IMBALANCED MULTI-CLASS PREDICTION OF STUDENT DROP-OUT AND GRADUATION: A SYSTEMATIC LITERATURE REVIEW

Ridwan Setiawan[1,2]✉
Edi Noersasongko[2]
Abdul Syukur[2]
Fikri Budiman[2]
Dede Kurniadi[1]

[1]Institut Teknologi Garut, Indonesia

[2]Universitas Dian Nuswantoro, Indonesia

✉ ridwan.setiawan@itg.ac.id

ABSTRACT

Student study status prediction, including drop-out and graduation, is a widely studied topic in higher education. Yet, evidence across studies remains difficult to compare due to differences in targets, imbalance treatment, metrics, and validation strategies. This systematic literature review synthesizes 70 peer reviewed articles published between 2017 and 2025 that apply machine learning or deep learning to predict study outcomes under class imbalance. Results reveal a strong dominance of binary targets, while multi class experiments are relatively rare, though they better reflect institutional categories and expose larger performance gaps across classes. Reported imbalance handling includes data level resampling, algorithm level class weighting, and ensemble or hybrid designs, but many studies lack sufficient procedural detail. Evaluation practices vary considerably; studies reporting per-class measures and imbalance-aware metrics, such as macro F1 and balanced accuracy, provide more decision-relevant evidence than those relying mainly on accuracy. Validation strategies range from hold out and stratified cross validation to nested validation, temporal splits, and external testing, shaping the credibility of reported performance for deployment. We propose an integrative taxonomy linking target formulation, imbalance degree, handling strategy, and evaluation design to enhance intervention efficiency through capacity aware prioritization, while strengthening responsibility through transparent reporting, defensible validation, and explicit attention to minority class performance.

KEYWORDS

Higher education, imbalanced classification, learning analytics, machine learning, responsibility, student dropout

HOW TO CITE

Setiawan R., Noersasongko E., Syukur A., Budiman F., Kurniadi D. (2026) 'Imbalanced Multi-class Prediction of Student Drop-out and Graduation: A Systematic Literature Review', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 72–90. http://dx.doi.org/10.7160/eriesj.2026.190106

Article history
Received
January 12, 2026
Received in revised form
February 27, 2026
Accepted
March 12, 2026
Available on-line
March 31, 2026

Highlights

- *Synthesizes 70 open-access Scopus journal studies from 2017 to 2025 on predicting student drop-out and graduation using machine learning and deep learning.*
- *Finds strong dominance of binary targets, while multi-class targets remain rare yet better reflect institutional status categories and reveal larger per-class performance gaps.*
- *Shows imbalance handling is often under-reported; when reported, options include resampling, class weighting, cost-sensitive learning, and ensemble or hybrid designs, with implications for transparent and accountable study design.*
- *Highlights that accuracy alone is insufficient under imbalance; per-class metrics and imbalance-aware summaries improve decision relevance for efficient interventions and improve responsibility through clearer minority class reporting and more defensible evidence.*

INTRODUCTION

Student dropout and delayed graduation remain persistent challenges in higher education across different countries and institutional contexts. Their consequences extend beyond individual students and entail economic, social, and managerial implications for institutions and policymakers, particularly regarding service sustainability, academic quality, and accountability in higher education management (Csalódi and Abonyi, 2021; Véliz Palomino and Ortega, 2023; Quimiz-Moreira et al., 2025). Early identification of students at risk is therefore a strategic necessity because it enables interventions to be delivered in a timelier, more targeted, and more evidence-based manner (Setiawan et al., 2025). In this review, this practical need is closely linked to *efficiency*, understood as the extent to which predictive evidence can support the prioritization

of interventions under constrained institutional resources. At the same time, *responsibility* refers to the defensibility, transparency, and fairness of the evidence used in educational decision support and scientific reporting. Accordingly, this review is positioned not merely as an inventory of modelling approaches, but as an examination of the study design decisions that shape whether research findings are useful for intervention prioritization and defensible for educational and scientific use.

In line with the growing availability of academic records and learning activity data, learning analytics and educational data mining are increasingly used to understand risk patterns and support data-driven decision-making. A systematic review by de Oliveira et al. (2021) showed that dropout risk is influenced by a combination of academic and nonacademic factors, while Andrade-Girón et al. (2023) confirmed the relevance of machine learning and deep learning for building pattern-based predictions and early warning systems. However, interpretability and variation in evaluation practices remain important methodological concerns.

Several review studies have mapped predictive approaches in the education domain, including the use of artificial neural networks in educational data mining by Okewu et al. (2021), data mining practices for academic performance prediction by Daza et al. (2022), and a review of graduation prediction that highlighted limitations in algorithm coverage, database selection, and the transparency of data collection procedures in earlier systematic literature reviews by Pelima et al. (2024). In the context of dropout prediction, Salinas-Chipana et al. (2024) reported a PRISMA-based review that showed the dominance of random forest models and confirmed the importance of academic, demographic, economic, and health-related attributes as predictors. Although the review literature has enriched the understanding of predictive methods and feature sets, class imbalance and evaluation consistency, especially in multiclass scenarios, have not yet been addressed as central issues in an integrated manner.

The distribution of labels in student retention data is commonly imbalanced. Cases of dropout are often substantially fewer than cases of persistence or graduation. As a result, a model may appear strong on aggregate metrics while remaining weak at detecting minority groups, which are often the primary targets of intervention. Budiman et al. (2022) and Villar and de Andrade (2024) emphasized that class imbalance is a recurring issue in educational data mining and requires both appropriate imbalance handling strategies and the use of more representative evaluation metrics. In this respect, Martins et al. (2021) noted that balance-sensitive metrics based on precision and recall, including the F1 score, are more informative than accuracy alone, particularly when the objective is to identify at-risk students in a fairer and more actionable way.

Beyond the challenge of imbalance, the complexity of academic status often requires richer target formulations than binary labels alone (Kurniadi et al., 2021). Categories such as still enrolled, graduated, delayed graduation, and dropout reflect different academic pathways and may call for different forms of intervention. Within this context, *efficiency* concerns whether predictive models can support the targeted allocation of mentoring and intervention resources so that unnecessary alarms do not consume institutional capacity and missed risk cases are managed in line with policy priorities. In parallel, *responsibility* concerns the adequacy of minority-class evaluation, methodological transparency, and the avoidance of biased conclusions arising from inappropriate validation designs or performance metrics that obscure failures to detect students at risk. Thus, evaluation quality is not merely a technical matter, but one with direct implications for the effectiveness of retention policies and the fairness of institutional services.

This study addresses this gap by providing a systematic literature review of the application of machine learning and deep learning techniques to predict student academic status in higher education, with particular attention to dropout and graduation outcomes and to the methodological challenges of multiclass classification under class imbalance. It makes three specific contributions. First, it offers a systematic framework for examining how prior studies formulate prediction targets as binary or multiclass problems and for clarifying the methodological consequences of these choices for model evaluation in higher education contexts (de Oliveira et al., 2021; Pelima, Sukmana, and Rosmansyah, 2024; Salinas-Chipana et al., 2024). Second, it comparatively synthesizes the class imbalance mitigation strategies reported in the literature, at both the data and algorithmic levels, and relates them to evaluation and validation practices that determine whether performance on minority classes can meaningfully support intervention prioritization and responsible educational decision support (Andrade-Girón et al., 2023; Martins et al., 2023; Villar and de Andrade, 2024). Third, as an integrative outcome, it proposes a taxonomy of study design decisions that links target formulation, imbalance severity, mitigation strategies, and the selection of performance metrics and validation schemes. This taxonomy is intended both as an audit framework for reviewing prior studies and as a guide for designing future research that is more transparent, accountable, and methodologically aligned with the complementary aims of *efficiency* and *responsibility* in educational science (Page et al., 2021; Rethlefsen et al., 2021).

The purpose of this manuscript is to systematically synthesize empirical evidence on predicting student dropout and graduation in higher education, with particular emphasis on the relationships among binary and multiclass target formulations, class imbalance handling strategies, and evaluation and validation practices. This synthesis leads to a taxonomy of study design decisions intended to support intervention efficiency and the responsible use of predictive models. In line with this objective, the review addresses six research questions: RQ1, what machine learning and deep learning algorithms have been used to predict student dropout and graduation; RQ2, how have targets been formulated as binary or multiclass classifications, and what implications do these choices have for modelling and reporting; RQ3, what class imbalance handling strategies have been reported at the data, algorithmic, and decision levels; RQ4, to what extent have ensemble and hybrid methods been used, and what application patterns emerge in this corpus; RQ5, which evaluation metrics and validation methods have been used, and how suitable are

they for imbalanced data; and RQ6, what methodological gaps and challenges remain, and what recommendations can strengthen future research. The systematic review procedure and the reporting of the study selection flow follow the principles of Systematic Literature Review (SLR) and PRISMA to ensure procedural traceability and replicability (Kitchenham, 2004; Page et al., 2021).

This manuscript is organized as follows. The Materials and Methods section presents the systematic review procedure, search strategy, selection criteria, and data extraction and synthesis processes. The Results section presents the main findings in response to the research questions. The Discussion section relates these findings to the broader literature, explains their implications for intervention efficiency and responsible model use, and discusses the limitations of the review. Finally, the Conclusion section summarizes the principal contributions and outlines directions for future research.

MATERIALS AND METHODS

This study employs a systematic literature review to synthesize empirical evidence on predicting student dropout and graduation in higher education, with an emphasis on multi-class formulations under class imbalance conditions (Kitchenham, 2004; Page et al., 2021). The scope of the study and research questions were determined using the PICOC framework to maintain consistency in selection and extraction decisions and to link the population, intervention, comparison, outcome, and context elements to the research questions. The operationalization of Population, Intervention, Comparison, Outcomes, and Context (PICOC) and its mapping to the research questions are presented in Table 1.

PICOC Element	Operationalization	RQ	Motivation
Population	Student academic data in higher education institutions includes imbalanced binary and multi-class scenarios.	RQ1; RQ2; RQ5	Mapping data characteristics and variations in the number of classes to identify multi-class research gaps.
Intervention	Machine learning and deep learning algorithms, including single, ensemble, and hybrid configurations.	RQ1; RQ3; RQ4	Inventorying modelling approaches and consistency in handling imbalanced data.
Comparison	Baseline without handling imbalance; resampling technique variations; cost sensitivity; ensemble; hybrid; and target formulation variations.	RQ1; RQ2; RQ3; RQ4	Mapping variations in experimental design without making it a single claim of superiority.
Outcome	Aggregate metrics and per-class metrics, the confusion matrix, and computational efficiency are reported.	RQ5	Assessing the representativeness of metrics for minority classes and potential evaluation bias.
Context	Formal higher education, public or institutional datasets, focusing on dropout rates, academic performance, graduation, and early warning systems.	RQ1; RQ2; RQ3; RQ4; RQ5; RQ6	Maintaining domain relevance and facilitating cross-institutional application mapping.

Table 1: PICOC framework, motivation, dan RQ

The process of identifying, screening, and reporting the study selection flow follows PRISMA. A summary of the selection process and reasons for exclusion are presented in Figure 1 (Page et al., 2021; Rethlefsen et al., 2021).

Data source and search strategy

A literature search was conducted using Scopus as the sole database to ensure query repeatability and selection consistency across a single, cross-publisher, cross-disciplinary index. Scopus was selected because it provides standardized bibliographic metadata and DOI linkage across peer-reviewed journals, which supports protocol auditability and replicable retrieval under a fixed query string. This choice improves procedural consistency, yet it is also a limitation because reliance on a single index may introduce coverage bias and may under represent relevant journal outlets that are more visible through other curated indexes or discipline specific libraries (Mongeon and Paul-Hus, 2016; Baas et al., 2020) To enhance independent verifiability, the corpus was limited to open access journal articles with DOIs. This restriction enables readers to access full texts and verify data extraction, but it may exclude relevant evidence available outside open access or outside Scopus indexing. Future extensions of this review may integrate additional sources, such as Web of Science and discipline-focused libraries, to quantify overlap and assess whether distributional findings differ across indexes (Helbach et al., 2022).

