19), conditional on the interaction-trace criterion |
| Half of the reports compare ML and DL in parallel; none relies on DL alone (RQ2) | 19 | ✓ | ✓ | ✓ | ✓ | ✓ | High | Direct count (10/19), coded from all models evaluated |
| Explainability is used by a majority of reports, mainly for global risk factors (RQ3) | 19 | ✓ | ↓ | ✓ | ✓ | ✓ | Moderate | 11/19 (10/17 without one group); heterogeneous techniques and functions |
| Sample-size justification and ethical approval are rarely reported (RQ4) | 19 | ✓ | ✓ | ✓ | ✓ | ↓ | Moderate | 2/19 and 8/19; non-reporting cannot be distinguished from non-performance |
| No report evaluates fairness across protected subgroups (RQ4) | 19 | ✓ | ✓ | ✓ | ✓ | ↓ | Moderate | 0/19 in all analyses; absence of reporting, not evidence of biased models |
| Most reports lack temporal or external validation (RQ4) | 19 | ✓ | ✓ | ✓ | ✓ | ↓ | Moderate | 16/19; quality signals coded with limited agreement before consensus |
| Relative predictive performance of model families | 17 | ↓ | ↓ | ↓ | ✓ | ↓ | Very low | Heterogeneous targets, metrics, and validation; near-ceiling results after oversampling; selective metric reporting |

## 4. Discussion

This section interprets the findings along seven lines. The first examines the mismatch between the dropout framing of the reports and the outcome they predict. The second identifies the cross-cutting patterns that link the analytical categories, and the third examines the theoretical grounding of dropout prediction in the corpus. The fourth reports the audit of reporting completeness against the adapted TRIPOD+AI items and the analytical quality signals. The fifth examines the ethical issues raised by predicting dropout from digital

traces, the sixth considers what these findings imply for sustainable and equitable higher education, and the seventh discusses the limitations of the review. Future research directions are integrated into the corresponding lines rather than treated as a separate agenda.

### *4.1. What Is Predicted: Course Failure in Place of Dropout*

The most consistent finding of the review concerns the outcome itself. The reports were retrieved with a search that required a dropout, attrition, or retention term, and each presents its model as a contribution to dropout prediction or early warning. Yet 15 of the 19 model course failure or academic risk, three model programme dropout, and one models course withdrawal (Section 3.2); the proportion is similar when the flagged reports are removed (9 of 11; Table 10). The literature indexed as dropout prediction from digital traces is therefore, for the most part, a literature on predicting course failure, in which dropout is the motivating outcome rather than the observed one. We describe this pattern as target substitution. It emerged when the targets were made explicit during the re-assessment of the corpus, not from a research question defined in advance, and it describes this body of evidence; whether it holds in the wider literature, including online education, would require a search designed for that purpose.

The pattern has a plausible explanation in institutional infrastructure. Course grades are recorded on the platform within weeks, whereas programme dropout is observed only after months or years and is defined differently across institutions, so a course-level target can be built from the data of a single semester and a single platform. Failing courses is widely regarded as an antecedent of leaving, and the proxy is defensible as a first step. It changes, however, the meaning of the evidence. A model that predicts failure in one course identifies students at academic risk, not students who will leave the institution, and its value as a dropout warning depends on a link that the 15 reports do not evaluate, because their models are trained and assessed against course outcomes rather than observed departure. The conditions under which the warnings were evaluated reinforce this: six of the 15 use data from the whole period, so they cannot show how early a warning could be issued, and neither of the two temporal or cohort-based validations in the corpus concerns a course-failure target (Section 3.2).

Target substitution also bears on equity. Students who leave in good academic standing, for financial, family, or work reasons or because of weak social integration [1], produce no failure signal, so a warning built on course failure cannot identify them by construction. If such reasons are more frequent among the groups that SDG target 4.5 seeks to protect, the blind spot would fall on them. This is an implication of how the target is defined, not a result observed in the corpus, but it compounds the absence of evaluation across protected subgroups (Section 4.4): the reports neither test whether their models work equally well for all students nor model the outcome that would reveal students who leave for non-academic reasons.

Three consequences follow for future studies. The target should be defined operationally and named for what it is, so that course-failure models are reported as academic-risk prediction rather than as dropout prediction (TRIPOD+AI item 8a); the level of the target (programme, course, or failure) should be stated explicitly; and the value of academic-risk warnings for retention should be tested by linking predicted failure to later enrolment status on subsequent cohorts. For reviews of this field, the finding implies that counts of "dropout prediction" studies overstate the evidence on dropout itself, and that the target of each study should be extracted and reported.

### *4.2. Cross-Cutting Patterns Across the Analytical Categories*

The analytical categories are linked by a common dependence on institutional infrastructure, which shapes both what can be predicted and how predictions can be used. LMS records appear in 16 of the 19 reports because they are generated automatically by systems that universities already operate, but they describe activity within the formal virtual classroom and omit other dimensions of the student experience, such as social integration or administrative interaction. Combining interaction traces with institutional records (8 reports) or external data (3) recovers part of that space, at the cost of additional demands on data integration, consent, and governance. The three reports that do not use an LMS show that other platforms, such as an institutional mobile application [30], a programming environment [67], or an educational simulator [48], can also carry predictive signal, but they are too few to establish whether richer or different traces improve early identification.

The same infrastructure shapes the prediction target, which explains in part why most reports model course failure rather than dropout (Section 4.1).

The distribution of model families follows a similar logic. Half of the reports compare classical ML with DL models in parallel, and none relies on DL alone or combines the two families in a pipeline. In these comparisons, the neural model is one candidate among several rather than a design choice motivated by the structure of the data, and the best result is a tree-based ensemble more often than a neural network. The corpus therefore does not support the conclusion that DL adds value for dropout prediction from digital traces, nor the opposite: the comparisons are made on small, single-institution datasets, under different targets and validation schemes, and often with resampling that inflates the results of all models (Section 4.4).

Explainability is reported in 11 of the 19 reports, and this majority does not depend on a single research group (10 of 17 without it). Its use, however, is narrow. Explanations describe which variables matter for the model as a whole, or serve to select variables; individual explanations for the people who must act on a warning are illustrated in two reports and evaluated in none. The translation of an explanation into an action, such as a counselling appointment or a curricular adjustment, is not documented in the corpus. Dropout prediction is thus treated as a classification problem with an interpretability layer added afterwards, rather than as a decision-support system designed with its users.

Together, these patterns trace a single tension. The reports privilege what is accessible and technically defensible within existing infrastructure: LMS data, course-level targets, comparisons of familiar algorithms, and global feature importance. This is a reasonable response to institutional conditions, but it explains why the field has accumulated predictive results without establishing the conditions under which they would be useful for retention. Progress would require targets defined at programme level or explicitly linked to later dropout, validation on later cohorts, comparisons of model families designed around specific data structures, and explanations evaluated with the advisers and students who receive them.

### *4.3. Theoretical Grounding of Dropout Prediction*

Another pattern that emerges from the corpus concerns the limited use of explicit theoretical frameworks for understanding and contextualizing the dropout phenomenon. The majority of the reviewed studies approach dropout prediction as a technical classification problem, specifying data sources, model architectures, and performance metrics, without situating their predictive targets within established theoretical models of student departure.

The seminal contributions in this area include Tinto's Student Integration Model [1], which proposes that departure is shaped by students' academic and social integration into

the institutional community; Bean's Student attrition model [9], which draws on organizational turnover theory to explain the role of institutional fit, satisfaction, and external environmental variables on withdrawal intentions; and the integrated model presented in Cabrera et al. [69], which combines elements of both frameworks through structural equation modeling to account for financial and environmental pressures alongside academic and social integration. Braxton's subsequent synthesis [70] interrogated the empirical viability of these frameworks and proposed refinements appropriate to the diverse institutional and student landscapes of the twenty-first century.

Among the reviewed studies, only Matz et al. [30] explicitly grounds its predictive design in Tinto's [1,71] conceptualization of social integration. The study cites Cabrera et al. [69] to motivate the inclusion of behavioral engagement features as proxies for students' social embeddedness within their institution, and interprets campus event attendance, peer network centrality, and community interaction metrics as operationalizations of social integration. This is precisely the type of explicit theory-to-data mapping that the broader corpus lacks. Two studies offer partial grounding: Poellhuber et al. [24] conceptualize dropout as the administrative culmination of a disengagement process and use the multidimensional definition of behavioral engagement [72] to select Moodle variables, establishing an implicit process-oriented lens without invoking the canonical models of Tinto or Bean; Vaarma and Li [25] cite Tinto's foundational work in their reference list and acknowledge in their limitations section that survey-based psychosocial variables are absent from their predictive models, implying awareness of the gap between available data and theoretically motivated predictors, although the study's analytic design does not engage with Tinto's framework. The remaining 16 reports approach dropout prediction without reference to any theoretical model of student departure [23,26–29,47–49,61–68]. These studies treat dropout prediction as a classification or regression task, specifying features, algorithms, and performance metrics without articulating which theoretical constructs their variables are intended to operationalize or how their findings confirm, challenge, or refine existing models of student attrition.

