ISSN: 3139-1141

Stage Aware Student Dropout Prediction in Higher Education

A Comparative Synthesis of Machine Learning and Statistical Approaches

Maseeha Banu[#1], Sinchana C K[*2], Manya S R[#3], S B Sharath Gowda[#], Anjali Sharma[#]

[#]*Department of CSE (AI&ML), Malnad College of Engineering, Hassan, India*

[1]mb@mcehassan.ac,[2]sinchanackay12@gmail.com,[3]msr610742@gmail.com

***Abstract*—Student dropout remains a persistent challenge for higher education institutions, and predictive analytics promises earlier and better targeted support. Published studies, however, differ widely in what they call dropout, when they predict it, which data they use and how they evaluate their models. This paper presents a comparative synthesis of four recent studies that span the main design choices in the field: a class level platform built on attendance data at the University of Coimbra; a multi stage cascade of models built on academic data only for Spanish engineering students; a large scale comparison of random forest and a feature tokenizer transformer, with explainability and fairness analysis, at an Italian university; and a four step logistic regression study of dropout intention among student teachers in Ethiopia. We compare their dropout definitions, prediction timing, data, imbalance handling, algorithms, metrics and explainability practices. The synthesis shows that predictive power rises sharply once post enrolment academic or engagement data become available, that tree based and ensemble models remain competitive on tabular institutional data, that recall oriented evaluation is often reported without calibrated precision, and that results are rarely comparable across studies. We propose a stage aware, human in the loop framework and a minimum reporting checklist to improve the comparability, fairness and practical usefulness of future dropout prediction research.**

Keywords—student dropout; early warning systems; educational data mining; machine learning; explainable AI; higher education

I. INTRODUCTION

Dropout from higher education is a long-standing concern. Tinto's student integration model framed attrition as the outcome of a student's academic and social integration into an institution [1], and later work has added individual, economic and institutional predictors. Universities now hold large volumes of administrative and learning data, which makes educational data mining a natural route to identifying at-risk students early enough for intervention.

The literature is nevertheless fragmented. Some studies predict whether a student will leave a degree programme, others whether a student will abandon a single course, and others model the intention to drop out measured through questionnaires. Some use only data available at enrolment; others use grades, attendance or virtual-learning-environment logs. Models range from logistic regression to gradient boosting and transformers. Practitioners therefore find it hard to know which approach suits their context, and researchers find it hard to compare results.

This paper addresses that gap through a comparative synthesis of four studies [2]– [5], each representing a distinct design philosophy. The contributions are: (i) a side-by-side comparison of dropout definitions, prediction stages, data, preprocessing, models and evaluation; (ii) a synthesis of which predictors and prediction stages matter most; (iii) a critical discussion of methodological risks, including label ambiguity, imbalance handling, metric choice and fairness; and (iv) a stage-aware, human-in-the-loop framework with a minimum reporting checklist. Section II gives background, Section III the synthesis method, Section IV the four studies, Section V the comparative analysis, Section VI the proposed framework, and Section VII concludes with limitations and future work.

II. RELATED WORK AND BACKGROUND

There is no single definition of dropout. Programme-level dropout, in which a student leaves a degree without completing it, is the classical definition associated with Tinto [1]. Course-level dropout is operationalised differently, for example as missing the final assessments of a subject, and intention-based definitions measure whether a student reports thinking about leaving. Each definition supports a different decision: programme-level labels inform institutional retention policy, course-level labels inform instructors during a semester, and intention scores inform counselling.

Prediction can be made at admission or enrolment, at the end of a semester or year, or at milestones within a class. Earlier prediction leaves more time for intervention but uses less information; later prediction is more accurate but may arrive when the student has already disengaged.

Logistic regression, decision trees and naive Bayes remain common baselines. Ensemble methods such as random forests [6] and gradient boosting [7] are widely used on tabular data, and attention-based architectures for tabular data, such as the feature tokenizer transformer [8], have recently been explored. Because predictions may affect individual students, post hoc explanation methods such as permutation importance and Shapley additive explanations (SHAP) [9] are increasingly reported, together with fairness analysis. Dropout is a minority class, so oversampling methods such as the synthetic minority over-sampling technique (SMOTE) [10] are frequently applied.

III. METHODOLOGY OF THE SYNTHESIS

This is a qualitative, comparative synthesis rather than a systematic review or a new experiment. It is guided by three research questions. **RQ1:** how do dropout definitions and

prediction timing differ across studies, and how do they shape what the models can be used for? **RQ2:** which predictors and data sources contribute most at different stages? **RQ3:** which modelling and evaluation practices are shared, and where do they limit comparability, fairness or deployment?