The inquiries aim to encompass three dimensions: the modelling methodology, the higher education context, and outcomes on dropout or graduation rates. The inquiry was executed utilizing the advanced search function within the title, abstract, and keywords boxes, imposing a temporal constraint from 2015 to 2025, specifying journal source type, English language, and final publication status. The search was conducted on August 31, 2025, and updated on December 1, 2025. The Scopus query string used is listed in this section to ensure the reproducibility of the procedure.

> TITLE-ABS-KEY ((“machine learning” OR “deep learning” OR “neural network*” OR “artificial intelligence” OR “predictive analytics” OR “educational data mining”) AND (“student dropout” OR “dropout prediction” OR “student retention” OR “graduation prediction” OR “student graduation” OR “academic performance” OR “student success” OR “student attrition”) AND (“university” OR “college” OR “higher education”)) AND (LIMIT-TO (SRCTYPE,

"j")) AND (LIMIT-TO (OA, "all")) AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (LANGUAGE, "english"))

Inclusion and exclusion criteria

Inclusion and exclusion criteria are established before the selection process begins to prevent ad hoc changes. In summary, studies were included if they were conducted in a higher education context, used student academic data, had a prediction target related to institutional study status, employed machine learning or deep learning as the primary approach, and reported evaluations with clear metrics and procedures. Studies were excluded if they focused on a non-institutional context, such as MOOCs, courses, training, or semester completion; were not aligned with the definition of study status outcome; did not use ML or DL; did not report evaluations; or were from discontinued journals. The criteria are presented in Table 2.

Type	Criteria
Inclusion	(1) the context of higher education and the use of student academic data. (2) the prediction or classification target relates to institutional study status, including dropout or study cessation, retention, graduation, timeliness, or other academic risk categories explicitly mapped to student study status. (3) Using ML or DL as the primary approach. (4) Providing sufficient information on the target formulation, features, or dataset for synthesis. (5) Reporting model performance with clear evaluation metrics or procedures.
Exclusion	(1) Focus on levels apart from student study status. (2) Focus on MOOCs, courses, training, or semester completion. (3) the study status outcomes are not aligned with the definition provided by the institution. (4) Does not use ML or DL. (5) Does not report evaluation. (6) the study was published in a journal that has since been discontinued.

Table 2: Inclusion and exclusion criteria

Study selection process and records management

The selection was conducted in stages: title and abstract screening, followed by full-text assessment. Out of 845 records screened at the title and abstract stage, 268 studies proceeded to full-text assessment. Subsequently, 198 studies were excluded for documented reasons, leaving 70 for inclusion in the final corpus. The selection process and reasons for exclusion are summarized in Figure 1.

At the record management stage, records exported from Scopus are checked for duplicate entries using a combination of DOI, title, and bibliographic metadata. Entries identified as duplicates are retained as the most complete record to maintain consistency in corpus calculations at subsequent filtering stages. Screening and extraction were performed by one reviewer. Then decisions at critical points, especially exclusions at the full-text stage and the appropriateness of the exclusion reasons, were checked by a second reviewer. If there are differences in assessment at critical decision points, the final decision is determined through discussion until consensus is reached.

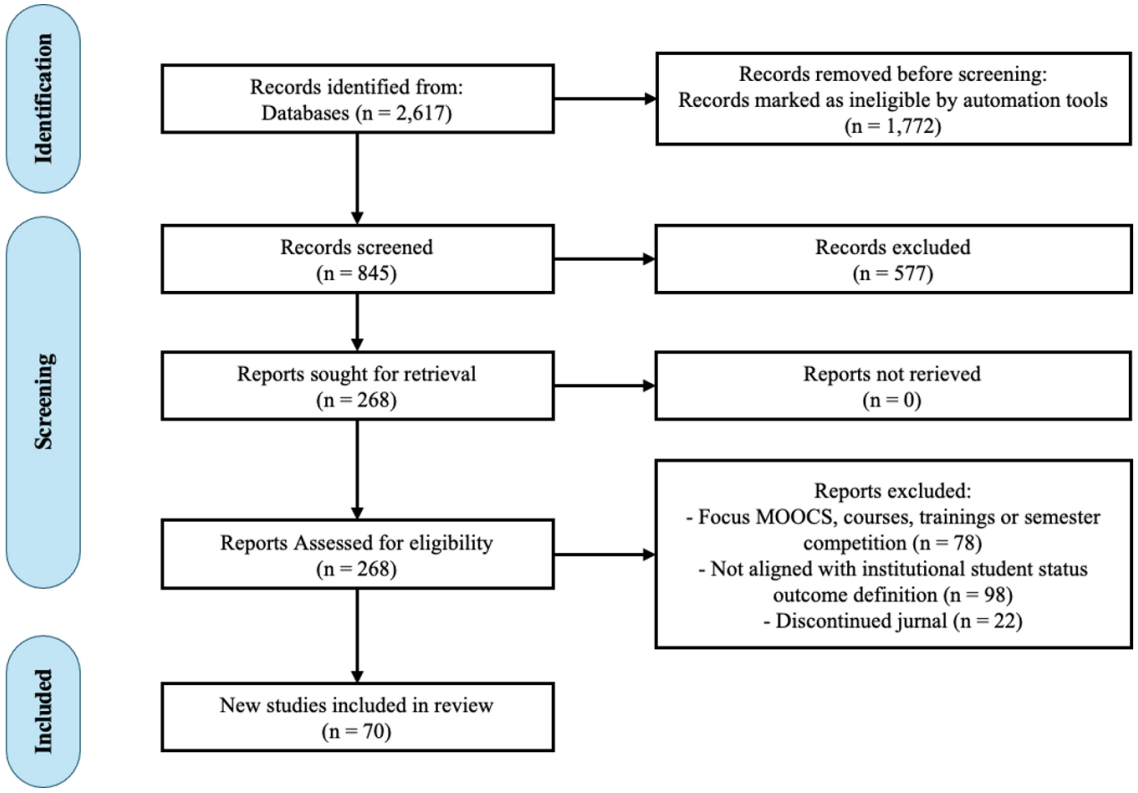

Figure 1: PRISMA flow diagram of study identification

Corpus Study Description and Data Characteristics

The final corpus consists of 70 journal articles that meet the inclusion criteria. The publication range is from 2017 to 2025. The distribution of articles by year of publication is shown in Figure 2 to illustrate the temporal distribution of research on predicting student study status within the corpus.

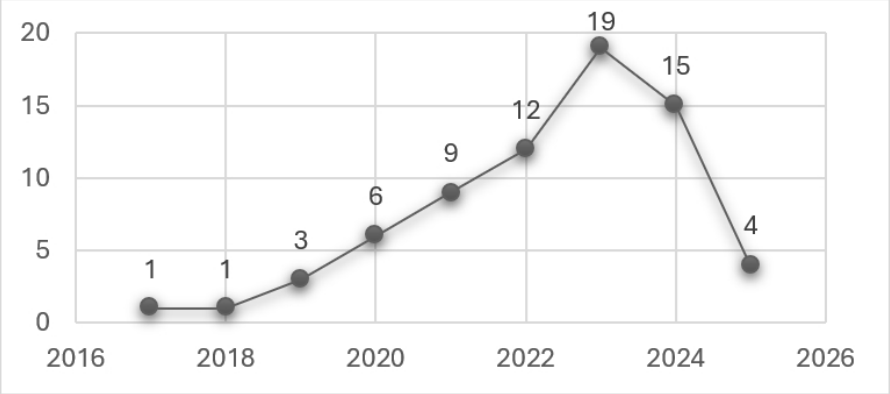

Figure 2: Distribution of articles by year of publication, 2017-2025

The distribution of articles by journal quartile is presented in Figure 3. This visualization describes the corpus, not the quality of each study, which is assessed through the primary studies' methodological components.

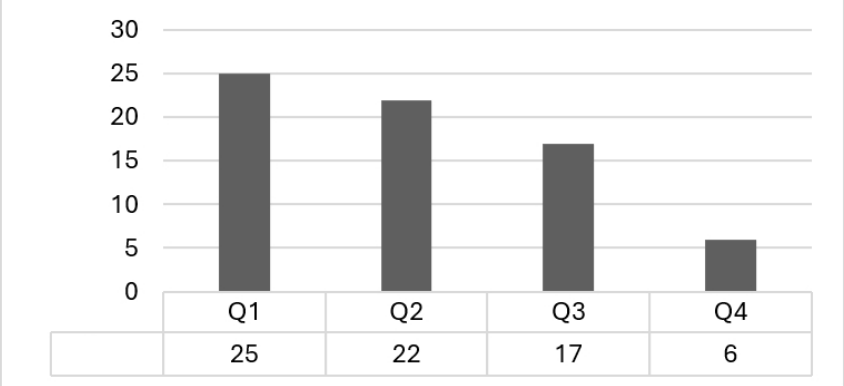

Figure 3: Article distribution across journal quartiles 2017-2025

The characteristics of the corpus are also reviewed based on the dataset's source, as this aspect affects the traceability of the experiment and the possibility of replication. Across the 70 studies, 62 (88.5%) relied on private institutional datasets, while 8 (11.5%) used public or open-access datasets. To support transparency in the reviewed corpus, a complete list of the 70 included studies is provided in Table 4.

The eight studies utilizing public or open-access datasets are not widely distributed across many sources but rather concentrated on a few reference datasets. Three studies used the Instituto Politécnico de Portalegre, Portugal dataset published through the UCI Machine Learning Repository by Realinho et al. (2021), namely the research by Goran et al. (2024), Martins et al. (2023), and Villar and de Andrade (2024). Two studies used the Tecnológico de Monterrey, Mexico dataset available through the open data repository by Alvarado-Uribe et al. (2022), namely Cañete-Sifuentes et al. (2023) and Gonzalez-Nucamendi et al. (2023). Additionally, Rovira et al. (2017) and Deleña et al. (2025) used datasets published alongside their articles, whereas Csalódi and Abonyi (2021) referenced the dataset's public location in the sources they cited. The implications of private dataset dominance for experiment reproducibility and generalization of findings are discussed in the Discussion section.

Data extraction and synthesis procedures

Data were extracted using a structured form to maintain traceability to the primary source. Extraction includes bibliographic information, data context, target definition, and number of classes, indication of class imbalance if reported, algorithms used, strategies for handling imbalance at both the data and algorithm levels, the use of ensembles or hybrid approaches, evaluation metrics, and data validation and splitting schemes. If available, information regarding hyperparameter tuning and computational efficiency is also recorded. This summary of extraction variables is used as the basis for descriptive and thematic synthesis.

The synthesis was conducted in two stages. The first stage involved a descriptive summary to map the distribution of findings according to the research questions. The second stage involved a cross-research question thematic synthesis to link study design decisions, including target formulation, degree of imbalance, mitigation strategies, and evaluation and validation plans.