The consequence of this absence is substantive. Without theoretical grounding, predictive models may identify statistically significant associations without explaining the underlying causal mechanisms, which limits their contribution to theory development and their usefulness for designing targeted retention interventions [70,73]. The interpretive ambiguity is concrete: a login frequency variable may reflect academic disengagement, motivational withdrawal, or external socioeconomic pressures; without anchoring the variable to an established risk framework, it is difficult to assess which of these mechanisms the model is detecting and, consequently, what type of institutional intervention would be most appropriate following a risk alert [1,69]. The same ambiguity applies to engagement metrics, grade trajectories, and other digital trace variables that proliferate in the corpus.

Hybrid theory-data approaches contend that digital trace data can serve as operationalizations of theoretically motivated constructs, for instance engagement as a proxy for social integration and grade trajectories as a proxy for academic integration, but only when the mapping between data and theory is made explicit [1,69]. Matz et al. [30] illustrates the viability of this approach and may serve as a model for future studies seeking to combine predictive accuracy with theoretical interpretability. Future systematic reviews in this domain should incorporate theoretical grounding as an explicit analytical dimension; primary studies should articulate which theoretical constructs their predictive variables are intended to operationalize, and should discuss how their findings confirm, challenge, or refine existing theoretical models of student departure. Such integration would elevate dropout prediction research from a purely technical enterprise toward one that advances both scientific understanding and institutional practice.

### *4.4. Reporting Completeness Under TRIPOD+AI and Analytical Quality Signals*

The methodological gaps identified in the cross-cutting analysis are not unique to the learning analytics field. A structurally analogous set of reporting deficiencies in clinical prediction research motivated the development of the TRIPOD+AI statement [33], an international reporting guideline published in 2024 that updates and supersedes the original TRIPOD 2015 statement. TRIPOD+AI was developed through systematic literature reviews, two rounds of a modified Delphi exercise involving more than 200 experts from 27 countries, and an international consensus meeting; the resulting checklist contains 27 items and 52 subitems covering all stages of a prediction model study, from title and methods through to data and code availability, with explicit emphasis on fairness, open science practices, and end-user involvement. Its stated purpose is to ensure that prediction model studies are complete, transparent, and reproducible, so that they can be critically appraised and safely implemented.

The gaps identified in the present review map directly onto the reporting dimensions addressed by TRIPOD+AI. Heterogeneity in the definition of the target outcome (dropout, retention, academic risk) corresponds to items 8a–8b, which require a precise operational definition of the outcome, the time horizon of prediction, and a justification for the choice of definition. The restricted transferability of single-institution models relates to items 5a, 6a, and 16, which mandate explicit description of the data sources, the institutional setting, and any differences between development and evaluation datasets that may affect external validity. Incomplete reporting of data provenance, validation procedures, and model tuning aligns with items 7, 11, 12a–12g, and 18c–18f, which require documentation of data preprocessing, handling of missing values, all model-building steps including hyperparameter tuning, internal validation strategy, study registration, and availability of analytical code and data. The gap between prediction and intervention, defined as the limited translation of model outputs into institutional retention actions, is addressed by items 15 and 27a–27c, which specify that studies must describe the model output format, the expertise required of end users, and the practical conditions under which model predictions are intended to inform decisions.

Beyond these direct correspondences, the adoption of TRIPOD+AI in the learning analytics literature would introduce two reporting dimensions that are largely absent from current studies. Item 14 requires authors to describe approaches used to assess and promote fairness: whether the model performs equitably across student subgroups defined by gender, socioeconomic status, ethnicity, study modality, or first-generation status. This is particularly consequential in dropout prediction, where digital trace data may systematically underrepresent students with limited platform access, thereby producing biased risk estimates for already-vulnerable populations [74–77]. Algorithmic bias in educational AI has been documented across grading, admissions, recommender systems, and engagement modeling, with consistent findings that models trained on platform-mediated traces tend to penalize students whose engagement patterns deviate from majority profiles for reasons unrelated to academic disengagement [74,78]. Item 19 introduces a requirement for stakeholder involvement, prompting authors to report whether students, academic advisors, or institutional coordinators participated in the design, interpretation, or dissemination of the model. This dimension bears directly on the practical uptake and equitable deployment of predictive systems in higher education [79]. Both clinical and educational prediction models draw on longitudinal observational data, apply machine learning algorithms to estimate the probability of an outcome in a target population, and aim to inform a timely intervention. Given this conceptual symmetry, TRIPOD+AI offers a reasonable starting point for reporting AI-based dropout prediction, provided that its adaptation to education is examined and refined.

To examine the completeness of reporting, the 16 Methods items of TRIPOD+AI were adapted to AI-based dropout prediction from digital traces and applied to the 19 included reports. The adapted items are listed in Table 12; the decision rule used for each item and the scoring procedure are given in Table 13; and the correspondence between each adapted item and the original TRIPOD+AI wording, with the justification for every change, is provided as Supplementary Table S4. Each item was coded as reported (1), not reported (0), or not applicable (NA); the score of a report is the number of items reported divided by the number of applicable items, with no weighting and no partial credit. A score expresses how much of the information needed to appraise a study is present in the report. It does not measure methodological quality: a report that describes a weak missing-data strategy fulfills item 11 as fully as one that describes a sound strategy, and an item coded as not reported does not show that the step was omitted. The compliance matrix is reported in Table 14.

**Table 12.** TRIPOD+AI checklist, Methods items adapted to AI-based university student dropout prediction using digital traces.

| Item | Scope | Topic | TRIPOD+AI Item Description |
|---|---|---|---|
| 5a | D; E | Data sources | Describe digital trace sources separately for development and evaluation sets (LMS, e.g., Moodle, Canvas; academic information system; videoconferencing platforms; digital library; electronic attendance system); justify their use; and discuss the representativeness of the data with respect to the diversity of the student population. |
| 5b | D;E | Study period | Specify the academic period(s) covered (semesters or academic years), the start and end dates of digital trace collection, and, if applicable, the follow-up period to confirm the dropout outcome. |
| 6a | D; E | Setting | Specify the institutional setting (public/private, face-to-face/virtual/hybrid; country or region; number of campuses) and the academic program(s) or courses included in the study. |
| 6b | D; E | Eligibility | Describe student inclusion and exclusion criteria (e.g., first-year cohorts, actively enrolled students, minimum registered digital activity, exclusion of students on temporary leave or with documented exceptional circumstances). |
| 7 | D; E | Data preparation | Describe the preprocessing of digital traces (log cleaning, temporal aggregation, feature engineering, normalization) and data quality verification, including whether the process was consistent across sociodemographically differentiated student groups and across platforms or data sources. |
| 8a | D; E | Outcome | Provide an operational definition of student dropout according to institutional criteria (e.g., non-enrollment in the following semester, definitive withdrawal, failure to complete minimum credits); specify the prediction time horizon; justify the definition; and verify that it is applied consistently across student groups. |
| 9ab | D | Predictors | Describe the selection of digital traces as initial predictors (based on learning analytics literature, prior models, or the full set of available variables) and any pre-selection filter applied before model construction (e.g., data completeness threshold, minimum variance, theoretical relevance). Define all predictors clearly, specifying how and when each trace was recorded and the observation window used. |