Four studies were selected to maximise diversity in institutional context (Portugal, Spain, Italy, Ethiopia), unit of prediction (class, programme, student intention), data type (attendance, grades and admission data, demographics and questionnaires) and modelling approach (classical machine learning, ensembles, deep tabular models, logistic regression). For each study we extracted the dropout definition, sample, prediction timing, features, preprocessing and imbalance handling, algorithms, validation scheme, metrics, headline results, and explainability or fairness analysis. All figures below are taken from the original publications. Because datasets, labels and validation designs differ, the numerical results are not directly comparable and are reported for context only.

IV. The Four Studies

A. *Borges et al.: Class-Level Prediction and a Deployed Platform*

Using anonymised University of Coimbra records from 2000/2001 onwards (excluding COVID-affected years), Borges et al. [2] predict subject-level dropout, defined as failing to attend any scheduled exam or final assessment (Approach A), with a stricter variant for non-working students absent from more than 90% of classes who also missed assessments (Approach B). A baseline dataset without attendance information has 1,612,651 records with 15.45% positives. Attendance-based datasets contain 327,504 entries per milestone, with milestones at the start, midpoint and end of the lecture period, and dropout prevalence of 12.95% (Approach A) and 4.76% (Approach B). Naive Bayes (NB), decision tree (DT), k-nearest neighbours (KNN), random forest (RF) and XGBoost were tuned by grid search with five-fold cross-validation, prioritising recall, and SMOTE was applied to training data only.

Results improve as attendance accumulates. Statistical tests show significant gains from the start to the midpoint of the class period, and no significant gain from midpoint to end under the stricter labelling. At the later milestones decision trees and XGBoost both reach recall above 0.99 on Approach B datasets, and XGBoost achieves the highest area under the receiver operating characteristic curve (AUC-ROC), 0.947 at the midpoint and 0.955 at the end. Attendance ratio and re-enrolment in the same subject are the most important features; in the baseline dataset, re-enrolled students dropped out at 32.34% against 12.40% for first-time enrollees. The model is integrated into the BATINA OnBoard dashboard, where instructors see risk scores with a default 66% threshold under General Data Protection Regulation (GDPR) oriented safeguards and a human-in-the-loop design. The authors acknowledge reliance on attendance data, false positives from recall prioritisation, potential structural bias and single-institution generalisability.

B. *Fernández-García et al.: A Cascade of Stage-Specific Models*

Fernández-García et al. [3] deliberately restrict themselves to academic data available to any university (access records, grades, scholarship status), excluding personal and learning-management data to ease reproducibility and avoid privacy complications. Five models are built for engineering students at a Spanish public university: at enrolment and after each of the first four semesters. Feature and instance engineering includes removing redundant, high-cardinality and unavailable-at-prediction-time features, deriving success-rate, not-shown-rate and median-mark features, and removing obsolete degrees and incomplete records. The final dataset has 1,418 observations (783 dropouts, 635 completers), balanced with SMOTE combined with Tomek links, with a 75/25 train-test split. Gradient boosting (GB), RF, support vector machine (SVM) and a weighted ensemble are compared, and each stage's model output is fed as a feature to the next stage.

At enrolment, accuracy is close to 70% and the best dropout recall is 74.79% (ensemble), although the authors prefer gradient boosting (recall 72.34%) for its higher precision. After the first semester the ensemble reaches a recall of 82.24%, and by the fourth semester the best recall is 91.55% (Fig. 1). No single algorithm dominates at every stage.

C. *Zanellati et al.: Random Forest versus a Feature Tokenizer Transformer*

Zanellati et al. [4] study 44,875 first-year students across 110 programmes at a large Italian university (cohorts 2018 to 2021). Dropout is defined as not re-enrolling in the same programme the following year, which includes withdrawal, internal transfer and transfer to another university, with a prevalence of about 15.4%. Features cover personal data, secondary-school background, programme characteristics and academic performance (weighted mark average, honours and credits) measured at 0, 3, 6, 9 and 12 months after enrolment. Models are trained on the 2018–2019 and 2019–2020 cohorts and tested on the 2020–2021 cohort, a temporally honest split.

The best model overall is the feature tokenizer transformer at 12 months, with accuracy 0.87 and sensitivity 0.81, against a majority-class baseline of accuracy 0.84 and sensitivity 0.25. Credits earned dominate random forest explanations under grouped permutation importance (GPI), while attention-based importance for the transformer is much flatter. SHAP local explanations show that both models rely on credits, but that the transformer treats a student with a satisfactory credit count in a high-transfer programme (pharmacy and biotechnology) as high risk, which the authors read as better contextualisation. A fairness analysis stratified by gender, geographical origin and economic status uses demographic parity, equalised odds and predictive parity with bootstrap confidence intervals.