No	Reff	Year	No	Reff	Year
1.	(Rabelo and Zárate, 2025)	2025	36.	(Haerani et al., 2023)	2023
2.	(Deleña et al., 2025)	2025	37.	(Mouchantaf and Chamoun, 2023)	2023
3.	(Oqaidi, Aouhassi, and Mansouri, 2025)	2025	38.	(Hoyos Osorio and Daza Santacoloma, 2023)	2023
.4	(Hooper, Ragland, and Artemiou, 2025)	2025	39.	(Niyogisubizo et al., 2022)	2022
5.	(Roslan et al., 2024)	2024	40.	(Vidal et al., 2022)	2022
6.	(Delogu et al., 2024)	2024	41.	(Segura, Mello and Hernández, 2022)	2022
7.	(Nguyen Thi Cam, Sarlan and Arshad, 2024)	2024	42.	(Barramuño, Meza-Narváez and Gálvez-García, 2022)	2022
8.	(Villar and de Andrade, 2024)	2024	43.	(Moreira da Silva et al., 2022)	2022
9.	(Vaarma and Li, 2024)	2024	44.	(Cannistrà et al., 2022)	2022
10.	(Goran et al., 2024)	2024	45.	(Nuanmeesri et al., 2022)	2022
11.	(Zanellati, Zingaro and Gabbrielli, 2024)	2024	46.	(Hammoodi and Al-Azawei, 2022)	2022
12.	(Okoye et al., 2024)	2024	47.	(Vega et al., 2022)	2022
13.	(Nagy and Molontay, 2024)	2024	48.	(Canto, De Oliveira and De Mattos Veroneze, 2022)	2022
14.	(Ndunagu et al., 2024)	2024	49.	(Yaqin, Rahardi, and Abdulloh, 2022)	2022
15.	(Setiadi et al., 2024)	2024	50.	(Rose and Mary.T, 2022)	2022
16.	(Herianto et al., 2024)	2024	51.	(Fernandez-Garcia et al., 2021)	2021
17.	(Darenoh, Bachtiar, and Perdana, 2024)	2024	52.	(Opazo et al., 2021)	2021
18.	(Sayed, 2024)	2024	53.	(Csalódi and Abonyi, 2021)	2021
19.	(Anagnostopoulos et al., 2024)	2024	54.	(Uliyan et al., 2021)	2021
20.	(Delen, Davazdahemami and Rasouli Dezfouli, 2024)	2024	55.	(Palacios et al., 2021)	2021
21.	(Cho, Yu and Kim, 2023)	2023	56.	(Fontana et al., 2021)	2021
22.	(Phan, De Caigny and Coussement, 2023)	2023	57.	(Nanglae et al., 2021)	2021
23.	(Hammoudi Halat et al., 2023)	2023	58.	(Cuizon, 2021)	2021
24.	(Matz et al., 2023)	2023	59.	(Yaqin, Laksito, and Fatonah, 2021)	2021
25.	(Villegas-Ch, Govea, and Revelo-Tapia, 2023)	2023	60.	(Tsai et al., 2020)	2020
26.	(Cañete-Sifuentes et al., 2023)	2023	61.	(Sandoval-Palis et al., 2020)	2020
27.	(Kaensar and Wongnin, 2023)	2023	62.	(Freitas et al., 2020)	2020
28.	(Salam and Zeniarja, 2023)	2023	63.	(Alvarez, Callejas and Griol, 2020)	2020
29.	(Gutierrez-Pachas et al., 2023)	2023	64.	(Sani et al., 2020)	2020
30.	(Won et al., 2023)	2023	65.	(Bedregal-Alpaca et al., 2020)	2020
31.	(Kim et al., 2023)	2023	66.	(Rodríguez-Muñiz et al., 2019)	2019
32.	(Martins et al., 2023)	2023	67.	(Ortigosa et al., 2019)	2019
33.	(Anggrawan, Hairani, and Satria, 2023)	2023	68.	(Febro, 2019)	2019
34.	(Gonzalez-Nucamendi et al., 2023)	2023	69.	(Arumugam, Vinodhini, and Chandrasekaran, 2018)	2018
35.	(Song et al., 2023)	2023	70.	(Rovira, Puertas and Igual, 2017)	2017

Table 4: List of studies included in the SLR 2017-2025

Quality of reporting assessment

Quality assessment is positioned as an evaluation of the quality of reporting relevant to the auditability of the methodology, rather than as an absolute quality assessment. A study is considered adequate if it includes at least five components: the modeling method, class balancing strategies or an explicit statement regarding imbalance, evaluation metrics, validation design or data splitting, and the use or absence of hybrid or ensemble approaches if claimed as a contribution (Page et al., 2021). The assessment was conducted using a binary checklist for each component, and a summary of the results provided context for interpreting the findings in the results and discussion sections.

RESULTS

RQ1: What machine learning and deep learning algorithms are used to predict student dropout and graduation?

To answer RQ1, this review synthesizes the machine learning and deep learning algorithms used in the study corpus to predict students' academic status in higher education, including dropout, retention, and graduation. A summary of algorithm usage distribution at the corpus level is presented in Table 5. One study can report more than one model, so the frequencies in the table represent the number of studies reporting that model and are not mutually exclusive.

Algorithm / Family	Number of Studies	Percentage
Random Forest	34	48.6%
Decision Tree (C4.5/CART)	28	40.0%
Logistic Regression	27	38.6%
Support Vector Machine (SVM)	27	38.6%
KNearest Neighbors (KNN)	16	22.9%
Naïve Bayes	10	14.3%
MLP / Deep Neural Network	10	14.3%
XGBoost	15	21.4%
LightGBM	6	8.6%
CatBoost	5	7.1%
Gradient Boosting lain	9	12.9%
AdaBoost	6	8.6%
CNN	4	5.7%
LSTM / RNN	4	5.7%
Ensemble eksplisit (stacking, dsb)	6	8.6%

Table 5: Distribution of model usage in the study corpus

Based on Table 5, tree-based and tree ensemble models are the most frequently reported, particularly Random Forest and Decision Tree. Baseline models are also reported in a significant proportion, including logistic regression and SVM. Within the boosting group, XGBoost is reported in several studies, along with other boosting variations such as Gradient Boosting, LightGBM, CatBoost, and AdaBoost. In the deep learning group, MLPs or deep neural networks are reported in a portion of the studies, while CNNs, LSTMs, or RNNs are reported in smaller proportions. Additionally, some studies report the use of explicit ensembles such as stacking as a model combination configuration.

To provide context for reporting at the study level, representative study examples, classification scenarios, main models, and reported metrics are presented in Table 6. The examples in Table 6 are used to illustrate the variations in modeling scenarios and selected metrics, not to compare performance across studies.

Main Scenarios and Models	Summary of study metrics and context notes
Binary, Random Forest	Accuracy: 95.93 percent; F1 score: approximately 0.88. Dropout prediction in the B40 context (Sani et al., 2020).
Binary, XGBoost	AUC is approximately 0.97; accuracy is approximately 94.1 percent (Canto, De Oliveira, and De Mattos Veroneze, 2022; Haerani et al., 2023; Kim et al., 2023)
Binary, Decision Tree C4.5	Accuracy is approximately 89 percent. Predicting dropout in the Malaysian context (Roslan et al., 2024).
Binary, Stacking SMLOS with SMOTE and Optuna	Accuracy is 95.5 percent. Configuration using resampling and hyperparameter tuning (Herianto et al., 2024).
Binary, Voting ensemble LR, DT, and ANN	Recall dropout is about 98 percent. Reporting emphasizes dropout metrics (Rabelo and Zárate, 2025).
Multiclass: 3 classes; XGBoost-tuned AGbSCHO.	Accuracy 88.00 percent; Cohen's kappa 0.666. Metaheuristic tuning in multi-class scenarios (Goran et al., 2024).
Multiclass for three classes, LightGBM and CatBoost tuned with Optuna	F1 dropout 0.88; F1 graduate 0.86; F1 enrolled 0.83. Reporting metrics per class (Villar and de Andrade, 2024).
Multiclass for three classes, Random Forest with SVMSMOTE	Balanced accuracy: 74.8 percent; global F1: 0.745. Reporting balanced accuracy and global F1 in multi-class scenarios (Martins et al., 2023).
Multiclass for three classes, C4.5	F-measure continues at 99.6 percent; dropout at 72.0 percent; change at 44.4 percent. The F-measure is reported for each class, including the change class (Rodríguez-Muñiz et al., 2019).

Table 6: Representative studies, classification scenarios, main models, and metrics reported on predicting student academic status, 2017 to 2025

In line with Table 6, in the binary scenario, the corpus includes Random Forest and XGBoost reporting with metrics such as accuracy, F1, and AUC, as well as Decision Tree C4.5 with the accuracy metric. The corpus also includes studies that report ensemble configurations through stacking and voting, with metric reporting emphasizing dropout classes such as recall. In a three-class multiclass scenario, the corpus includes studies reporting XGBoost with Cohen's kappa, studies reporting LightGBM and CatBoost with F1 per-class, and studies reporting Random Forest with SVMSMOTE using balanced accuracy and global F1. In C4.5-based multiclass studies, F-measure reporting varies across classes and lists values for each status.

RQ2: How are targets formulated as binary or multiclass classification in predicting student study status, and what implications follow for modelling and reporting?

To answer RQ2, this review synthesizes how studies in the SLR corpus frame the target of predicting student study status as binary or multiclass classification, including those that test both scenarios within a single experimental design. A summary of the target formulation distribution at the study level is presented in Table 7. Quantitatively, 64 out of 70 studies only conducted binary experiments, 3 studies only conducted multi-class experiments, and 3 studies ran hybrid binary and multi-class scenarios (Rodríguez-Muñiz et al., 2019; Alvarez, Callejas, and Griol, 2020; Uliyan et al., 2021; Martins et al., 2023; Goran et al., 2024; Villar and de Andrade, 2024). Of the experiments conducted, 67 evaluated binary scenarios, while 6 evaluated multi-class scenarios.

Study Category	Number of Studies	Percentage
Binary only (experiment with only 2 classes)	64	91.4%
Multiclass only (experiments with more than 2 classes)	3	4.3%
Hybrid (runs binary and multi-class)	3	4.3%
Total number of studies conducting binary experiments	**67**	**95.7%**
Total number of studies conducting multi-class experiments	**6**	**8.6%**

Table 7: Distribution of target class formulations in the SLR corpus

Based on Table 7, the majority of studies conducted binary experiments, either as the sole scenario or as part of a hybrid scenario. In binary formulations, the target is typically expressed as dropout versus non-dropout, or graduation versus non-graduation, within a specific time horizon. Examples of the dropout-versus-non-dropout binary formulation are reported in the studies by Roslan et al. (2024) and Sani et al. (2020), which model dropout as the primary output. Examples of binary formulations for graduation or dropout output are also found in the studies by Canto et al. (2022), Haerani et al. (2023), and Kim et al. (2023), which report graduation or dropout predictions using metrics such as AUC and accuracy.

In contrast, multi-class formulations represent the target as multiple states, each corresponding to the student's study path in greater detail. In the corpus, a frequently occurring example is three study status classes, such as Graduate, Enrolled, and Dropout, which are evaluated by reporting metrics per-class (Villar and de Andrade, 2024), as well as three classes of Dropout, Enrolled, and Graduate in another multi-class configuration (Goran et al., 2024). Variations in the definition of multi-class labels are also found in other contexts, such as Continue, Dropout, and Change (Rodríguez-Muñiz et al., 2019); Promotion, Repetition, and Dropout (Alvarez, Callejas, and Griol, 2020); and learning progress-based labels such as Ongoing or Normal, At Risk, and Unsurpassed (Uliyan et al., 2021).

RQ3: What class imbalance handling strategies are reported, and what application patterns emerge?

To answer RQ3, this review examines the class imbalance handling strategies reported in studies predicting student study status, including interventions at the data, algorithm, and decision levels. The distribution of strategies reported at the corpus level is presented in Table 8. In this synthesis, the category "not reported" refers to studies that do not explicitly address class imbalance, making it impossible to trace mitigation strategies from the primary studies.

Reported strategies	Number of Studies	Percentage
Not reported	33	47.1%
Oversampling	16	22.9%
Undersampling	6	8.6%
Combined sampling and cost-sensitive sampling	5	7.1%
Cost-sensitive or class-weighted sampling	3	4.3%
Hybrid resampling and cleaning	2	2.9%
Decision threshold adjustment	2	2.9%
Balanced or stated design	2	2.9%
Ensemble-based sampling	1	1.4%

Table 8: Distribution of class imbalance handling strategies reported in the SLR corpus, 2017-2025

Based on Table 8, almost half of the studies did not explicitly report on the treatment of class imbalance. Among the studies reporting strategies, resampling at the data-level is the most frequently mentioned approach, particularly oversampling and undersampling. Additionally, the corpus also includes strategies at the algorithm and decision level, such as cost-sensitive learning or class weighting, decision threshold adjustment, and sampling-based ensembles.