**Table 12.** *Cont.*

| Item | Scope | Topic | TRIPOD+AI Item Description |
|---|---|---|---|
| 10 | D; E | Sample size | Explain how the study size (number of students and dropout events) was determined for model development and evaluation; justify its sufficiency for the complexity of the model employed; report the events-per-predictor ratio; and include any formal sample-size calculation. |
| 11 | D; E | Missing data | Describe how missing digital trace data were handled, recognizing that the absence of records may not be random (students at dropout risk may generate less digital activity). Specify whether imputation, case deletion, or missing-data-tolerant models were used, and justify the strategy adopted. |
| 12a | D | Analytical methods | Describe how the digital trace data were partitioned for model development and evaluation (e.g., chronological split by cohort, stratified random partition, *k*-fold cross-validation); justify the strategy based on sample size and the risk of temporal data leakage. |
| 12c | D | Analytical methods | Specify the model type (e.g., logistic regression, decision tree, random forest, XGBoost, LSTM neural network), the justification for its selection, all model-building steps, the hyperparameter tuning procedure, and the internal validation method employed (e.g., *k*-fold cross-validation, leave-one-semester-out). |
| 12e | D; E | Analytical methods | Specify all metrics and plots used to evaluate model performance: discrimination (AUC-ROC, AUC-PR), calibration (calibration curve, Brier score), and practical utility (decision curve analysis, F1 score, sensitivity and specificity at selected classification thresholds). Justify the choice of metrics based on class imbalance and the intended model use. |
| 13 | D; E | Class imbalance | If the proportion of dropout students is a minority relative to non-dropouts, indicate which class-imbalance techniques were applied (e.g., SMOTE, undersampling, minority-class overweighting, threshold modification) and describe any subsequent model recalibration. |
| 14 | D; E | Fairness | Describe the approaches used to evaluate and promote model fairness across student groups (by gender, socioeconomic status, ethnicity, study modality, first-generation student status, technology access). Specify whether fairness metrics (e.g., demographic parity, equal opportunity) were assessed and what actions were taken to mitigate identified biases in the digital traces. |
| 15 | D | Model output | Specify the model output (continuous dropout probability, binary dropout/non-dropout classification, or risk categories: low/medium/high). Provide details and justification for any classification threshold adopted, indicating whether it was optimized for sensitivity, specificity, F1, or another metric, and on which dataset it was determined. |
| 17 | D; E | Ethical approval | Name the committee or institutional body that approved the use of student digital trace data. Describe the informed consent mechanism or institutional privacy policy that covers the use of these data, including the applicable regulatory framework (e.g., GDPR in Europe, FERPA in the US, national personal data protection legislation). |

Note: Methods items adapted to AI-based university student dropout prediction using digital traces. Scope: D = development only; D; E = development and evaluation.

**Table 13.** Decision rules for the adapted TRIPOD+AI Methods items and scoring procedure. Each item was coded 1 (reported), 0 (not reported), or NA (not applicable). Score of a report = items coded 1/items applicable; no weights and no partial credit. The rules were fixed before coding (rule P7) and clarified during consensus for items 12c, 13, and 14 (dates in the codebook).

| Item | Topic | Coded 1 When the Report ... |
|---|---|---|
| 5a | Data sources | names the platforms or systems from which the predictors were obtained |
| 5b | Study period | states the academic years, semesters, or dates of the data; "several periods" without detail = 0 |
| 6a | Setting | states the country or institution and the programme or course |
| 6b | Eligibility | gives at least one explicit inclusion or exclusion criterion for students; the number of students alone = 0 |
| 7 | Data preparation | describes cleaning, aggregation, feature engineering, or normalisation of the traces |
| 8a | Outcome | defines the predicted event with a criterion or threshold and its timing; a label alone ("pass/fail", "dropout") = 0 |
| 9ab | Predictors | lists and defines the predictors used |
| 10 | Sample size | gives a calculation, events per predictor, or an explicit argument of sufficiency; reporting *n* alone = 0 |
| 11 | Missing data | describes how missing data were handled, or states that they were checked and absent |
| 12a | Data partitioning | states the partitioning scheme (proportions or number of folds) |
| 12c | Model building | names the models and describes hyperparameter tuning or internal validation; an explicit statement that default hyperparameters were used counts as reported; naming the models alone = 0 |
| 12e | Performance metrics | specifies the metrics and at least one is not accuracy, or justifies the choice |
| 13 | Class imbalance | reports a technique (resampling, weights, threshold); NA when the authors show balanced classes or the task is a regression |
| 14 | Fairness | reports performance by protected subgroup, a fairness metric, or mitigation; discussing the risk, or subgroups defined by behaviour, = 0 |
| 15 | Model output | states the form of the output (probability, class, risk category) |
| 17 | Ethics | names an ethics committee or approval, or states that consent was obtained; anonymisation or an exemption alone = 0 |

Across the 19 reports, 215 of the 300 applicable item ratings were coded as reported (71.7%). The mean score was 11.3 items (SD 1.8), or 71.6% of the applicable items, with individual scores ranging from 7 of 15 [65] to 14 of 16 [25] and 14 of 15 [27]. The three reports published before 2022 reached a mean of 68% of applicable items, compared with 72% for those published from 2023 onward; with three reports in the earlier group, this difference does not establish a trend, and any change over time would more plausibly reflect the diffusion of reporting norms than an improvement in the underlying methods.

Seven items were reported by all or nearly all reports: data sources (5a), preprocessing (7), predictor definition (9ab), data partitioning (12a), and model output (15) by all 19, performance metrics (12e) by 18, and model building with tuning or internal validation (12c) by 17. These items describe the core technical workflow, whose reporting has become standard practice.

**Table 14.** TRIPOD+AI Methods reporting matrix for the 19 included reports, ordered by year. 1 = reported; 0 = not reported (which does not imply that the step was not performed); NA = not applicable. Score = items reported/items applicable.

| Report | 5a | 5b | 6a | 6b | 7 | 8a | 9ab | 10 | 11 | 12a | 12c | 12e | 13 | 14 | 15 | 17 | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Corrigan et al. [63] | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 11/16 |
| Oreshin et al. [64] | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 11/16 |
| Domladovac [47] | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | NA | 0 | 1 | 0 | 10/15 |
| Fahd and Miah [48] | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 9/16 |
| Matz et al. [30] | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 13/16 |
| Poellhuber et al. [24] | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 13/16 |
| Kustitskaya et al. [27] | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | NA | 0 | 1 | 1 | 14/15 |
| Pecuchova and Drlik [23] | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 11/16 |
| Perkash et al. [62] | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 9/16 |
| Rico-Juan et al. [29] | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | NA | 0 | 1 | 0 | 10/15 |
| Vaarma and Li [25] | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 14/16 |
| Al-Gahmi [61] | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 11/16 |
| Bettahi et al. [26] | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 11/16 |
| Fahd and Miah [49] | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 12/16 |
| Madahana and Ekoru [28] | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 11/16 |
| Mutka et al. [67] | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 13/16 |
| Nnamoko et al. [68] | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 13/16 |
| Pratiwi et al. [65] | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | NA | 0 | 1 | 0 | 7/15 |
| Schnedlitz et al. [66] | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 12/16 |
| Reported | 19/19 | 13/19 | 16/19 | 12/19 | 19/19 | 12/19 | 19/19 | 2/19 | 12/19 | 19/19 | 17/19 | 18/19 | 10/15 | 0/19 | 19/19 | 8/19 | – |

Six items were reported by between 63% and 84% of the reports to which they applied. The institutional context (6a) was reported by 16. The study period (5b) was reported by 13. Student eligibility criteria (6b) and the operational definition of the outcome with its threshold and timing (8a) were each reported by 12, so a third of the reports do not state precisely which students were modelled or what counted as the event predicted. Missing-data handling (11) was reported by 12. Class-imbalance handling (13) was reported by 10 of the 15 reports in which it applied. The omission of missing-data handling matters in this domain, because the absence of activity records is unlikely to be random: students at highest risk tend to disengage from the platform and leave fewer traces.

Three items were reported by fewer than half of the reports. Ethical approval or consent (17) was documented by 8 of the 19, so most reports do not name an ethics committee or a consent procedure for the use of student behavioural records; this is evidence about reporting, and approval may have been obtained without being stated. A justification of sample size (10) was given by two. No report described an evaluation of model performance across demographic, socioeconomic, or other protected subgroups (14; 0 of 19). Five reports discuss the risk of bias in their limitations, and one reports performance across engagement terciles, which describes how the model behaves for more and less active students but not for groups defined by protected characteristics [26].

The quality signals in Table 15 complement the audit by recording design problems that a reporting checklist does not capture. The most frequent was the absence of temporal or external validation (Q4; 16 of 19): most models were evaluated on students from the same cohort and institution as those used for training, which says little about performance on the next cohort, where an early-warning system would actually operate. Internal contradictions in data or results (Q7) were found in 16 reports, mostly discrepancies between sample sizes or between results stated in the abstract, text, and tables; they are often minor but reduce confidence in the headline figures. Resampling, feature selection, or normalisation applied to the whole dataset before splitting (Q2) appeared in 11 reports, and predictors containing the outcome or information close to it (Q1) in 9, typically grade components of the final grade used to predict that grade. Model or configuration selection with the data used for reporting (Q5) appeared in five reports, repeated observations of the same students on both sides of the split (Q3) in four, and few minority-class events (Q6) in three; for Q3 and Q6 the information needed was often missing (10 and 9 reports not determinable). These signals were coded from the reports with limited agreement before consensus (Section 2.3), and they are descriptive flags rather than a risk-of-bias judgement; their joint frequency nevertheless suggests that many headline results may be optimistic.