D. *Singh and Alhulail: Explanatory Logistic Regression of Dropout Intention*

Singh and Alhulail [5] survey 1,723 first-year student-teachers in five Ethiopian public teacher training colleges with a 39-item questionnaire (Cronbach's alpha 0.769). The outcome is self-reported dropout intention, a dichotomous variable. Predictors are grouped into personal, academic and socioeconomic dimensions and entered in four steps. Explained variation (Nagelkerke R^2) is 0.149 for personal variables, 0.206 for academic variables, 0.254 for personal plus academic variables, and 0.317 with socioeconomic variables added. Academic performance, post-high-school aspirations, the academic reason for choosing the college and relationships with teachers are the strongest predictors, while cultural values,

parental education and financing costs are not significant; the authors attribute the latter to government financial support and the perceived value of higher education. The study is cross-sectional and self-reported, and it assesses explanatory rather than out-of-sample predictive performance.

V. Comparative Analysis

Tables I–III compare the four studies on design, methods and findings, as reported by the original authors.

TABLE I. Study Design: Context, Label and Timing

Study	Context	Dropout definition	Prediction timing
Borges [2]	Univ. of Coimbra, Portugal	Missing final assessments of a subject	Start, midpoint, end of class period
Fernández-García [3]	Engineering school, Spain	Not completing the degree (includes transfers)	Enrolment and four semester ends
Zanellati [4]	Large Italian university	Not re-enrolling in the same programme after one year	0, 3, 6, 9, 12 months
Singh and Alhulail [5]	Teacher training colleges, Ethiopia	Self-reported intention to drop out	Single point in first year

TABLE II. Methods: Data, Models and Validation

Study	Data and sample	Models and imbalance	Validation
Borges [2]	Enrolment, grades, attendance; 327,504 per milestone	NB, DT, KNN, RF, XGBoost; SMOTE on training data	5-fold cross-validation; Mann-Whitney and Wilcoxon tests
Fernández-García [3]	Access, grades, scholarship; 1,418 after cleaning	GB, RF, SVM, ensemble (cascade); SMOTE + Tomek links	Hold-out 75/25, grid-search cross-validation
Zanellati [4]	Demographics, school, credits, marks; 44,875 students	RF, feature tokenizer transformer; class weights, weighted sampling	Temporal split (train 2018–20, test 2020–21)
Singh and Alhulail [5]	Questionnaire; 1,723 students	Logistic regression; not applicable	Nagelkerke R^2, Wald tests

TABLE III. Findings, Explainability and Deployment

Study	Key predictors	Explainability / fairness	Deployment
Borges [2]	Attendance ratio, re-enrolment	Feature importance; bias discussed, audits planned	BATINA OnBoard dashboard
Fernández-García [3]	Success rate, marks, credits	Limited; fairness not addressed	Proposed decision support
Zanellati [4]	Credits, weighted average mark	GPI, attention, SHAP; three fairness metrics	Not deployed
Singh and Alhulail [5]	Academic performance, aspirations, college choice	Coefficients, odds ratios; fairness not addressed	Not applicable

A. *Definitions and Timing (RQ1)*

The four studies answer different questions. Borges et al. support instructors during a semester, Fernández-García et al. and Zanellati et al. support programme-level retention policy at stages of the student journey, and Singh and Alhulail support counselling of students with elevated intentions. Treating all four as measuring the same thing would be unsafe. Borges et al. partially address this by showing a statistical association between subject-level and programme-level dropout (odds ratio 4.82), but describe their labels as approximations. The definition of Zanellati et al. lumps together withdrawal and transfer, which may call for different interventions, as their pharmacy case illustrates.

All studies that vary timing show the same pattern: performance rises as post-enrolment data accumulate, and the gain flattens near the end of the observation window. Borges et al. find no significant improvement from midterm to end of term under their stricter labelling, and Fernández-García et al. reach 91.55% recall by the fourth semester but note that by then dropout is fairly easy to anticipate. This is an earliness–accuracy trade-off: the most accurate models are the least useful for prevention, because the student may already have left or disengaged.