At the data level, the studies by Song et al. (2023) and Yaqin et al. (2021, 2022) reported the use of family-based oversampling techniques such as SMOTE and its variations. In the study by Villar and de Andrade (2024)

on imbalanced multiclass data, the corpus also included comparisons of oversampling techniques like SMOTE and ADASYN. Undersampling was reported in several studies to reduce the dominance of the majority class or to explore class ratios in rare dropout conditions (Opazo et al., 2021; Cañete-Sifuentes et al., 2023).

At the algorithm and decision level, the corpus includes approaches such as class weighting or cost-sensitive learning, probability threshold adjustment, and sampling-based ensembles (Barramuño, Meza-Narváez, and Gálvez-García, 2022; Gonzalez-Nucamendi et al., 2023; Villegas-Ch, Govea, and Revelo-Tapia, 2023; Delen, Davazdahemami, and Rasouli Dezfouli, 2024). Additionally, studies by Alvarez et al. (2020) and Martins et al. (2023) report problem transformation through class merging to reduce label imbalance in extreme minority conditions. In certain configurations, the corpus also includes hybrid techniques that combine resampling and cleaning (Kim et al., 2023).

RQ4: To what extent are ensemble and hybrid methods used, and what application patterns emerge in this corpus?

To answer RQ4, this review examines how studies in the corpus apply ensemble and hybrid approaches to predicting student study status and identifies the application patterns that emerge across study designs. In this RQ, the term "ensemble" refers to the strategy of combining multiple models to produce a final prediction, such as stacking, voting, bagging, or boosting (Cuizon, 2021). The term "hybrid" refers to designs that combine different methods within a single workflow, such as combining deep learning and machine learning, or building a staged pipeline based on clustering or feature selection (Phan, De Caigny, and Coussement, 2023).

A summary of the ensemble and hybrid approach categorizations identified in the corpus is presented in Table 9 to illustrate the variety of configurations and examples of studies within each category.

Approach Categories in RQ4	Number of Studies	Recorded Studies
Ensemble Stacking	2	(Niyogisubizo et al., 2022; Herianto et al., 2024)
Ensemble Voting	4	(Cuizon, 2021; Cañete-Sifuentes et al., 2023; Okoye et al., 2024; Rabelo and Zárate, 2025)
Weighted and Cascading Voting	1	(Fernandez-Garcia et al., 2021)
Random Forest-Based Ensemble Bagging	4	(Sani et al., 2020; Palacios et al., 2021; Kaensar and Wongnin, 2023; Matz et al., 2023)
Boosting as an Ensemble	1	(Hammoudi Halat et al., 2023)
Ensembles for Imbalanced Data	1	(Martins et al., 2023)
Hybrids of Deep Learning and Machine Learning or Model Combination	3	(Kim et al., 2023; Delen, Davazdahemami, and Rasouli Dezfouli, 2024; Nguyen Thi Cam, Sarlan, and Arshad, 2024)
Hybrid Stepwise Clustering or Feature Selection-Based Hybrids	2	(Nanglae et al., 2021; Nuanmeesri et al., 2022)
Hybrids Across Analytical Paradigms	1	(Csalódi and Abonyi, 2021)
Dual Classification and Survival Modeling	1	(Gutierrez-Pachas et al., 2023)

Table 9: Summary of ensemble and hybrid approach categories in the study corpus, 2017 To 2025

Based on Table 9, the most frequently occurring categories are voting ensembles and random forest-based bagging, each appearing in 4 studies. The corpus also includes hybrid configurations that combine modeling paradigms or stages, such as integrating deep learning and machine learning, as well as cluster-based or feature-selection-based staged pipelines. To provide context for reporting at the study level, the following description summarizes the configuration, comparators, and performance findings as reported in the primary studies.

Herianto et al. (2024) reported hybrid stacking via SMLOS, which combines multiple base models with a meta model and compares it to individually optimized base models, with the highest reported accuracy in the tested configuration. Cañete-Sifuentes et al. (2023) reported on a voting ensemble based on machine learning automation, where VotingEnsemble combines multiple tree-based models and is compared to single models and models specifically designed to handle class imbalance, reporting the combination of true positive rate and false positive rate at a specific ratio.

Fernandez-Garcia et al. (2021) reported a proportional weighted ensemble combining gradient boosting, random forests, and support vector machines, with individual model comparisons. They observed changes in recall and precision in the first semester. Martins et al. (2023) reported on ensembles for imbalanced data by comparing Balanced Random Forest and Easy Ensemble with resampling pipelines and standard models, namely SMOTE with Random Forest and SVMSMOTE with Random Forest. They reported F1 scores and balanced accuracy for the minority class in a multi-class scenario. Rabelo and Zárate (2025) reported a classic voting ensemble that combines CART, logistic regression, and artificial neural networks, with individual model comparisons, and found higher prediction stability in the study context. Hammoudi Halat et al. (2023) reported boosting as an ensemble with the comparators used in the study and reported performance results on the tested configurations.

RQ5: Which evaluation metrics and validation methods are used, and how suitable are they for imbalanced data?

To answer RQ5, the corpus shows that evaluating the prediction of student study status becomes less representative when relying solely on accuracy, especially when class distributions are

imbalanced. Several studies improve accuracy using confusion matrix-based metrics, including precision, recall, and F1 score, as well as metrics that are more sensitive to imbalance, such as balanced accuracy and G-mean (Yaqin, Rahardi, and Abdulloh, 2022; Kim et al., 2023; Martins et al., 2023).

To complement the tabular synthesis, Figure 4 presents two representative confusion matrices from a primary study to illustrate class-wise error patterns in a multi-class setting. This visualization supports the argument that per-class reporting and confusion matrix reading are essential under class imbalance, particularly when minority class detection is central to intervention decisions. The figure is provided as a representative example and does not imply an aggregated benchmark across heterogeneous studies.

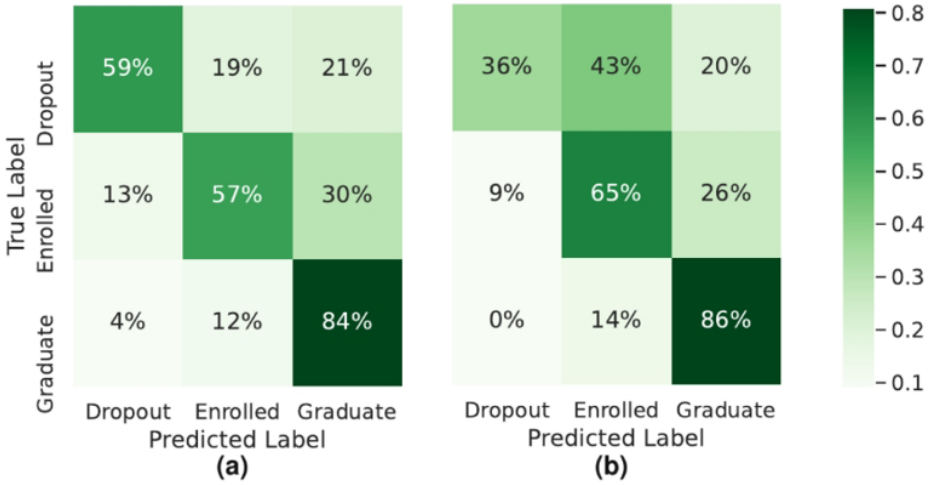

Figure 4: Representative confusion matrices for multi-class study status prediction illustrating class-wise error patterns. (A) Example from dataset S0. (B) Example from dataset S2. Reproduced from (Martins et al., 2023)

Some studies report correlation and agreement metrics, such as the Matthews correlation coefficient and Cohen's kappa, as performance summaries that account for all parts of the confusion matrix (de la Cruz Huayanay, Bazán, and Russo, 2024). In the corpus, this metric appears as a summary alternative when studies want to present a performance measure that does not rely on simple aggregation.

In area-based metrics, the corpus includes reporting of AUC-ROC and precision-recall-based metrics, including AUC-PRPR (Luque et al., 2019; Palacios et al., 2021; Martins et al., 2023; Delogu et al., 2024; Vaarma and Li, 2024). Several studies highlight the precision-recall curve and AUC PR as important complements when the evaluation focus is directed toward rare positive classes, while AUC ROC is still reported to maintain comparability with more common reporting in the literature (Luque et al., 2019; Palacios et al., 2021; Delogu et al., 2024).

The corpus also includes error metric reporting, such as false-positive and false-negative rates, to describe more specific error patterns within the institution's classes of interest. In this context, some studies emphasize the false negative rate when the evaluation focus is on the risk of failing to detect at-risk students (Cañete-Sifuentes et al., 2023; Okoye et al., 2024).

Beyond the metrics, the validation designs used in the primary studies varied. The corpus reports the use of hold-out, stratified k-fold cross-validation, cross-validation with external validation, nested cross-validation, and temporal validation (Moreira da Silva et al., 2022; Niyogisubizo et al., 2022; Kim et al., 2023; Martins et al., 2023; Phan, De Caigny, and Coussement, 2023). Stratified k-fold is said to keep the representation of minority classes in each fold, while nested cross-validation is used in studies that combine evaluation with more systematic hyperparameter tuning (Martins et al., 2023; Phan, De Caigny, and Coussement, 2023). Temporal validation is also reported when studies test the consistency of models across cohorts or academic periods (Moreira da Silva et al., 2022; Kim et al., 2023).

In some studies, the class balancing technique is also described as part of the evaluation pipeline. The corpus includes reports indicating that resampling techniques, such as SMOTE, were applied to the training data within cross-validation schemes by implementing them in each training fold, which allows the evaluation procedure to be traced without merging the training and test data during the balancing stage (Kim et al., 2023; Martins et al., 2023; Song et al., 2023).

RQ6: What methodological gaps and challenges remain, and what recommendations strengthen future research?

To answer RQ6, the corpus synthesis shows that research on predicting student study status is developing rapidly, but still leaves methodological gaps that can affect the validity of conclusions, especially when the problem is formulated as multiclass classification with imbalanced label distributions. In the corpus, binary formulations remain dominant, while studies explicitly testing multi-class classification are relatively limited. In available multi-class studies, the prominent challenges are not only the decline in aggregate performance but also the performance disparity between classes, especially when one class becomes an extreme minority or when classes are conceptually close and easily confused. Therefore, reporting metrics by class and

analyzing error patterns emerged as a crucial methodological need for understanding performance readability at relevant study statuses for intervention. Confusion matrix analysis should be used to identify systematic confusions between conceptually adjacent statuses and to quantify minority-class errors that aggregate summaries may hide. Figure 4 is included as a representative example to illustrate how such confusions can be inspected in multi-class settings (Rodríguez-Muñiz et al., 2019; Uliyan et al., 2021; Martins et al., 2023; Goran et al., 2024; Villar and de Andrade, 2024).

The next gap concerns the traceability of methodological decisions for handling class imbalance. Several studies have reported data-level strategies, including resampling (e.g., SMOTE and its derivatives), hybrid approaches that combine oversampling and cleaning, and undersampling. Other studies have reported algorithm-level strategies such as class weighting and cost-sensitive learning. However, the corpus also shows that reporting on the treatment of imbalance is not always explicit, making it difficult for readers to assess whether improvements in minority classes stem from balancing strategies, model selection, or procedural consequences of the evaluation. The key concern at this juncture is the placement of resampling within the assessment pipeline, as implementing it before data partitioning may yield excessively optimistic performance estimates due to information leakage. Therefore, this review recommends that resampling techniques such as SMOTE be applied strictly to the training folds within the cross-validation scheme, after the split is created. The validation fold and any held-out test set must remain untouched to prevent information leakage and overly optimistic estimates (Cañete-Sifuentes et al., 2023; Kim et al., 2023; Song et al., 2023).