These gaps do not depend on the flagged reports or on the research group with two reports (Table 10): no report evaluates fairness across protected subgroups in any sensitivity analysis, and the absence of temporal or external validation remains the most frequent signal in all of them.

The audit therefore shows three reporting gaps and one design gap. Sample-size justification and ethical approval are rarely reported (items 10 and 17), the evaluation of performance across protected subgroups is absent (item 14), and most models are evaluated only within a single cohort (Q4). These findings concern what the reports document. They do not show that the models are biased, that ethical approval was missing, or that the models would fail on new cohorts; they show that the published evidence does not allow readers, institutions, or regulators to verify these properties. Future studies should align their reporting with TRIPOD+AI, validate on later cohorts, and report performance by subgroup, and journals in learning analytics and educational data mining could encourage this through their submission requirements.

**Table 15.** Signals of analytical quality by report. Q1: predictors containing the outcome or posterior information; Q2: resampling, feature selection, or normalisation before the split; Q3: repeated observations of the same student on both sides of the split; Q4: no temporal or external validation; Q5: tuning or selection with the reporting data; Q6: few minority-class events; Q7: internal contradictions. ND = not determinable from the report. Signals are descriptive flags, not a risk-of-bias judgement.

| Report | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 |
|---|---|---|---|---|---|---|---|
| Corrigan et al. [63] | ND | ND | ND | Yes | Yes | ND | Yes |
| Oreshin et al. [64] | No | ND | No | No | No | ND | Yes |
| Domladovac [47] | Yes | Yes | No | Yes | Yes | No | Yes |
| Fahd and Miah [48] | No | Yes | ND | Yes | Yes | ND | Yes |
| Matz et al. [30] | ND | Yes | No | No | No | No | Yes |
| Poellhuber et al. [24] | Yes | Yes | ND | Yes | Yes | ND | Yes |
| Kustitskaya et al. [27] | Yes | Yes | Yes | Yes | ND | No | ND |
| Pecuchova and Drlik [23] | Yes | Yes | Yes | Yes | ND | ND | Yes |
| Perkash et al. [62] | Yes | Yes | ND | Yes | No | Yes | Yes |
| Rico-Juan et al. [29] | No | Yes | ND | Yes | No | ND | Yes |
| Vaarma and Li [25] | No | No | No | No | Yes | ND | Yes |
| Al-Gahmi [61] | Yes | ND | Yes | Yes | ND | ND | Yes |
| Bettahi et al. [26] | Yes | No | ND | Yes | ND | No | Yes |
| Fahd and Miah [49] | No | Yes | ND | Yes | No | No | Yes |
| Madahana and Ekoru [28] | Yes | ND | ND | Yes | No | ND | Yes |
| Mutka et al. [67] | No | Yes | Yes | Yes | No | Yes | Yes |
| Nnamoko et al. [68] | No | No | ND | Yes | No | Yes | No |
| Pratiwi et al. [65] | ND | Yes | ND | Yes | No | No | Yes |
| Schnedlitz et al. [66] | Yes | ND | No | Yes | ND | No | No |
| Present (Yes) | 9 | 11 | 4 | 16 | 5 | 3 | 16 |
| Not determinable | 3 | 5 | 10 | 0 | 5 | 9 | 1 |

Seen through the normative lens of the 2030 Agenda, which motivates this review rather than serving as an analytical category, the fairness gap is also a sustainability risk. An early-warning system that performs worse for students with limited platform access would reproduce the disparities that SDG target 4.5 seeks to eliminate, and it could do so while reporting high aggregate accuracy. A recent regional fairness audit of dropout prediction on the Open University Learning Analytics Dataset illustrates the possibility [80]: logistic regression and random forest classifiers reached an overall AUC of 0.91–0.92, yet their true positive rate ranged from 0.55 to 0.80 across regions, an equal-opportunity gap of 0.25. Whether the models reviewed here behave in this way cannot be determined, because none of the 19 reports presents performance by protected subgroup. The risk is therefore plausible but unverified: the corpus provides no evidence that the models are unfair, and no evidence that they are fair.

Techniques to mitigate such disparities exist at every stage of the modeling pipeline. Before training, reweighing assigns instance weights that make the protected attribute and the outcome statistically independent [81]. During training, adversarial debiasing penalizes a predictor whose outputs allow an adversary to recover the protected attribute [82]. After training, group-specific decision thresholds can equalize true positive rates across subgroups, which satisfies the equal opportunity criterion [83]. No study in the reviewed corpus reported any of these techniques. All three require subgroup information, which is precisely what item 14 asks studies to collect and report.

### *4.5. Ethical Issues in Predicting Dropout from Digital Traces*

The ethical questions raised by this literature extend beyond the reporting of approval, and the corpus engages with them only briefly. Four issues deserve explicit attention,

because they bear on whether predictive systems of this kind should be deployed and on what terms.

The first concerns consent and the secondary use of traces. Interaction records are generated for teaching and administration, and their reuse for risk prediction is a secondary purpose that students rarely anticipate when they log in. Learning analytics ethics has long held that such reuse requires transparency about which data are collected, for what purpose, and with what consequences, and that students should be treated as agents with a stake in the process rather than as data subjects only [84,85]. Eight of the reviewed reports document ethical approval or consent, but they rarely explain what students were told about the use of their activity for risk prediction, and in one case consent was obtained by the original collectors of a public dataset for a different study [49]. Whether students in the remaining settings knew that their activity was being used to estimate their risk cannot be determined from the reports.

The second concerns surveillance and the proxies that traces represent. Logins, clicks, and time on task are behavioural proxies whose meaning depends on context: low activity may reflect disengagement, but also paid work, caring duties, limited connectivity, or a preference for studying offline. Treating these proxies as indicators of risk extends institutional monitoring into students' study habits and can disadvantage those whose legitimate patterns of study differ from the majority [74,78]. The reports that add external sources, such as public social-network profiles or neighbourhood statistics [64,68], widen this monitoring beyond the institution, and the principle of data minimisation, which asks whether each source is necessary for the purpose pursued, is not discussed in the corpus.

The third concerns the effects of the prediction itself. A risk label can help a student receive support, but it can also lower expectations, stigmatise, or channel students away from demanding courses, and the effects of a warning depend on how it is communicated and what follows it [86]. Once an institution knows that a student is at risk, it acquires a responsibility to act, and the obligation to provide support becomes part of the ethics of prediction [86]. Only two of the reviewed reports describe interventions triggered by their predictions, both without a randomised comparison [27,63], and none examined how students experienced being identified as at risk.

The fourth concerns governance and accountability. Decisions about which target to predict, which threshold to apply, and which errors to tolerate are value judgements: a threshold that maximises recall flags more students who would not have left, and one that maximises precision misses more who will. The reports present these choices as technical, and few involve students, advisers, or other stakeholders in making them. Regulation is beginning to formalise these responsibilities. Under the European Union's Artificial Intelligence Act [87], AI systems intended to evaluate learning outcomes, including when those outcomes are used to steer the learning process, are classified as high-risk (Annex III, point 3); dropout prediction is not named explicitly, but systems whose risk scores trigger interventions in a student's learning trajectory may fall within this category and would then be subject to requirements on data governance, including the examination of training data for possible biases (Article 10), and on human oversight. Checklists for trusted learning analytics translate similar principles into institutional practice [85].

Taken together, these issues suggest that the ethical adequacy of a dropout prediction system cannot be inferred from its accuracy or from the existence of an approval. It depends on whether students know and can contest how their data are used, whether the data collected are proportionate to the purpose, whether warnings lead to support rather than to labelling, and whether the people affected take part in defining what the system predicts and how it is used. These are questions that future primary studies could address directly,

for example by reporting consent procedures, justifying each data source, evaluating the consequences of warnings, and involving students and advisers in design.

### *4.6. Implications for Sustainable and Equitable Higher Education*

The gaps identified in the reporting audit are not only reporting deficiencies but also conditions for the sustainable use of predictive systems in universities. Three implications follow, concerning data governance, regulation, and institutional capacity.