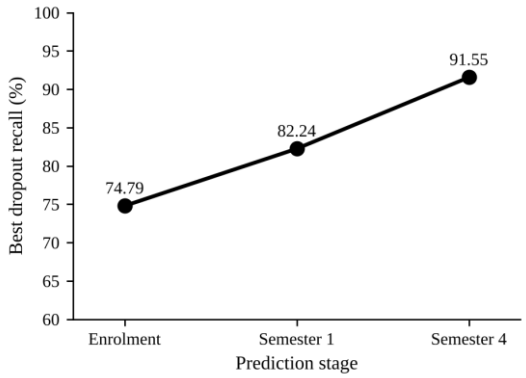

Fig. 1. Best dropout recall by prediction stage, as reported by Fernández-García et al. [3] (reported stages only).

B. *Predictors and Data (RQ2)*

At or before enrolment, predictive power is modest. The baseline dataset of Borges et al. gives a best recall of 0.712 and a best AUC-ROC below 0.7, and Fernández-García et al. report accuracy near 70% at enrolment. Pre-enrolment variables such as demographics and background add limited information, consistent with the weak personal-dimension result of Singh and Alhulail (R^2 0.149).

Once the student is in the institution, behavioural and academic-progress indicators dominate: attendance ratio [2], credits earned and marks [3], [4], and perceived academic performance [5]. Re-enrolment in the same subject is a strong signal of difficulty [2]. Three very different datasets and one survey thus converge on academic progress and engagement as the core signal. By contrast, socioeconomic predictors behave inconsistently across contexts: Singh and Alhulail find parental education and financing insignificant in a setting where education is publicly funded and valued as a route to mobility, whereas other literature they review reports the opposite. Predictor importance is context-dependent and should not be assumed to transfer across institutions or countries.

C. *Modelling, Imbalance and Evaluation (RQ3)*

Algorithms. Tree-based models (decision trees, random forests, XGBoost, gradient boosting) are consistently strong on institutional tabular data, and no study finds a universal winner. Zanellati et al. report a transformer advantage of at least about one percentage point on their metrics, at higher computational cost and on a single dataset. The most defensible conclusion is that model choice matters less than data timing, feature quality and label design.

Imbalance. Dropout is a minority class (about 5% to 15% in the large datasets). Two studies use SMOTE-based oversampling and one uses class weighting and weighted sampling. Borges et al. apply SMOTE only to training sets so that test distributions stay realistic, which is good practice. Fernández-García et al. balance a dataset of about 1,400 observations to 1:1 before a hold-out split, so readers should check that synthetic samples were not generated before splitting, since that would give optimistic estimates.

Metrics. Recall is prioritised in three of the four studies because missing an at-risk student is considered costlier than flagging one unnecessarily. This is reasonable, but recall-optimised models can approach 1.0 by flagging many students, as the very high recall at late milestones in Borges et al. invites readers to check against precision, and instructors then face alert fatigue and possible stigmatisation. The majority-class baseline of Zanellati et al. (accuracy 0.84, sensitivity 0.25) shows why accuracy must be reported with sensitivity and specificity. Precision–recall curves, calibration and threshold analysis are largely absent and should be standard.

Validation. Only Zanellati et al. use a strictly temporal split, which best mimics deployment, where models trained on past cohorts score future students. Random cross-validation can leak information across cohorts and overstate performance when curricula or policies change over time.

Explainability and fairness. Explanation is secondary in two studies and a core contribution in Zanellati et al., where different techniques tell partly different stories (credits dominate under permutation importance but not under attention weights), which argues for triangulating several methods. Fairness analysis is rare: only Zanellati et al. perform a stratified evaluation, and Borges et al. flag bias as an ongoing concern, noting that associations with re-enrolment or socioeconomic indicators may reflect structural inequality. An accurate model that systematically over-flags certain groups can reinforce disadvantage.

Threats to Validity: The synthesis covers four studies and is not a systematic review, so it cannot claim to represent the whole literature. The studies differ in context, labels and validation, so quantitative comparisons are indicative only. We rely on results as reported by the original authors and did not reproduce their experiments, and the studies are concentrated in specific regions and institution types, so findings may not transfer to online, vocational or other settings.

VI. A Stage-Aware, Human-in-the-Loop Framework

Building on the strengths of the four studies, we propose a framework with five principles (Fig. 2). **(1) Align the label with the decision.** Define dropout per use case (course disengagement, programme withdrawal, transfer or intention) and, where possible, separate withdrawal from transfer, since the appropriate response differs. **(2) Use stage-specific models.** Train separate models at defined stages (pre-enrolment, early term, midterm, end of term or year) and, as in the cascade of [3], let later models consume earlier predictions. Report performance per stage so that institutions can choose the earliest stage at which accuracy is operationally acceptable.