To improve auditability and to avoid overly optimistic estimates caused by information leakage, Figure 5 summarizes a leakage-free evaluation workflow for imbalanced multi-class classification. The workflow emphasizes that preprocessing and resampling must be performed only within each training fold after splitting, while the validation fold remains untouched until evaluation. Procedural guidance for leakage-free evaluation under class imbalance is as follows. First, define the target mapping rules and report the class distribution. Second, select a validation design that matches the intended deployment scenario, such as stratified cross-validation, temporal validation, or external validation. Third, within each training fold only, fit preprocessing steps, apply resampling, and then fit the model, while keeping the validation fold untouched. Fourth, when hyperparameter tuning is performed, use a nested cross-validation design or an inner loop to prevent information leakage into the evaluation. Fifth, evaluate on the untouched validation fold using per-class precision, recall, and F1, together with imbalance-sensitive summaries such as macro F1 and balanced accuracy, and interpret results through confusion matrix diagnostics.

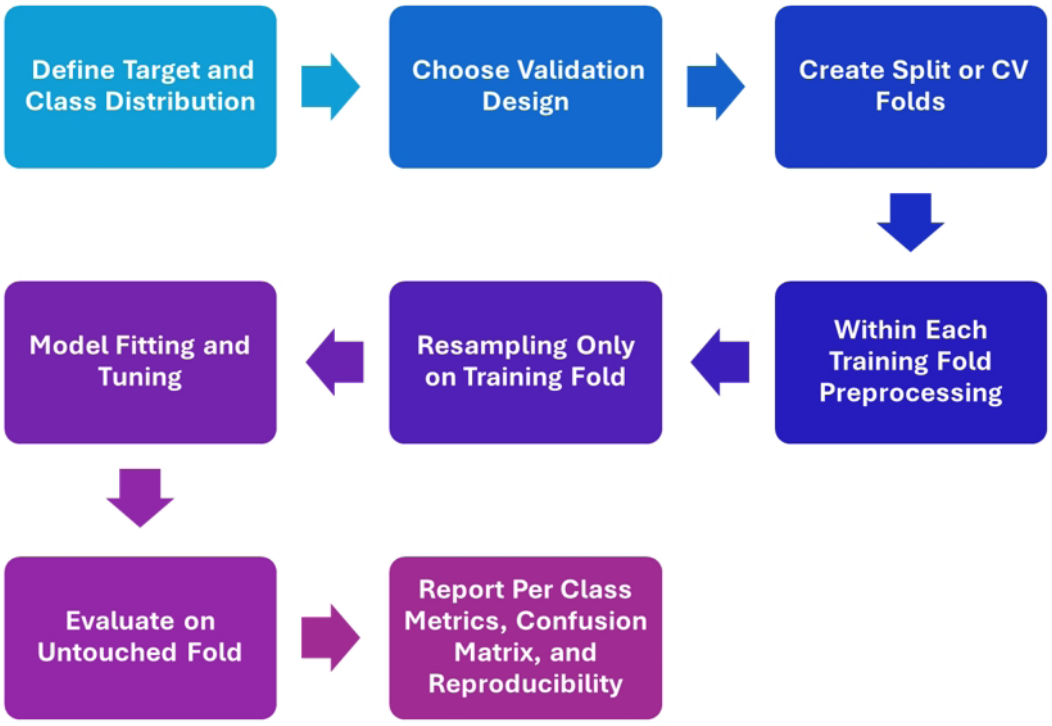

Figure 5: Workflow for leakage-free evaluation under multi-class class imbalance, emphasizing preprocessing and resampling only within training folds, while validation folds remain untouched until evaluation

In answering RQ5, the corpus indicates that research closely linked to the traits of imbalanced data typically improves accuracy by using measures derived from confusion matrices, imbalance-sensitive metrics, and area-based and error-based metrics. This variation is closely associated with the model's intended application, particularly when its output informs risk rating and intervention prioritization. The corpus presents a variety of validation methodologies, including hold-out, stratified k-fold cross-validation, external validation, nested cross-validation, and temporal validation. This design variation demonstrates that the reliability of performance estimates is significantly affected by the data partitioning, the management of hyperparameter tuning, and the inclusion of cross-cohort or cross-temporal assessments (Moreira da Silva et al., 2022; Niyogisubizo et al., 2022; Kim et al., 2023; Martins et al., 2023; Phan, De Caigny, and Coussement, 2023).

Another recurring challenge in the corpus concerns the repeatability of experiments and opportunities for cross-institutional replication. The corpus demonstrates a dominance of internal private institutional datasets over public or open-access datasets. This condition enriches the institutional context but, at the same time, limits the exact replication of experiments by other researchers due to data access, the definition of study status labels, and feature structures that are often tied to local academic policies. In the corpus, the distribution of dataset sources shows a very high proportion of private datasets relative to public or open-access datasets, making replication and generalizability issues important contexts for interpreting cross-study findings.

To promote a more open future research agenda despite private data constraints, this review recommends standardizing feature structures and reporting artifacts so that models developed on internal datasets can still be tested for cross-institutional generalizability. Concretely, studies should publish a feature schema listing feature names, definitions, types, allowable ranges, missing-value handling, and encoding rules, along with a mapping from local variables to the shared schema. In addition, studies should document target mapping rules and prediction horizons using consistent terminology and provide evaluation scripts or pseudocode that enable external teams to replicate preprocessing, splitting protocols, and metric computation on their own institutional data. These steps do not require releasing identifiable student records, yet they enable reproducible comparisons and cross-institutional validation through a shared representation of the problem.

To strengthen traceability and auditability, the corpus also indicates the need for consistent minimal reporting standards across studies. In this manuscript, quality assessment is positioned as an evaluation of reporting quality relevant to methodological auditability, with minimal components including modeling methods, strategies for handling imbalance or explicit statements regarding imbalance, evaluation metrics, validation designs or data splitting, and clarity of use or the absence of hybrid or ensemble approaches when claimed as contributions (Page et al., 2021). Minimal reporting checklist for auditability and reproducibility: (1) Target definition and mapping rules, including the institutional policy assumptions used to label dropout or graduation related outcomes; (2) Class distribution reported for the overall dataset and for each split or fold, not only at the full dataset level; (3) Data splitting and validation protocol, including whether the split is stratified, temporal, or uses an external test set, and the rationale for the choice; (4) Pipeline specification that explicitly states the sequence of preprocessing, resampling, and model fitting, and explicitly states that resampling is performed only within training folds; (5) Hyperparameter tuning protocol, including whether nested cross-validation or an inner loop is used, and which data are used for selection; (6) Evaluation metrics including per-class precision, recall, and F1, together with at least one imbalance-sensitive summary such as macro F1 or balanced accuracy, and an additional summary such as MCC or kappa when applicable; (7) Confusion matrix analysis describing systematic confusions between conceptually adjacent classes and the error profile of the minority class; (8) Reproducibility artifacts including random seed control, software libraries, and a description of the feature schema sufficient for replication or cross-institutional testing. This framework helps read the findings of RQ1-RQ5 in a more structured way, while also highlighting the methodological sections that are most often a source of uncertainty in the interpretation of primary studies. As a cross-RQ1 to RQ6 summary, the dimensions of target formulation, degree of imbalance, handling strategies, and evaluation and validation can be mapped as a study design taxonomy synthesized in the manuscript and presented in Figure 6. This taxonomy is presented as a summary of findings that facilitates pattern reading and as a framework for formulating a methodological strengthening agenda for subsequent research, without making any single configuration a claim of universal superiority across all institutional contexts.

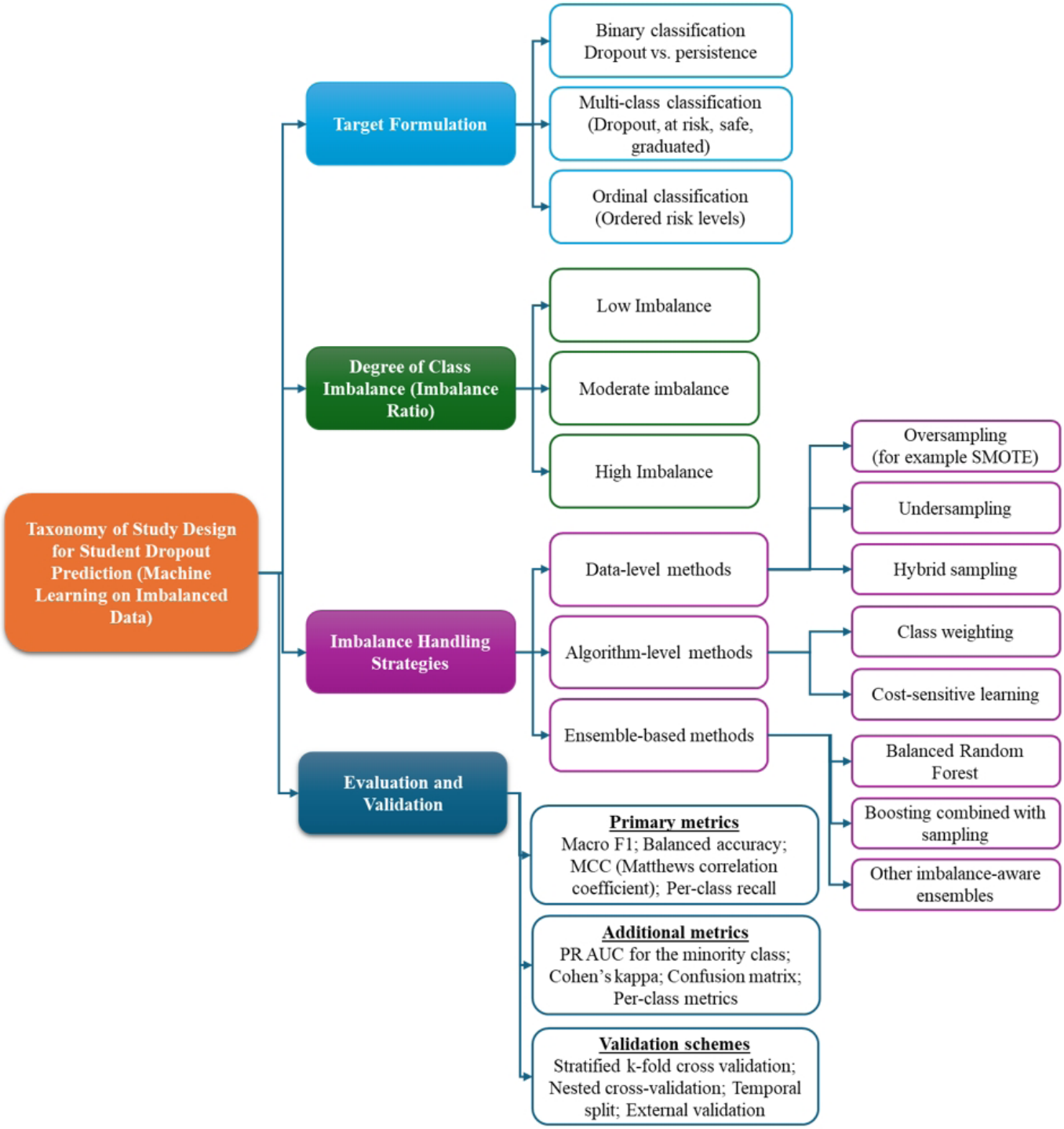

Figure 6: Study design taxonomy for predicting student dropout on imbalanced data, synthesized from findings RQ1 to RQ6

DISCUSSION

The synthesis in the Results section indicates that predicting students' study status in higher education has become an established research theme, characterized by a variety of publication venues and the widespread use of machine learning, including tree-based models, boosting, and some forms of deep learning. However, across the RQs, it also appears that the maturity of modeling is not always matched by that of evaluation or the traceability of study design decisions. This inequality is important within the framework of efficiency and responsibility in education and science. From an efficiency perspective, indefensible evaluation designs can lead to inefficient allocation of limited support capacity due to misleading error trade-offs. From a responsibility perspective, the same indefensible designs can lead to conclusions that are not defensible for educational decision support and not accountable as scientific evidence (Vidal et al., 2022; Villegas-Ch, Govea, and Revelo-Tapia, 2023). Accordingly, this Discussion interprets the results in terms of operational implications for efficient interventions and methodological

implications for responsible evidence, and it concludes with actionable recommendations derived from the taxonomy, including reporting, evaluation, and reproducibility priorities.