Data governance is the first condition. Ethical approval or consent was documented in only 8 of the 19 reports, so for most models the basis for using student behavioural records is not stated. Institutions that intend to deploy early-warning systems need explicit rules on consent, retention, and access before student traces feed a risk model. Such rules protect students and also make the resulting systems auditable over time.

Regulation is the second condition. As more institutions adopt AI-based early-warning systems, the absence of reporting on differential model performance leaves regulators, accreditation bodies, and student organisations without the information needed to assess whether such systems comply with non-discrimination principles. The requirements of the European Union's Artificial Intelligence Act discussed in Section 4.5 call for precisely the documentation that the reviewed reports seldom provide. Researchers, institutional decision-makers, and regulators therefore converge on the same agenda: transparent, auditable, and equitable retention systems, not merely accurate ones.

Institutional capacity is the third condition. The configurations that dominate the corpus, LMS traces and familiar ML algorithms, run on data infrastructure that most universities already maintain, which lowers the cost of adoption and makes local development feasible for institutions with limited resources. The same accessibility places the responsibility for validation, fairness evaluation, documentation, and maintenance on the institution itself. A model that is validated on each new cohort, audited by subgroup, and documented against TRIPOD+AI can remain useful over successive academic years; a model deployed once and never re-evaluated may lose accuracy as student populations, platforms, and teaching practices change, and its errors may concentrate on the groups it already served least well.

### *4.7. Limitations of This Review*

The interpretation of the evidence reviewed in this study is constrained by limitations at three levels: corpus, primary study, and review.

At the corpus level, the comparability of reported results is limited by heterogeneity in how studies define and operationalize the target variable. Dropout, attrition, retention, inactivity, academic risk, delayed progression, and student success are not defined consistently across the corpus: some studies model permanent withdrawal, whereas others predict course failure, non-enrollment in the following semester, or risk of poor academic performance. Models that use similar terminology may therefore be estimating different educational outcomes, and predictive performance is not directly comparable across studies. A related limitation concerns the transferability of models trained on data from a single institution, program, course, or learning platform: although such data are useful for local early-warning systems, they may reflect specific institutional practices, platform configurations, grading policies, or student populations, and models trained on them may require retraining and validation before being applied to other educational contexts. The treatment of data sources is also uneven; LMS records, SIS data, demographic variables, academic history, and administrative records are often combined or compared without consistent reporting of their provenance, temporal structure, or level of granularity, which makes it difficult to determine which types of digital traces contribute most to early dropout prediction

and whether richer multi-source datasets provide systematic advantages over single-source LMS data.

Two additional corpus-level limitations relate to the conditions under which performance metrics should be read. Many studies report accuracy, precision, recall, F1 score, AUC, or similar indicators, but fewer examine whether model outputs are interpretable, actionable, or integrated into actual retention interventions; high predictive performance should therefore not be interpreted as evidence that a model can directly support pedagogical or administrative decision-making. Reporting of validation and optimization procedures also varies: some studies apply systematic hyperparameter optimization and cross-validation, whereas others use default configurations or provide limited information about model tuning, which affects the reliability of comparisons between algorithms, since differences in performance may reflect differences in tuning strategy rather than genuine superiority of one model family over another. Finally, although some studies include explainability techniques such as SHAP, LIME, or feature importance analysis, explainability is not consistently incorporated as a core component of model evaluation; in several cases, models identify students at risk without providing sufficient information about the factors driving the prediction, which limits the educational value of the system when instructors or student-support teams need to decide what intervention should follow a risk alert.

The geographic composition of the corpus is itself a corpus-level limitation. Although author affiliations span 21 countries, 13 of the 19 reports involve at least one of seven countries: Australia, Spain, the United States, Croatia, Russia, Canada, and Ireland. Several higher-education systems with substantial undergraduate enrollment, particularly China, India, Brazil, and the broader Sub-Saharan African region, are absent or represented by single contributions. Latin America, which the Introduction cites for some of the highest dropout rates reported, contributes no study led from the region. Its presence is limited to secondary affiliations in Mexico and the Dominican Republic within a single multinational study [62], and no study from Colombia, Ecuador, or Chile is part of the final corpus. This asymmetry reflects the indexing biases of Scopus and WoS toward English-language venues and high-income research ecosystems, and it constrains the external validity of the synthesis. The predictive patterns, model preferences, and reporting practices documented in this review may not generalize to higher-education systems operating under different institutional infrastructures, digital adoption levels, or student populations. Future reviews could mitigate this limitation by including regional databases such as SciELO, IndexLat, or the China National Knowledge Infrastructure, and by accepting non-English publications when robust translation or bilingual reviewing capacity is available.

At the primary-study level, privacy, institutional access, and data-governance constraints limit the size, diversity, and reproducibility of the datasets used. Student data are sensitive, and restrictions on sharing institutional records reduce opportunities for external benchmarking and replication, making it difficult to determine whether reported results would hold across different universities, countries, academic programs, or student cohorts. These constraints are legitimate, as they protect students and comply with applicable regulatory frameworks, but they amplify the corpus-level limitations on transferability and external validity.

At the review level, six limitations affect the interpretation of the synthesis. First, the review was registered retrospectively, after the searches, screening, and first round of extraction had been completed, and no separate protocol was prepared; the registration discloses this foreknowledge, but it cannot rule out decisions influenced by knowledge of the studies, and the six documented deviations, including the specification of the object of study during peer review, should be read in that light. Second, the search was limited to English-language records indexed in Scopus and WoS, and its `AND NOT` block on

online-learning terms may have removed campus-based or blended studies that mention an online component in their title, abstract, or keywords; the magnitude and direction of this potential coverage error could not be determined from the excluded records. Separately, because online-only education was excluded by design, the review does not establish whether the patterns reported here, including the near-absence of fairness evaluation, would persist in that setting; addressing that question would require a separately designed and documented search. Third, the eligibility rule for T3 targets accepted a link to dropout stated anywhere in a report. This rule is permissive: two reports were flagged for weak T3 framing, and these, together with the six reports flagged for other reasons, were removed in a sensitivity analysis that did not change the main findings. The finding on target substitution is conditional on this rule, because reports predicting failure without any link to dropout were excluded; it holds, however, when the flagged reports are removed (9 of 11). Fourth, the first round of extraction was coded by consensus without retaining pre-consensus records. The re-coding provides pre-consensus agreement for every variable, but agreement on the inclusion decision and on several quality signals was low before consensus; consensus by discussion against the full text and written rules mitigates this, but the coding of borderline cases remains a matter of judgement. Fifth, both reviewers used generative AI tools to pre-fill their coding sheets before verifying every code against the full text; different tools were used to limit common errors, but the agreement coefficients describe AI-assisted codings rather than unaided ones. Sixth, the synthesis consists of counts that treat each report as one unit and ignore sample size and precision, and the TRIPOD+AI scores measure the completeness of reporting rather than methodological quality; the percentages reported should therefore be read as descriptive patterns within this corpus, not as estimates for the field.

## 5. Conclusions

This systematic review examined how AI models use interaction-generated digital traces to predict student dropout, directly or through course failure or academic risk framed as an early warning of dropout, in campus-based and blended higher education. After a reassessment of the corpus with written rules and independent double coding, 19 reports on 18 datasets were included. They were classified by prediction target, horizon, trace source, model family, explainability, and validation, and their reporting was audited against items adapted from TRIPOD+AI and complemented by signals of analytical quality. Because the corpus is small, recent, concentrated in a few countries, and heterogeneous in targets and designs, the conclusions below describe this body of evidence rather than the field as a whole; the confidence ratings likewise apply to this corpus under the stated eligibility and coding rules.

The main finding concerns what is predicted. Although every report presents its model as dropout prediction, 15 of the 19 model course failure or academic risk, and only four model programme dropout or course withdrawal. The evidence therefore concerns academic risk more than departure, the value of these models as dropout warnings is not evaluated within the reports, and a warning based on course failure cannot, by construction, identify students who leave in good academic standing. The research questions add a consistent picture. LMS records are the most common source of interaction traces (RQ1: 6 of 19), often combined with institutional records. Half of the reports compare classical ML and DL models in parallel and the rest use classical ML only; none relies on DL alone, and the comparisons do not establish the superiority of any family (RQ2). Explainability is reported in 11 of the 19 reports, a majority that does not depend on one research group, but it serves mainly to describe global risk factors; explanations for the people who act on a warning are seldom produced and never evaluated (RQ3).