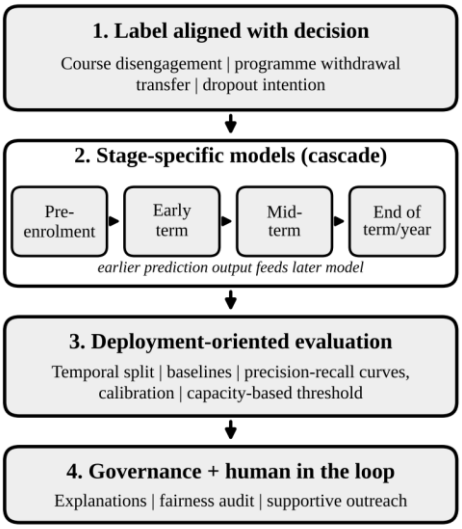

Fig. 2. Proposed stage-aware, human-in-the-loop framework.

(3) Prefer robust, interpretable baselines first. Start with logistic regression and gradient-boosted trees, and add deep tabular models only if they deliver a validated, material gain on temporally held-out data. **(4) Evaluate for deployment.** Use temporal splits, report sensitivity, specificity, precision, precision-recall (PR) curves and calibration, compare against a naive baseline, and select thresholds from intervention capacity, for example how many students advisers can actually contact. **(5) Embed governance.** Obtain consent where required, restrict and log access, use local explanations such as SHAP alongside global importance, run stratified fairness audits, and treat scores as prompts for supportive outreach, never for punitive action. Collect instructor feedback and monitor for feedback loops in which interventions change the very patterns the model learned.

To improve comparability, Table IV lists a minimum reporting checklist for dropout-prediction studies.

TABLE IV. Minimum Reporting Checklist

Item	What to report
Label	Exact dropout definition, time horizon, treatment of transfers
Timing	Prediction point and which data were available at that point
Population	Cohorts, exclusions and prevalence of dropout
Leakage control	Features unavailable at prediction time removed; resampling applied after splitting
Validation	Temporal hold-out and, where relevant, cross-validation details
Baselines	Majority-class and simple model baselines
Metrics	Recall, precision, specificity, AUC-ROC, precision-recall AUC, calibration, confidence intervals
Explainability	At least two complementary methods
Fairness	Stratified metrics for protected or sensitive groups
Deployment	Threshold choice, intervention pathway, human oversight

VII. Conclusion and Future Work

Across four methodologically different studies a consistent picture emerges. Prediction before or at enrolment is possible but modest. Academic progress and engagement indicators such as attendance, credits earned and marks sharply improve predictions, and the improvement flattens late in the observation window, creating a trade-off between earliness and accuracy. Tree-based and ensemble models are reliable choices for tabular institutional data, while deep tabular models are promising but not yet shown to be decisive. The main obstacles to progress are not algorithmic: they are inconsistent dropout definitions, optimistic validation, incomplete metric reporting and limited fairness analysis. Stage-aware modelling, deployment-oriented evaluation and a common reporting checklist would make future work more comparable and more useful to the students and staff it is meant to help.

The main limitations are the small number of studies, the reliance on reported rather than reproduced results, and the absence of any prospective test of the proposed framework. Future work should test the framework prospectively across multiple institutions, evaluate whether interventions triggered by model alerts actually reduce dropout, and explore privacy-preserving approaches such as federated learning for cross-institutional validation.

References

[1] V. Tinto, “Dropout from higher education: A theoretical synthesis of recent research,” Rev. Educ. Res., vol. 45, no. 1, pp. 89–125, 1975.

[2] G. A. Borges et al., “A platform for early class dropout prediction of university students,” IEEE Access, vol. 13, pp. 109116–109133, 2025, doi: 10.1109/ACCESS.2025.3581751.

[3] A. J. Fernández-García et al., “A real-life machine learning experience for predicting university dropout at different stages using academic data,” IEEE Access, vol. 9, pp. 133076–133090, 2021, doi: 10.1109/ACCESS.2021.3115851.

[4] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, “Balancing performance and explainability in academic dropout prediction,” IEEE Trans. Learn. Technol., vol. 17, pp. 2086–2099, 2024, doi: 10.1109/TLT.2024.3425959.

[5] H. P. Singh and H. N. Alhulail, “Predicting student-teachers dropout risk and early identification: A four-step logistic regression approach,” IEEE Access, vol. 10, pp. 6470–6482, 2022, doi: 10.1109/ACCESS.2022.3141992.

[6] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794.

[8] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, “Revisiting deep learning models for tabular data,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, 2021, pp. 18932–18943.

[9] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 4768–4777.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.