The findings in Table 5 and the case studies in Table 6 confirm the dominance of the pragmatic tabular approach, particularly tree-based methods such as tree families and tree ensembles, with boosting. This pattern aligns with the generally administrative and tabular nature of academic data and the relatively easy implementation requirements. However, for cross-institutional generalization, the primary value of this SLR is better characterized as mapping the patterns of study design decisions that determine the strength of performance claims, rather than ranking cross-study models. Variations in the definition of study status and prediction horizon, and differences in curriculum structure and academic policies, often make performance figures between studies incomparable, so consistency in defining targets and the defensibility of evaluations are prerequisites for interpretation (Rodríguez-Muñiz et al., 2019; Palacios et al., 2021). The private dataset dominance observed in the corpus further constrains cross-institutional generalizability because target definitions, feature construction, and academic policies are often institution-specific. Therefore, an open research agenda in this domain should prioritize the portability of research artifacts rather than that of raw data. A practical path is to standardize feature structures through shared feature schemas and data dictionaries that preserve privacy while enabling cross-institutional testing. Under this agenda, different institutions can implement a compatible feature schema locally and evaluate the same modelling and validation protocol on their own cohorts, which supports stronger external validation evidence without requiring sensitive record sharing.

From the perspective of target formulation, Table 7 shows that binary experiments dominate, while multi-class experiments are relatively limited. The dominance of binary experiments can be understood as a simplification of the early warning task. Still, multi-class studies in the corpus show a complexity closer to the reality of the institutional study status. In multi-class scenarios, performance differences between classes become more apparent, especially when there are extreme minority classes or conceptually close status categories that are easily confused (Rodríguez-Muñiz et al., 2019; Martins et al., 2023; Goran et al., 2024; Villar and de Andrade, 2024). The implication is that a single-number performance summary based on aggregate data becomes increasingly risky for multiple classes because it can mask weaknesses in the status most relevant for intervention.

Regarding class imbalance, Table 8 shows that the reported handling strategies included resampling, class weighting, cost-sensitive learning, and other variations, but the proportion of studies that did not explicitly state the imbalance treatment remained high. This limitation is not marginal: Table 8 shows that 33 of 70 studies (47.1%) did not explicitly report class-imbalance handling strategies. This lack of procedural reporting is a significant obstacle to auditability and reproducibility because readers cannot verify whether reported improvements in the minority class arise from the intended imbalance strategy, from model choice, or from evaluation design side effects. This is a critical reporting issue because the lack of procedural information makes it difficult to assess whether performance is driven by the appropriate strategy or by estimation bias arising from a loose evaluation design. The literature emphasizes that the key issue is not just the use of resampling but its placement within the evaluation pipeline, specifically, whether it is applied only to the training data, including to each training fold during cross-validation, to prevent information leakage and overly optimistic estimates (Kim et al., 2023; Martins et al., 2023; Song et al., 2023). Therefore, defensible reporting needs to explicitly state the sequence of procedures, including data separation, validation schemes, hyperparameter tuning, and resampling positions.

On the metric side, the findings of RQ5 reinforce the conclusion that accuracy is an inadequate single summary measure for class imbalance. More responsible practices are evident in studies that complement accuracy with precision, recall, F1, and imbalance-sensitive metrics such as balanced accuracy and G-mean (Yaqin, Rahardi, and Abdulloh, 2022; Kim et al., 2023; Martins et al., 2023). The corpus also includes summary metrics that consider all components of the confusion matrix, such as the Matthews correlation coefficient and Cohen's kappa (de la Cruz Huayanay, Bazán, and Russo, 2024). In curve-based evaluation, the literature confirms that in extreme imbalances, ROC can appear satisfactory even though the performance of the rare positive class remains weak, making precision-recall-based metrics, including AUC PR, more informative when the institution's goal is to rank dropout risk (Luque et al., 2019; Palacios et al., 2021; Martins et al., 2023; Delogu et al., 2024; Vaarma and Li, 2024). At the implementation level, the cost of errors is reflected in the reporting of false-positive and false-negative rates, especially when the focus is on the risk of failing to detect at-risk students (Cañete-Sifuentes et al., 2023; Okoye et al., 2024). This confirms that the efficiency of interventions depends on managing the trade-off between errors and academic service capacity, not solely on average performance.

RQ4 shows that ensembles and hybrid configurations are present with varying intensities (Table 9). The emergence of voting ensembles and bagging based on Random Forest can be interpreted as a pragmatic strategy to improve prediction stability on tabular data. However, adding complexity through stacking, weighted voting, or hybrid designs does not always make implementation more efficient unless there is strict validation and clear reporting. Some studies show benefits in their own test settings, such as hybrid stacking (Herianto et al., 2024) and changes in the trade-off for automation-based voting (Cañete-Sifuentes et al., 2023). Other studies focus on changes in recall and precision in the early stages that are important for early warning (Fernandez-Garcia et al., 2021). Therefore, ensembles and hybrids are more accurately positioned as tools for managing trade-offs in specific contexts, rather than as guarantees of universal improvement (Vidal et al., 2022; Villegas-Ch, Govea, and Revelo-Tapia, 2023).

As a cross-RQ synthesis, Figure 6 summarizes the most critical design decisions for result reliability, namely target formulation, degree of imbalance, imbalance-handling strategies, and evaluation and validation. This taxonomy confirms that algorithm selection cannot be separated from

more fundamental decisions, particularly the definition of labels, the prediction horizon, and the validation scheme. Thus, Figure 6 can serve as an audit framework for reporting prior studies and as a checklist for designing new studies, thereby ensuring greater consistency and comparability.

Implications for efficiency and responsibility can be drawn directly from the taxonomy. For efficiency, the evidence base becomes more actionable when studies report class-specific error patterns and explicitly relate metric choices to intervention capacity, because early warning systems operate under constrained mentoring and support resources. For responsibility, the evidence base becomes more defensible when studies transparently report target definitions, validation designs, and minority class performance, because these elements determine whether results can be trusted for educational decision support and whether findings are reproducible and accountable as scientific contributions. Accordingly, this review emphasizes that efficiency-oriented deployment and responsibility-oriented research practice depend on consistent reporting and evaluation choices, not solely on algorithm selection. This link clarifies how study design decisions translate into both operational value and scientific accountability. A further implication of using a single database is that the descriptive distributions reported in this review, such as the relative prevalence of specific algorithms, imbalance-handling strategies, or validation designs, may be sensitive to index coverage. For example, education-oriented journals and applied analytics outlets may be indexed differently from engineering and computing venues, which could shift the observed proportions of methods even when the substantive methodological issues remain similar. Importantly, the main conclusions of this review emphasize study design transparency, leakage-free evaluation, and reporting on minority classes rather than absolute performance rankings. Therefore, while the inclusion of additional databases may alter some frequency-based summaries, the central recommendations on auditability and defensible evaluation are expected to remain applicable. Nevertheless, future work should replicate the protocol across multiple sources, such as Web of Science and discipline-specific libraries, and compare overlaps to assess the robustness of the observed patterns.

This study has three main limitations. First, the search relies on a single database, Scopus, which may introduce index coverage bias (Mongeon and Paul-Hus, 2016; Baas et al., 2020). Scopus was selected as a practical proxy for high-quality peer-reviewed journal literature because it provides broad multidisciplinary journal indexing with consistent metadata, enabling a reproducible and auditable protocol. However, relevant studies may still appear in other curated indexes or discipline-specific libraries, and their inclusion could shift some descriptive distributions, such as the relative frequencies of algorithms or validation designs. The main conclusions of this review are primarily methodological and focus on transparency, leakage-free evaluation, and minority class reporting, so they are less dependent on the exact distribution of methods. However, future work should extend retrieval to additional sources, such as the Web of Science and discipline-specific libraries, to quantify overlap and test robustness. Second, the protocol restricts the corpus to open-access journal articles with DOIs and English language, which improves verifiability but may omit relevant evidence that is not open-access, not written in English, or disseminated in alternative publication formats. Third, the dominance of private datasets limits exact replication and cross-institutional comparison, as target definitions, feature construction, and academic policies are often context-specific. To mitigate this limitation while respecting privacy constraints, future work should standardize feature structures through shared feature schemas and mapping documentation, so that models and evaluation protocols can be tested across institutions even when the underlying datasets cannot be released (Rodríguez-Muñiz et al., 2019; Palacios et al., 2021; Hooper, Ragland, and Artemiou, 2025).

The implied future research agenda is to strengthen evaluation and reporting standards for predicting study status, especially in imbalanced multi-class scenarios. Priorities include reporting metrics per-class and stable summaries such as the Matthews correlation coefficient or Cohen's kappa, affirming validation procedures and the position of resampling in the pipeline, and using more conservative validation when intensive hyperparameter tuning is performed, including nested cross-validation, temporal validation, or external validation on different cohorts (Niyogisubizo et al., 2022; Martins et al., 2023; Phan, De Caigny, and Coussement, 2023; Song et al., 2023; de la Cruz Huayanay, Bazán, and Russo, 2024).

CONCLUSIONS

This study synthesizes 70 studies on predicting student academic status in higher education using machine learning and deep learning approaches from 2017 to 2025, with an emphasis on addressing class imbalance and ensuring defensible evaluation. The findings indicate that the contribution of research in this field cannot be assessed solely on the basis of algorithm selection or aggregate performance metrics. Its reliability and usability are more determined by the study design decisions, particularly the formulation of the target, the degree of imbalance, the strategies for handling imbalance, and the evaluation and validation design. Binary formulation dominance still stands out, while multi-class studies are relatively limited but closer to the reality of institutional study status and tend to reveal clearer performance gaps between classes. At the same time, variations and incompleteness in reporting treatment imbalances, metrics, and validation procedures limit cross-study comparability and reduce confidence in cross-institutional generalizability.

Within the framework of efficiency and responsibility in education and science, the practical value of a prediction system depends on the transparency of error trade-offs, the reporting of metrics representing minority classes, and validation aligned with operational scenarios, because these elements determine whether the evidence supports capacity-aware interventions and whether the results remain defensible, reproducible, and accountable. Further research is needed to strengthen minimum reporting standards, correctly place resampling within the training and validation process, and expand class-wise metric reporting in multi-class scenarios. More conservative validation designs, including nested, temporal, or external validation when possible, are also needed to support stable implementation across cohorts.