The audit of reporting completeness (RQ4) shows that the core technical workflow is well reported, whereas sample-size justification (2 of 19), ethical approval or consent (8 of 19), and evaluation of performance across protected subgroups (0 of 19) are rarely or never reported. These figures describe what the reports document, not what their authors did, and they do not show that the models are biased; they show that the published evidence does not allow anyone to verify whether the models work equally well for all students. The signals of analytical quality add a design concern: most models were evaluated only within the cohort used for training (16 of 19), and resampling or feature selection before the data split, predictors close to the outcome, and internal contradictions were frequent. Headline performance figures, some above 98%, may therefore be optimistic, because within-cohort evaluation and several of these practices are known to inflate performance estimates. They should not be interpreted as the performance an institution could necessarily expect in a future cohort. These findings were stable across the five sensitivity analyses.

For researchers, three priorities follow: defining targets that are, or are explicitly linked to, programme dropout, naming course-failure models as academic-risk prediction, and validating models on later cohorts; reporting according to TRIPOD+AI, including performance by protected subgroup, ethical approval, and consent; and designing explanations with the advisers and students who receive them, and evaluating what happens after a warning. For institutions, a model that performs well on aggregate metrics has not yet shown that it identifies risk equally well among all students, and its deployment requires local validation, subgroup audits, and clear rules on consent and data use. For the review community, the low agreement observed before consensus on some eligibility decisions and quality signals suggests that reliability should be reported for every coded variable.

Read through the normative lens of the 2030 Agenda, the review indicates that the technical capacity to predict academic risk from institutional data exists, whereas the evidence that this capacity advances equal access to tertiary education (SDG target 4.5) is not yet available. Producing that evidence would require future studies to report performance by subgroup, document consent and approval, treat missing traces as information rather than noise, and evaluate whether early warnings lead to support.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/su182111428/s1 and in the OSF project at https://osf.io/289vz/ (accessed on 4 November 2026): Table S1: PRISMA 2020 checklist [36]; Table S2: PRISMA 2020 for Abstracts checklist [36]; Table S3: Pre-consensus inter-rater agreement by coded variable [45,46]; Material S1: Worked examples of borderline eligibility decisions; Table S4: Correspondence between the original TRIPOD+AI Methods items and the adapted items [33].

**Author Contributions:** Conceptualization, M.A.; methodology, M.A., R.S., B.C. and R.O.; investigation, M.A. and A.J.; formal analysis, M.A.; data curation, M.A. and A.J.; writing—original draft preparation, M.A.; writing—review and editing, M.A., A.J., R.S., B.C. and R.O.; supervision, R.S., B.C. and R.O. All authors have read and agreed to the published version of the manuscript.

**Funding:** The work of B.C. was supported by the Agencia Nacional de Investigación y Desarrollo (ANID), Chile, through FONDECYT Regular under grant number 1262511.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data supporting the findings of this review are derived from the 19 included reports, which are publicly available and listed in the references. The eligibility rules, the codebook with its dated clarifications, the two pre-consensus coding spreadsheets with their SHA-256 fingerprints, the consensus spreadsheets with the justification of every resolved disagreement, the agreement coefficients, the adapted TRIPOD+AI checklist with its decision rules, the Rayyan

screening decisions, and the scripts used for comparison and consolidation are openly available in the Open Science Framework project associated with the registration at https://osf.io/289vz/ (accessed on 4 November 2026). The Boolean search strings and the PRISMA flow documentation are reproduced in Tables 3 and 5.

**Acknowledgments:** The authors thank the Universidad de Valparaíso and the Doctoral Program in Intelligent Industry of the Pontificia Universidad Católica de Valparaíso for their institutional support during the development of this work. During the preparation of this work the authors used Claude Opus 5.5 (Anthropic) and Z.AI GLM-5.3 (Zhipu AI; accessed 30 October 2026) to prepare evidence dossiers and pre-fill the data-extraction sheets, as described in Section 2.3, and Claude Sonnet 5.5 (Anthropic) to assist in drafting and editing the manuscript and the response to reviewers. MinerU 3.4.5 (OpenDataLab) was used to convert PDF files to text. The authors verified every extracted code against the original reports, reviewed and edited all content, and take full responsibility for the content of the publication.

**Conflicts of Interest:** The authors declare no conflicts of interest. No author of this review is an author of any of the included studies. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Appendix A. Countries of Author Affiliation by Study

Table A1 lists the countries of author affiliation for each included study, taken from the affiliations printed in each publication. A country is counted once per publication when at least one author lists an affiliation there, including secondary affiliations; the column totals reproduce Table 6 and Figure 3.

**Table A1.** Countries of author affiliation for each of the 19 included reports.

| Study | Countries of Affiliation |
|---|---|
| Al-Gahmi [61] | United States |
| Bettahi et al. [26] | Morocco |
| Corrigan et al. [63] | Ireland |
| Domladovac [47] | Croatia |
| Fahd and Miah [48] | Australia |
| Fahd and Miah [49] | Australia |
| Kustitskaya et al. [27] | Russia |
| Madahana and Ekoru [28] | South Africa |
| Matz et al. [30] | United States, Germany, Canada, Switzerland |
| Mutka et al. [67] | Croatia |
| Nnamoko et al. [68] | United Kingdom |
| Oreshin et al. [64] | Russia |
| Pecuchova and Drlik [23] | Slovakia |
| Perkash et al. [62] | Pakistan, Ireland, Spain, South Korea, United States (Puerto Rico), Mexico, Dominican Republic, Angola |
| Poellhuber et al. [24] | Canada |
| Pratiwi et al. [65] | Indonesia |
| Rico-Juan et al. [29] | Spain |
| Schnedlitz et al. [66] | Austria, Australia |
| Vaarma and Li [25] | Finland |

## References

1. Tinto, V. *Leaving College: Rethinking the Causes and Cures of Student Attrition*; University of Chicago Press: Chicago, IL, USA, 1987.
2. Panwar, S.S.; Khandale, K.; Chourasia, A.; Patidar, A.; Mandloi, G. A review on understanding the patterns of student dropout. In *Proceedings of the World Congress on Smart Computing (WCSC 2024)*; Studies in Smart Technologies; Shukla, P.K., Sharma, H., Mallipeddi, R., Eds.; Springer: Singapore, 2025; pp. 235–250. [CrossRef]
3. Brunner, J.J.; Alarcón, M. Higher education governance as a social contract: Challenges for Latin America and the Caribbean. *Prospects* **2024**, *54*, 411–417. [CrossRef]

4. Arévalo-Cordovilla, F.E.; Peña, M. AI-driven predictive models and chatbots for early intervention and student success in higher education: A systematic review. *Int. J. Eng. Educ.* **2025**, *41*, 1473–1488.
5. Mduma, N.; Kalegele, K.; Machuve, D. A survey of machine learning approaches and techniques for student dropout prediction. *Electron. J. Inf. Syst. Dev. Ctries.* **2019**, *85*, e12089. [CrossRef]
6. Parra-Sánchez, J.S.; Pardo, I.D.T.; de Merino, C.Y.M. Explanatory factors of university dropout explored through artificial intelligence. *Rev. Electron. Investig. Educ.* **2023**, *25*, 1–17. [CrossRef]
7. OECD. *Benchmarking Higher Education System Performance*; Higher Education, OECD Publishing: Paris, France, 2019. [CrossRef]
8. Ferreira, F.H.G.; Avitabile, C.; Botero, J.; Haimovich, F.; Urzúa, S. *Momento Decisivo: La Educación Superior en América Latina y el Caribe*; Banco Mundial: Washington, DC, USA, 2017.
9. Bean, J.P. Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Res. High. Educ.* **1980**, *12*, 155–187. [CrossRef]
10. United Nations General Assembly. *Transforming Our World: The 2030 Agenda for Sustainable Development*; Resolution A/RES/70/1; United Nations General Assembly: New York, NY, USA, 2015.
11. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *WIREs Data Min. Knowl. Discov.* **2020**, *10*, e1355. [CrossRef]
12. Cardona, T.; Cudney, E.A.; Hoerl, R.; Snyder, J. Data mining and machine learning retention models in higher education. *J. Coll. Stud. Retent. Res. Theory Pract.* **2023**, *25*, 51–75. [CrossRef]
13. Pelima, L.R.; Sukmana, Y.; Rosmansyah, Y. Predicting university student graduation using academic performance and machine learning: A systematic literature review. *IEEE Access* **2024**, *12*, 23451–23465. [CrossRef]
14. Okewu, E.; Adewole, P.; Misra, S.; Maskeliunas, R.; Damasevicius, R. Artificial neural networks for educational data mining in higher education: A systematic literature review. *Appl. Artif. Intell.* **2021**, *35*, 983–1021. [CrossRef]
15. Attiya, W.M.; Shams, M.B. Predicting student retention in higher education using data mining techniques: A literature review. In *Proceedings of the International Conference on Cyber Management and Engineering (CyMaEn 2023)*; IEEE: Piscataway, NJ, USA, 2023; pp. 171–177. [CrossRef]
16. Luo, J.; Zheng, C.; Yin, J.; Teo, H.H. Design and assessment of AI-based learning tools in higher education: A systematic review. *Int. J. Educ. Technol. High. Educ.* **2025**, *22*, 42. [CrossRef]
17. Quimiz-Moreira, M.; Delgadillo, R.; Parraga-Alava, J.; Maculan, N.; Mauricio, D. Factors, Prediction, Explainability, and Simulating University Dropout Through Machine Learning: A Systematic Review, 2012–2024. *Computation* **2025**, *13*, 198. [CrossRef]
18. Shi, H.; Zhang, N.; Caskurlu, S.; Na, H. Applications of Machine Learning for At-Risk Student Prediction in Online Education: A 10-Year Systematic Review of Literature. *J. Comput. Assist. Learn.* **2025**, *41*, e70058. [CrossRef]
19. Liu, X.; Shariff, S.A.; Wong, D.H.T.; Xie, Z. A Systematic Literature Review of Practical and Explainable AI (XAI) Student Performance Prediction System in Higher Education Institutions. *Educ. Technol. Res. Dev.* **2026**, 1–44, *advance online publication*. [CrossRef]
20. Rodrigues, H.S.; Santiago, E.D.S.; Wanderley, G.M.C.X.; Moraes, L.O.; Mello, C.E.; Alvares, R.V.; Dos Santos, R.P. Artificial intelligence algorithms to predict college students' dropout: A systematic mapping study. In Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024), Rome, Italy, 24–26 February 2024; Volume 3, pp. 344–351. [CrossRef]
21. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
22. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
23. Pecuchova, J.; Drlik, M. Enhancing the early student dropout prediction model through clustering analysis of students' digital traces. *IEEE Access* **2024**, *12*, 159336–159367. [CrossRef]
24. Poellhuber, L.V.; Poellhuber, B.; Desmarais, M.; Léger, C.; Manh-Chien Vu, M. Cluster-based performance of student dropout prediction as a solution for large scale models in a Moodle LMS. In *Proceedings of the LAK '23: Proceedings of the 13th International Learning Analytics and Knowledge Conference*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 592–598. [CrossRef]
25. Vaarma, M.; Li, H. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technol. Soc.* **2024**, *76*, 102474. [CrossRef]
26. Bettahi, A.; Harroud, H.; Belouadha, F.Z. Early student risk detection using CR-NODE: A completion-focused temporal approach with explainable AI. *Algorithms* **2025**, *18*, 781. [CrossRef]
27. Kustitskaya, T.A.; Esin, R.V.; Vainshtein, Y.V.; Noskov, M.V. Hybrid approach to predicting learning success based on digital educational history for timely identification of at-risk students. *Educ. Sci.* **2024**, *14*, 657. [CrossRef]

28. Madahana, M.C.I.; Ekoru, J.E.D. Predicting at-risk students in a blended learning environment via machine learning and deep learning approach with practical implications. In *Proceedings of the 2025 IEEE World AI IoT Congress (AIIoT)*; IEEE: Piscataway, NJ, USA, 2025; pp. 398–403. [CrossRef]
29. Rico-Juan, J.R.; Cachero, C.; Macià, H. Study regarding the influence of a student's personality and an LMS usage profile on learning performance using machine learning techniques. *Appl. Intell.* **2024**, *54*, 6175–6197. [CrossRef]
30. Matz, S.C.; Bukow, C.S.; Peters, H.; Deacons, C.; Stachl, C. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Sci. Rep.* **2023**, *13*, 5705. Correction in *Sci. Rep.* **2023**, *13*, 10069. https://doi.org/10.1038/s41598-023-36579-2. [CrossRef]
31. Greller, W.; Drachsler, H. Translating learning into numbers: A generic framework for learning analytics. *Educ. Technol. Soc.* **2012**, *15*, 42–57.
32. Alyahyan, E.; Düştegör, D. Predicting academic success in higher education: Literature review and best practices. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 3. [CrossRef]
33. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. Correction in *BMJ* **2024**, *385*, q902. https://doi.org/10.1136/bmj.q902. [CrossRef]
34. Andaur Navarro, C.L.; Damen, J.A.A.; Takada, T.; Nijman, S.W.J.; Dhiman, P.; Ma, J.; Collins, G.S.; Bajpai, R.; Riley, R.D.; Moons, K.G.M.; et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: A systematic review. *BMC Med. Res. Methodol.* **2022**, *22*, 12. [CrossRef]
35. Kapoor, S.; Cantrell, E.M.; Peng, K.; Pham, T.H.; Bail, C.A.; Gundersen, O.E.; Hofman, J.M.; Hullman, J.; Lones, M.A.; Malik, M.M.; et al. REFORMS: Consensus-based recommendations for machine-learning-based science. *Sci. Adv.* **2024**, *10*, eadk3452. [CrossRef]
36. Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*, n160. [CrossRef]
37. Aravena, M.; Olivares, R.; Jeria, A.; Crawford, B.; Soto, R. Artificial intelligence for predicting student dropout from digital traces in higher education: A systematic literature review of methods, explainability, and fairness gaps [Registration]. *OSF Regist.* **2026**. [CrossRef]
38. Pawson, R.; Greenhalgh, T.; Harvey, G.; Walshe, K. Realist review—A new method of systematic review designed for complex policy interventions. *J. Health Serv. Res. Policy* **2005**, *10*, 21–34. [CrossRef]
39. Dixon-Woods, M.; Agarwal, S.; Jones, D.; Young, B.; Sutton, A. Synthesising qualitative and quantitative evidence: A review of possible methods. *J. Health Serv. Res. Policy* **2005**, *10*, 45–53. [CrossRef]
40. Campbell, M.; McKenzie, J.E.; Sowden, A.; Katikireddi, S.V.; Brennan, S.E.; Ellis, S.; Hartmann-Boyce, J.; Ryan, R.; Shepperd, S.; Thomas, J.; et al. Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ* **2020**, *368*, l6890. [CrossRef]
41. Wolff, R.F.; Moons, K.G.; Riley, R.D.; Whiting, P.F.; Westwood, M.; Collins, G.S.; Reitsma, J.B.; Kleijnen, J.; Mallett, S. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **2019**, *170*, 51–58. [CrossRef]
42. Moons, K.G.M.; Damen, J.A.A.; Kaul, T.; Hooft, L.; Andaur Navarro, C.; Dhiman, P.; Beam, A.L.; Van Calster, B.; Celi, L.A.; Denaxas, S.; et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **2025**, *388*, e082505. [CrossRef]
43. Whiting, P.F.; Rutjes, A.W.; Westwood, M.E.; Mallett, S.; Deeks, J.J.; Reitsma, J.B.; Leeflang, M.M.; Sterne, J.A.; Bossuyt, P.M. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann. Intern. Med.* **2011**, *155*, 529–536. [CrossRef]
44. Hong, Q.N.; Fàbregues, S.; Bartlett, G.; Boardman, F.; Cargo, M.; Dagenais, P.; Gagnon, M.P.; Griffiths, F.; Nicolau, B.; O'Cathain, A.; et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Educ. Inf.* **2018**, *34*, 285–291. [CrossRef]
45. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
46. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 2nd ed.; Sage: Thousand Oaks, CA, USA, 2004.
47. Domladovac, M. Comparison of neural network with gradient boosted trees, random forest, logistic regression and SVM in predicting student achievement. In *Proceedings of the 44th International Convention on Information, Communication and Electronic Technology (MIPRO 2021)*; IEEE: Piscataway, NJ, USA, 2021; pp. 211–216. [CrossRef]
48. Fahd, K.; Miah, S.J. Designing and evaluating a big data analytics approach for predicting students' success factors. *J. Big Data* **2023**, *10*, 159. [CrossRef]
49. Fahd, K.; Miah, S.J. Enhanced predictive performance: A comparative analysis of ML and DL models using on augmented LMS interaction data. *Contemp. Educ. Technol.* **2025**, *17*, ep606. [CrossRef]