REFERENCES

Alvarado-Uribe, J., Mejía-Almada, P., Masetto Herrera, A. L., Molontay, R., Hilliger, I., Hegde, V., Montemayor Gallegos, J. E., Ramírez Díaz, R. A. and Ceballos, H. G. (2022) 'Student dataset from Tecnologico de Monterrey in Mexico to predict dropout in higher education', *Data*, Vol. 7, No. 9, p. 119. https://doi.org/10.3390/data7090119

Alvarez, N. L., Callejas, Z. and Griol, D. (2020) 'Predicting computer engineering students' dropout in Cuban higher education with pre-enrollment and early performance data', *Journal of Technology and Science Education*, Vol. 10, No. 2, pp. 241–258. https://doi.org/10.3926/jotse.922

Anagnostopoulos, T., Papakyriakopoulos, D., Psaromiligkos, Y. and Retalis, S. (2024) 'Exploiting LSTM neural network algorithm potentiality for early identification of delayed graduation in higher education', *WSEAS Transactions on Information Science and Applications*, Vol. 21, pp. 524–532. https://doi.org/10.37394/23209.2024.21.48

Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodriguez, W., Susanibar-Ramirez, E., Toro-Dextre, E., Ausejo-Sanchez, J., Villarreal-Torres, H. and Angeles-Morales, J. (2023) 'Predicting student dropout based on machine learning and deep learning: A systematic review', *EAI Endorsed Transactions on Scalable Information Systems*, Vol. 10, No. 5, pp. 1–11. https://doi.org/10.4108/eetsis.3586

Anggrawan, A., Hairani, H. and Satria, C. (2023) 'Improving SVM classification performance on unbalanced student graduation time data using SMOTE', *International Journal of Information and Education Technology*, Vol. 13, No. 2, pp. 289–295. https://doi.org/10.18178/ijiet.2023.13.2.1806

Arumugam, S., Vinodhini, G. and Chandrasekaran, R. M. (2018) 'Predicting students' academic performance in the university using meta decision tree classifiers', *Journal of Computer Science*, Vol. 14, No. 5, pp. 654–662. https://doi.org/10.3844/jcssp.2018.654.662

Baas, J., Schotten, M., Plume, A., Côté, G. and Karimi, R. (2020) 'Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies', *Quantitative Science Studies*, Vol. 1, No. 1, pp. 377–386. https://doi.org/10.1162/qss_a_00019

Barramuño, M., Meza-Narváez, C. and Gálvez-García, G. (2022) 'Prediction of student attrition risk using machine learning', *Journal of Applied Research in Higher Education*, Vol. 14, No. 3, pp. 974–986. https://doi.org/10.1108/JARHE-02-2021-0073

Bedregal-Alpaca, N., Cornejo-Aparicio, V., Zárate-Valderrama, J. and Yanque-Churo, P. (2020) 'Classification models for determining types of academic risk and predicting dropout in university students', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 1, pp. 266–272. https://doi.org/10.14569/IJACSA.2020.0110133

Budiman, F., Saputro, I. A., Purwanto, P. and Andono, P. N. (2022) 'Optimization of classification results by minimizing class imbalance on decision tree algorithm', in: *International Seminar on Machine Learning, Optimization, and Data Science (ISMODE 2021)*, pp. 6–11. https://doi.org/10.1109/ISMODE53584.2022.9743062

Cañete-Sifuentes, L., Robles, V., Menasalvas, E. and Monroy, R. (2023) 'Comparing automated machine learning against an off-the-shelf pattern-based classifier in a class imbalance problem: Predicting university dropout', *IEEE Access*, Vol. 11, pp. 139147–139156. https://doi.org/10.1109/ACCESS.2023.3336596

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T. and Paganoni, A. M. (2022) 'Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques', *Studies in Higher Education*, Vol. 47, No. 9, pp. 1935–1956. https://doi.org/10.1080/03075079.2021.2018415

Canto, N. G., De Oliveira, M. A. and De Mattos Veroneze, G. (2022) 'Supervised learning applied to graduation forecast of industrial engineering students', *European Journal of Educational Research*, Vol. 11, No. 1, pp. 325–337. https://doi.org/10.12973/eu-jer.11.1.325

Cho, C. H., Yu, Y. W. and Kim, H. G. (2023) 'A study on dropout prediction for university students using machine learning', *Applied Sciences*, Vol. 13, No. 21, p. 12004. https://doi.org/10.3390/app132112004

Csalódi, R. and Abonyi, J. (2021) 'Integrated survival analysis and frequent pattern mining for course failure-based prediction of student dropout', *Mathematics*, Vol. 9, No. 5, p. 463. https://doi.org/10.3390/math9050463

Cuizon, J. C. (2021) 'Ensemble predictive model for academic churn risk using plurality voting', Mindanao *Journal of Science and Technology*, Vol. 19, No. 1, pp. 224–235. https://doi.org/10.61310/mndjsteect.1028.21

Darenoh, N. V., Bachtiar, F. A. and Perdana, R. S. (2024) 'Prediction of on-time student graduation with deep learning method', *Journal of ICT Research and Applications*, Vol. 18, No. 1, pp. 1–20. https://doi.org/10.5614/itbj.ict.res.appl.2023.18.1.1

Daza, A., Guerra, C., Cervera, N. and Burgos, E. (2022) 'Predicting academic performance through data mining: A systematic literature review', *TEM Journal*, Vol. 11, No. 2, pp. 939–949. https://doi.org/10.18421/TEM112-57

Delen, D., Davazdahemami, B. and Rasouli Dezfouli, E. (2024) 'Predicting and mitigating freshmen student attrition: A local-explainable machine learning framework', *Information Systems Frontiers*, Vol. 26, No. 2, pp. 641–662. https://doi.org/10.1007/s10796-023-10397-3

Deleña, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T. and Gulam, S. B. (2025) 'Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors', *Systems and Soft Computing*, Vol. 7, p. 200352. https://doi.org/10.1016/j.sasc.2025.200352

de la Cruz Huayanay, A., Bazán, J. L. and Russo, C. M. (2024) 'Performance of evaluation metrics for classification in imbalanced data', *Computational Statistics*, Vol. 39, No. 3, pp. 1447–1473. https://doi.org/10.1007/s00180-024-01539-5

Delogu, M., Lagravinese, R., Paolini, D. and Resce, G. (2024) 'Predicting dropout from higher education: Evidence from Italy', *Economic Modelling*, Vol. 130, p. 106583. https://doi.org/10.1016/j.econmod.2023.106583

Febro, J. D. (2019) 'Utilizing feature selection in identifying predicting factors of student retention', *International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 9, pp. 269–274. https://doi.org/10.14569/IJACSA.2019.0100934

Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M. and Sanchez-Figueroa, F. (2021) 'A real-life machine learning experience for predicting university dropout at different stages using academic data', *IEEE Access*, Vol. 9, pp. 133076–133090. https://doi.org/10.1109/ACCESS.2021.3115851

Fontana, L., Masci, C., Ieva, F. and Paganoni, A. M. (2021) 'Performing learning analytics via generalised mixed-effects trees', *Data*, Vol. 6, No. 7, p. 74. https://doi.org/10.3390/data6070074

Freitas, F. A. D. S., Vasconcelos, F. F. X., Peixoto, S. A., Hassan, M. M., Ali Akber Dewan, M., de Albuquerque, V. H. C. and Rebouças Filho, P. P. (2020) 'IoT system for school dropout prediction using machine learning techniques based on socioeconomic data', *Electronics*, Vol. 9, No. 10, p. 1613. https://doi.org/10.3390/electronics9101613

Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V. and García-Castelán, R. M. G. (2023) 'Predictive analytics study to determine undergraduate students at risk of dropout', *Frontiers in Education*, Vol. 8, p. 1244686. https://doi.org/10.3389/feduc.2023.1244686

Goran, R., Jovanovic, L., Bacanin, N., Stanković, M. S., Simic, V., Antonijevic, M. and Zivkovic, M. (2024) 'Identifying and understanding student dropouts using metaheuristic optimized classifiers and explainable artificial intelligence techniques', *IEEE Access*, Vol. 12, pp. 122377–122400. https://doi.org/10.1109/ACCESS.2024.3446653

Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G. and Gomez-Nieto, E. (2023) 'Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the Latin American context', *Education Sciences*, Vol. 13, No. 2, p. 154. https://doi.org/10.3390/educsci13020154

Haerani, E., Syafria, F., Lestari, F., Novriyanto, N. and Marzuki, I. (2023) 'Classification academic data using machine learning for decision making process', *Journal of Applied Engineering and Technological Science (JAETS)*, Vol. 4, No. 2, pp. 955–968. https://doi.org/10.37385/jaets.v4i2.1983

Hammoodi, M. S. and Al-Azawei, A. (2022) 'Using socio-demographic information in predicting students' degree completion based on a dynamic model', *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 2, pp. 107–115. https://doi.org/10.22266/ijies2022.0430.11

Hammoudi Halat, D., Abdel-Salam, A.-S. G., Bensaid, A., Soltani, A., Alsarraj, L., Dalli, R. and Malki, A. (2023) 'Use of machine learning to assess factors affecting progression, retention, and graduation in first-year health professions students in Qatar: A longitudinal study', *BMC Medical Education*, Vol. 23, No. 1, p. 909. https://doi.org/10.1186/s12909-023-04887-w

Helbach, J., Pieper, D., Mathes, T., Rombey, T., Zeeb, H., Allers, K. and Hoffmann, F. (2022) 'Restrictions and their reporting in systematic reviews of effectiveness: An observational study', *BMC Medical Research Methodology*, Vol. 22, No. 1, p. 230. https://doi.org/10.1186/s12874-022-01710-w

Herianto, H., Kurniawan, B., Hartomi, Z. H., Irawan, Y. and Anam, M. K. (2024) 'Machine learning algorithm optimization using stacking technique for graduation prediction', *Journal of Applied Data Sciences*, Vol. 5, No. 3, pp. 1272–1285. https://doi.org/10.47738/jads.v5i3.316

Hooper, S. E., Ragland, N. and Artemiou, E. (2025) 'Random forest models reveal academic and financial factors outweigh demographics in predicting completion of a year-round veterinary program', *Journal of the American Veterinary Medical Association*, Vol. 263, No. 2, pp. 1–9. https://doi.org/10.2460/javma.24.08.0501

Hoyos Osorio, J. K. and Daza Santacoloma, G. (2023) 'Predictive model to identify college students with high dropout rates', *Revista Electrónica de Investigación Educativa*, Vol. 25, pp. 1–10. https://doi.org/10.24320/redie.2023.25.e13.5398

Kaensar, C. and Wongnin, W. (2023) 'Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning', *Eurasia Journal of Mathematics, Science and Technology Education*, Vol. 19, No. 12, p. 2369. https://doi.org/10.29333/ejmste/13863

Kim, S., Choi, E., Jun, Y.-K. and Lee, S. (2023) 'Student dropout prediction for university with high precision and recall', *Applied Sciences*, Vol. 13, No. 10, p. 6275. https://doi.org/10.3390/app13106275

Kitchenham, B. (2004) *Procedures for performing systematic reviews*, Keele: Keele University, pp. 1–26.

Kurniadi, D., Abdurachman, E., Warnars, H. L. H. S. and Suparta, W. (2021) 'Predicting student performance with multi-level representation in an intelligent academic recommender system using backpropagation neural network', *ICIC Express Letters, Part B: Applications*, Vol. 12, No. 10, pp. 883–890. https://doi.org/10.24507/icicelb.12.10.883

Luque, A., Carrasco, A., Martín, A. and de las Heras, A. (2019) 'The impact of class imbalance in classification performance metrics based on the binary confusion matrix', *Pattern Recognition*, Vol. 91, pp. 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Martins, M. V., Baptista, L., Machado, J. and Realinho, V. (2023) 'Multi-class phased prediction of academic performance and dropout in higher education', *Applied Sciences*, Vol. 13, No. 8, p. 4702. https://doi.org/10.3390/app13084702

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T. and Realinho, V. (2021) 'Early prediction of student's performance in higher education: A case study', in: Rocha, Á., Adeli, H., Reis, L. P. and Costanzo, S. (eds.), *Trends and Applications in Information Systems and Technologies*, Cham: Springer, pp. 166–175. https://doi.org/10.1007/978-3-030-72657-7_16

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A. and Stachl, C. (2023) 'Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics', *Scientific Reports*, Vol. 13, No. 1, p. 5705. https://doi.org/10.1038/s41598-023-32484-w

Mongeon, P. and Paul-Hus, A. (2016) 'The journal coverage of Web of Science and Scopus: A comparative analysis', *Scientometrics*, Vol. 106, No. 1, pp. 213–228. https://doi.org/10.1007/s11192-015-1765-5

Moreira da Silva, D. E., Solteiro Pires, E. J., Reis, A., de Moura Oliveira, P. B. and Barroso, J. (2022) 'Forecasting students dropout: A UTAD university study', *Future Internet*, Vol. 14, No. 3, p. 76. https://doi.org/10.3390/fi14030076

Mouchantaf, N. and Chamoun, M. (2023) 'Predicting student dropout with minimal information', *Iraqi Journal of Science*, Vol. 64, No. 10, pp. 5265–5279. https://doi.org/10.24996/ijs.2023.64.10.33

Nagy, M. and Molontay, R. (2024) 'Interpretable dropout prediction: Towards XAI-based personalized intervention', *International Journal of Artificial Intelligence in Education,* Vol. 34, No. 2, pp. 274–300. https://doi.org/10.1007/s40593-023-00331-8

Nanglae, L., Iam-On, N., Boongoen, T., Kaewchay, K. and Mullaney, J. (2021) 'Determining patterns of student graduation using a bi-level learning framework', *Bulletin of Electrical Engineering and Informatics*, Vol. 10, No. 4, pp. 2201–2211. https://doi.org/10.11591/eei.v10i4.2502