50. Guyatt, G.H.; Oxman, A.D.; Vist, G.E.; Kunz, R.; Falck-Ytter, Y.; Alonso-Coello, P.; Schünemann, H.J. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* **2008**, *336*, 924–926. [CrossRef]
51. Ouzzani, M.; Hammady, H.; Fedorowicz, Z.; Elmagarmid, A. Rayyan—A web and mobile app for systematic reviews. *Syst. Rev.* **2016**, *5*, 210. [CrossRef]
52. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
53. Singh, A.K.; Karthikeyan, S. Heuristic technique to find optimal learning rate of LSTM for predicting student dropout rate. In *Proceedings of the 25th International Conference on Artificial Intelligence in Education (AIED 2024)*; Communications in Computer and Information Science; Springer: Cham, Switzerland, 2024; Volume 2151, pp. 47–54. [CrossRef]
54. Sharma, P.; Mehan, V. Risk of dropouts in education: Discussions and challenges. *AIP Conf. Proc.* **2023**, *2878*, 020009. [CrossRef]
55. Luza, J.C.J.; Rodriguez, C. Predictive attributes in machine learning for university academic performance: A feature engineering approach. In Proceedings of the 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN 2024), Indore, India, 22–23 December 2024; pp. 443–456. [CrossRef]
56. Albahli, S. Advancing sustainable educational practices through AI-driven prediction of academic outcomes. *Sustainability* **2025**, *17*, 1087. [CrossRef]
57. Chen, B.; Modry-Caron, I. Engineering student early dropout prediction in regional universities using multimodal AI. In Proceedings of the 2025 ASEE Annual Conference & Exposition, Montreal, QC, Canada, 22–25 June 2025. [CrossRef]
58. De Lima, L.M.; Krohling, R.A. Discovering an aid policy to minimize student evasion using offline reinforcement learning. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Virtually, 18–22 July 2021. [CrossRef]
59. Fahd, K.; Miah, S.J.; Ahmed, K. Predicting student performance in a blended learning environment using learning management system interaction data. *Appl. Comput. Inform.* **2025**, *21*, 220–231. [CrossRef]
60. Porras, J.M.; Lara, J.A.; Romero, C.; Ventura, S. A case-study comparison of machine learning approaches for predicting student's dropout from multiple online educational entities. *Algorithms* **2023**, *16*, 554. [CrossRef]
61. Al-Gahmi, A. Identifying struggling students using LMS data. In *Proceedings of the 2025 ASEE Annual Conference and Exposition*; American Society for Engineering Education: Washington, DC, USA, 2025. [CrossRef]
62. Perkash, A.; Shaheen, Q.; Saleem, R.; Rustam, F.; Villar, M.G.; Alvarado, E.S.; Torre Diez, I.D.L.; Ashraf, I. Feature optimization and machine learning for predicting students' academic performance in higher education institutions. *Educ. Inf. Technol.* **2024**, *29*, 21169–21193. [CrossRef]
63. Corrigan, O.; Glynn, M.; McKenna, A.; Smeaton, A.; Smyth, S. Student data: Data is knowledge—Putting the knowledge back in the students' hands. In *Proceedings of the 14th European Conference on e-Learning (ECEL 2015)*; Jefferies, A., Cubric, M., Eds.; Academic Conferences and Publishing International Limited: Hatfield, UK, 2015; pp. 165–172.
64. Oreshin, S.; Filchenkov, A.; Petrusha, P.; Krasheninnikov, E.; Panfilov, A.; Glukhov, I.; Kaliberda, Y.; Masalskiy, D.; Serdyukov, A.; Kazakovtsev, V.; et al. Implementing a machine learning approach to predicting students' academic outcomes. In *Proceedings of the CCRIS 2020: Proceedings of the 2020 International Conference on Control, Robotics and Intelligent System*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 78–83. [CrossRef]
65. Pratiwi, H.; Sa'ad, M.I.; Junirianto, E. Data-driven educational management in higher education: Predictive modeling of student learning outcomes using artificial intelligence. In *Proceedings of the 2025 International Conference on Informatics and Computing (ICIC)*; IEEE: Piscataway, NJ, USA, 2025. [CrossRef]
66. Schnedlitz, F.; Kerschbaumer, D.; Steinmaurer, A.; Lowe, D.; Gütl, C. To complete or not to complete: Prediction and classification of dropouts in a CS1 course. In *Proceedings of the 2025 IEEE Global Engineering Education Conference (EDUCON)*; IEEE: Piscataway, NJ, USA, 2025. [CrossRef]
67. Mutka, A.; Živković Mutka, F.; Žagar, M.; Tolić, D. Enhancing student retention in introductory programming courses: leveraging advanced learning validation tools and educational data mining. *IEEE Access* **2025**, *13*, 153614–153626. [CrossRef]
68. Nnamoko, N.; Barrowclough, J.; Onikoyi, B.; Liptrott, M. Predicting student academic outcome using online behavioural statistics and neighbourhood influences: A machine learning approach. *SN Comput. Sci.* **2025**, *6*, 661. [CrossRef]
69. Cabrera, A.F.; Nora, A.; Castañeda, M.B. College persistence: Structural equations modeling test of an integrated model of student retention. *J. High. Educ.* **1993**, *64*, 123–139. [CrossRef]
70. Braxton, J.M. (Ed.) *Reworking the Student Departure Puzzle*; Vanderbilt University Press: Nashville, TN, USA, 2000.
71. Tinto, V. Dropout from higher education: A theoretical synthesis of recent research. *Rev. Educ. Res.* **1975**, *45*, 89–125. [CrossRef]
72. Fredricks, J.A.; Blumenfeld, P.C.; Paris, A.H. School engagement: Potential of the concept, state of the evidence. *Rev. Educ. Res.* **2004**, *74*, 59–109. [CrossRef]
73. Seidman, A. (Ed.) *College Student Retention: Formula for Student Success*; Praeger: Westport, CT, USA, 2005.
74. Baker, R.S.; Hawn, A. Algorithmic bias in education. *Int. J. Artif. Intell. Educ.* **2022**, *32*, 1052–1092. [CrossRef]
75. Kizilcec, R.F.; Lee, H. Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*; Holmes, W., Porayska-Pomsta, K., Eds.; Routledge: New York, NY, USA, 2022; pp. 174–202. [CrossRef]

76. Chen, I.Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; Ghassemi, M. Ethical machine learning in healthcare. *Annu. Rev. Biomed. Data Sci.* **2021**, *4*, 123–144. [CrossRef]
77. Ibrahim, H.; Liu, X.; Zariffa, N.; Morris, A.D.; Denniston, A.K. Health data poverty: An assailable barrier to equitable digital health care. *Lancet Digit. Health* **2021**, *3*, e260–e265. [CrossRef]
78. Williamson, B.; Eynon, R. Historical threads, missing links, and future directions in AI in education. *Learn. Media Technol.* **2020**, *45*, 223–235. [CrossRef]
79. Camaradou, J.C.L.; Hogg, H.D.J. Commentary: Patient perspectives on artificial intelligence; what have we learned and how should we move forward? *Adv. Ther.* **2023**, *40*, 2563–2572. [CrossRef]
80. Elsayed, A.; Wardat, Y.; Kamalov, F.; Sulieman, H. Toward Sustainable and Equitable AI in Education Through a Regional Fairness Audit of Dropout Prediction Using the OULAD Dataset. *Sustainability* **2026**, *18*, 9440. [CrossRef]
81. Kamiran, F.; Calders, T. Data Preprocessing Techniques for Classification Without Discrimination. *Knowl. Inf. Syst.* **2012**, *33*, 1–33. [CrossRef]
82. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 335–340. [CrossRef]
83. Hardt, M.; Price, E.; Srebro, N. Equality of Opportunity in Supervised Learning. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016; pp. 3323–3331.
84. Slade, S.; Prinsloo, P. Learning analytics: Ethical issues and dilemmas. *Am. Behav. Sci.* **2013**, *57*, 1509–1528. [CrossRef]
85. Drachsler, H.; Greller, W. Privacy and analytics: It's a DELICATE issue. A checklist for trusted learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16), Edinburgh, UK, 2016*; ACM: New York, NY, USA, 2016; pp. 89–98. [CrossRef]
86. Prinsloo, P.; Slade, S. Ethics and learning analytics: Charting the (un)charted. In *Handbook of Learning Analytics*; Lang, C., Siemens, G., Wise, A., Gašević, D., Eds.; Society for Learning Analytics Research (SoLAR): Beaumont, AB, Canada, 2017; pp. 49–57. [CrossRef]
87. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, *L*, 2024/1689. Available online: http://data.europa.eu/eli/reg/2024/1689/oj (accessed on 12 July 2024).


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.