Ndunagu, J. N., Oyewola, D. O., Garki, F. S., Onyeakazi, J. C., Ezeanya, C. U. and Ukwandu, E. (2024) 'Deep learning for predicting attrition rate in open and distance learning (ODL) institutions', *Computers*, Vol. 13, No. 9, p. 229. https://doi.org/10.3390/computers13090229

Nguyen Thi Cam, H., Sarlan, A. and Arshad, N. I. (2024) 'A hybrid model integrating recurrent neural networks and the semi-supervised support vector machine for identification of early student dropout risk', *PeerJ Computer Science*, Vol. 10, p. e2572. https://doi.org/10.7717/peerj-cs.2572

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E. and Nshimyumukiza, P. C. (2022) 'Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization', *Computers and Education: Artificial Intelligence*, Vol. 3, p. 100066. https://doi.org/10.1016/j.caeai.2022.100066

Nuanmeesri, S., Poomhiran, L., Chopvitayakun, S. and Kadmateekarun, P. (2022) 'Improving dropout forecasting during the COVID-19 pandemic through feature selection and multilayer perceptron neural network', *International Journal of Information and Education Technology*, Vol. 12, No. 9, pp. 851–857. https://doi.org/10.18178/ijiet.2022.12.9.1693

Okewu, E., Adewole, P., Misra, S., Maskeliunas, R. and Damasevicius, R. (2021) 'Artificial neural networks for educational data mining in higher education: A systematic literature review', *Applied Artificial Intelligence*, Vol. 35, No. 13, pp. 983–1021. https://doi.org/10.1080/08839514.2021.1922847

Okoye, K., Nganji, J. T., Escamilla, J. and Hosseini, S. (2024) 'Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education', *Computers and Education: Artificial Intelligence*, Vol. 6, p. 100205. https://doi.org/10.1016/j.caeai.2024.100205

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J. and Moreira, F. (2021) 'How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review', *Big Data and Cognitive Computing*, Vol. 5, No. 4, p. 64. https://doi.org/10.3390/bdcc5040064

Opazo, D., Moreno, S., Álvarez-Miranda, E. and Pereira, J. (2021) 'Analysis of first-year university student dropout through machine learning models: A comparison between universities', *Mathematics*, Vol. 9, No. 20, p. 2599. https://doi.org/10.3390/math9202599

Oqaidi, K., Aouhassi, S. and Mansouri, K. (2025) 'Predicting graduation in Moroccan open-access bachelors: Early indicators and re-enrollment data', *Bulletin of Electrical Engineering and Informatics*, Vol. 14, No. 1, pp. 524–532. https://doi.org/10.11591/eei.v14i1.8580

Ortigosa, A., Carro, R. M., Bravo-Agapito, J., Lizcano, D., Alcolea, J. J. and Blanco, Ó. (2019) 'From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system', *IEEE Transactions on Learning Technologies*, Vol. 12, No. 2, pp. 264–277. https://doi.org/10.1109/TLT.2019.2911608

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P. and Moher, D. (2021) 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', *BMJ*, p. n71. https://doi.org/10.1136/bmj.n71

Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V. and Marchant, C. (2021) 'Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile', *Entropy*, Vol. 23, No. 4, p. 485. https://doi.org/10.3390/e23040485

Pelima, L. R., Sukmana, Y. and Rosmansyah, Y. (2024) 'Predicting university student graduation using academic performance and machine learning: A systematic literature review', *IEEE Access*, Vol. 12, pp. 23451–23465. https://doi.org/10.1109/ACCESS.2024.3361479

Phan, M., De Caigny, A. and Coussement, K. (2023) 'A decision support framework to incorporate textual data for early student dropout prediction in higher education', *Decision Support Systems*, Vol. 168, p. 113940. https://doi.org/10.1016/j.dss.2023.113940

Quimiz-Moreira, M., Delgadillo, R., Parraga-Alava, J., Maculan, N. and Mauricio, D. (2025) 'Factors, prediction, explainability, and simulating university dropout through machine learning: A systematic review, 2012–2024', *Computation*, Vol. 13, No. 8, p. 198. https://doi.org/10.3390/computation13080198

Rabelo, A. M. and Zárate, L. E. (2025) 'A model for predicting dropout of higher education students', *Data Science and Management*, Vol. 8, No. 1, pp. 72–85. https://doi.org/10.1016/j.dsm.2024.07.001

Realinho, V., Martins, M. V., Machado, J. and Baptista, L. (2021) Predict students' dropout and academic success [Dataset], UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., Blunt, H., Brigham, T., Chang, S., Clark, J., Conway, A., Couban, R., de Kock, S., Farrah, K., Fehrmann, P., Foster, M., Fowler, S. A., Glanville, J., Harris, E., Hoffecker, L., Isojarvi, J., Kaunelis, D., Ket, H., Levay, P., Lyon, J., McGowan, J., Murad, M. H., Nicholson, J., Pannabecker, V., Paynter, R., Pinotti, R., Ross-White, A., Sampson, M., Shields, T., Stevens, A., Sutton, A., Weinfurter, E., Wright, K. and Young, S. (2021) 'PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews', *Systematic Reviews*, Vol. 10, No. 1, p. 39. https://doi.org/10.1186/s13643-020-01542-z

Rodríguez-Muñiz, L. J., Bernardo, A. B., Esteban, M. and Díaz, I. (2019) 'Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques?', *PLoS ONE*, Vol. 14, No. 6, p. e0218796. https://doi.org/10.1371/journal.pone.0218796

Rose, A. L. P. J. and Mary, A. C. (2022) 'An early intervention technique for at-risk prediction of higher education students in cloud-based virtual learning environment using classification algorithms during COVID-19', *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 1, pp. 612–621. https://doi.org/10.14569/IJACSA.2022.0130174

Roslan, N., Jamil, J. M., Shaharanee, I. N. M. and Sultan Alawi, S. J. (2024) 'Prediction of student dropout in Malaysian's private higher education institute using data mining application', *Journal of Advanced Research in Applied Sciences and Engineering Technology*, Vol. 45, No. 2, pp. 168–176. https://doi.org/10.37934/araset.45.2.168176

Rovira, S., Puertas, E. and Igual, L. (2017) 'Data-driven system to predict academic grades and dropout', *PLoS ONE*, Vol. 12, No. 2, p. e0171207. https://doi.org/10.1371/journal.pone.0171207

Salam, A. and Zeniarja, J. (2023) 'Classification of deep learning convolutional neural network feature extraction for student graduation prediction', *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 32, No. 1, p. 335. https://doi.org/10.11591/ijeecs.v32.i1.pp335-341

Salinas-Chipana, J., Obregon-Palomino, L., Iparraguirre-Villanueva, O. and Cabanillas-Carbonell, M. (2024) 'Machine learning models for predicting student dropout—a review', in: Yang, X.-S., Sherratt, R. S., Dey, N. and Joshi, A. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Singapore: Springer Nature Singapore, pp. 1003–1014. https://doi.org/10.1007/978-981-99-3043-2_83

Sandoval-Palis, I., Naranjo, D., Vidal, J. and Gilar-Corbi, R. (2020) 'Early dropout prediction model: A case study of university leveling course students', *Sustainability*, Vol. 12, No. 22, p. 9314. https://doi.org/10.3390/su12229314

Sani, N. S., Fikri, A., Ali, Z., Zakree, M. and Nadiyah, K. (2020) 'Drop-out prediction in higher education among B40 students', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 11, pp. 550–559. https://doi.org/10.14569/IJACSA.2020.0111169

Sayed, M. (2024) 'Student progression and dropout rates using convolutional neural network: A case study of the Arab Open University', *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 28, No. 3, pp. 668–678. https://doi.org/10.20965/jaciii.2024.p0668

Segura, M., Mello, J. and Hernández, A. (2022) 'Machine learning prediction of university student dropout: Does preference play a key role?', *Mathematics*, Vol. 10, No. 18, p. 3359. https://doi.org/10.3390/math10183359

Setiadi, H., Sanjaya, K., Wijayanto, A., Wardhani, D. W. and Cahyono, H. D. (2024) 'Comparative analysis of classification algorithms using feature selection techniques to predict on-time student graduation', *Ingénierie des systèmes d'information*, Vol. 29, No. 4, pp. 1365–1379. https://doi.org/10.18280/isi.290412

Setiawan, R., Nursasongko, E., Syukur, A., Budiman, F. and Kurniadi, D. (2025) 'Handling class imbalance in student success prediction using machine learning: A comparison of SMOTE and SMOTETomek', in: *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, pp. 1–6. https://doi.org/10.1109/SIML65326.2025.11081128

Song, Z., Sung, S.-H., Park, D.-M. and Park, B.-K. (2023) 'All-year dropout prediction modeling and analysis for university students', Applied Sciences, Vol. 13, No. 2, p. 1143. https://doi.org/10.3390/app13021143

Tsai, S.-C., Chen, C.-H., Shiao, Y.-T., Ciou, J.-S. and Wu, T.-N. (2020) 'Precision education with statistical learning and deep learning: A case study in Taiwan', International Journal of Educational Technology in Higher Education, Vol. 17, No. 1, p. 12. https://doi.org/10.1186/s41239-020-00186-2

Uliyan, D., Aljaloud, A. S., Alkhalil, A., Amer, H. S. Al, Mohamed, M. A. E. A. and Alogali, A. F. M. (2021) 'Deep learning model to predict students retention using BLSTM and CRF', *IEEE Access*, Vol. 9, pp. 135550–135558. https://doi.org/10.1109/ACCESS.2021.3117117

Vaarma, M. and Li, H. (2024) 'Predicting student dropouts with machine learning: An empirical study in Finnish higher education', *Technology in Society*, Vol. 76, p. 102474. https://doi.org/10.1016/j.techsoc.2024.102474

Vega, H., Sanez, E., De La Cruz, P., Moquillaza, S. and Pretell, J. (2022) 'Intelligent system to predict university students dropout', International Journal of Online and Biomedical Engineering (iJOE), Vol. 18, No. 7, pp. 27–43. https://doi.org/10.3991/ijoe.v18i07.30195

Véliz Palomino, J. C. and Ortega, A. M. (2023) 'Dropout intentions in higher education: Systematic literature review', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 16, No. 2, pp. 149–158. https://doi.org/10.7160/eriesj.2023.160206

Vidal, J., Gilar-Corbi, R., Pozo-Rico, T., Castejón, J.-L. and Sánchez-Almeida, T. (2022) 'Predictors of university attrition: Looking for an equitable and sustainable higher education', *Sustainability*, Vol. 14, No. 17, p. 10994. https://doi.org/10.3390/su141710994

Villar, A. and de Andrade, C. R. V. (2024) 'Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study', *Discover Artificial Intelligence*, Vol. 4, No. 1, p. 2. https://doi.org/10.1007/s44163-023-00079-z

Villegas-Ch, W., Govea, J. and Revelo-Tapia, S. (2023) 'Improving student retention in institutions of higher education through machine learning: A sustainable approach', *Sustainability*, Vol. 15, No. 19, p. 14512. https://doi.org/10.3390/su151914512

Won, H.-S., Kim, M.-J., Kim, D., Kim, H.-S. and Kim, K.-M. (2023) 'University student dropout prediction using pretrained language models', *Applied Sciences,* Vol. 13, No. 12, p. 7073. https://doi.org/10.3390/app13127073

Yaqin, A., Laksito, A. D. and Fatonah, S. (2021) 'Evaluation of backpropagation neural network models for early prediction of student's graduation in XYZ University', *International Journal on Advanced Science, Engineering and Information Technology*, Vol. 11, No. 2, pp. 610–617. https://doi.org/10.18517/ijaseit.11.2.11152

Yaqin, A., Rahardi, M. and Abdulloh, F. F. (2022) 'Accuracy enhancement of prediction method using SMOTE for early prediction student's graduation in XYZ University', *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 6, pp. 418–424. https://doi.org/10.14569/IJACSA.2022.0130652

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