ISSN 2723-6471

Course-Disjoint Evaluation and Capacity-Aware Triage for Student Dropout Risk Prediction

Wijiyanto[1,2*], Aris Marjuni[3], Ahmad Zainul Fanani[4], Ruri Suko Basuki[5]

[1,3,4,5]*Department of Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia*

[2]*Faculty of Computer Science, Universitas Duta Bangsa Surakarta, Surakarta, Indonesia*

(Received: February 20, 2026; Revised: April 25, 2026; Accepted: June 10, 2026; Available online: June 28, 2026)

Abstract

Early-warning systems for student dropout prevention require evaluation protocols and outputs that remain reliable when applied across heterogeneous academic contexts. This study quantifies how conventional random splits can overestimate performance when models are expected to generalize across different courses and proposes a decision-support layer that translates predicted risk into capacity-aware intervention policies. Using a benchmark higher-education dataset (N=4,424; 34 predictors; three classes: Dropout, Enrolled, Graduate) with 17 Course groups, phased prediction is implemented to reflect increasing evidence availability: S0 (pre-enrollment), S1 (plus semester-1 academic evidence), and S2 (plus semester-2 academic evidence). Baseline results are replicated with leakage-safe preprocessing (imputation, one-hot encoding, scaling) and Synthetic Minority Over-sampling Technique (SMOTE) applied strictly within training folds, comparing multinomial logistic regression, random forest, and tree-based boosting models. Deployment-oriented performance is assessed using StratifiedGroupKFold by Course to enforce course-disjoint testing. Discrimination is reported with Macro-F1 and Balanced Accuracy, while probability quality is evaluated using LogLoss, Brier score, expected calibration error, maximum calibration error, and reliability diagrams. Calibrated probabilities are translated into capacity-aware risk bands (Top-k% triage), selective prediction is evaluated via abstention to defer low-confidence cases, and split conformal prediction sets are optionally reported for multiclass uncertainty communication. Results show consistent performance drops under course-disjoint validation, confirming a non-trivial generalization gap. Error decomposition indicates that Enrolled is the most ambiguous class and exhibits phase-dependent confusion toward both terminal outcomes. Calibration shows phase-specific trade-offs between likelihood-based and worst-case calibration metrics, while risk bands yield high-precision triage under limited capacity, and abstention improves decision quality at reduced coverage. Overall, the study provides a deployment-oriented evaluation and decision-support workflow for translating dropout risk models into actionable capacity planning.

Keywords: Student Dropout, Course-Disjoint Validation, Probability Calibration, Risk Bands, Selective Prediction, Conformal Prediction

1. Introduction

Early-warning systems (EWS) for student dropout prevention are increasingly important because dropout directly affects students' academic trajectories and institutions' effectiveness in delivering academic support [1], [2]. Early risk identification enables targeted allocation of limited support resources (e.g., advising, remedial programs, and financial assistance) so interventions can be delivered before risk escalates into study failure. From an applied data science perspective, the goal is therefore not only predictive accuracy but also actionable outputs that can be integrated into institutional decision processes [3], [4]. In addition, successful EWS deployment depends on institutional readiness and stakeholder acceptance of the system [5].

However, EWS success is not determined solely by high classification scores [6]. In real deployment, predictive models must remain robust under contextual shifts such as differences in cohorts, assessment practices, and learning dynamics across courses or programs [7]. Many prior studies still rely on random splits or stratified cross-validation that may mix highly similar contexts between training and testing, producing optimistic performance estimates that do not reflect out-of-course generalization required in

***Corresponding author: Wijiyanto (wijiyanto@udb.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i3.1419

practice. This issue is particularly salient in higher education, where course-level heterogeneity provides a realistic source of domain shift and can materially alter model performance once deployed [8], [9].

Moreover, operational EWS require decision-actionable outputs: reliable risk probabilities for ranking students, setting intervention priorities, and planning intervention capacity [10], [11]. Consequently, evaluation should go beyond discrimination alone by reporting probability quality (calibration) and translating model outputs into capacity-aware policies (e.g., Top-k% risk bands) [6], [10]. In this study, these requirements are integrated with a phased prediction setup (S0–S2) that mirrors increasing availability of academic evidence over time, from early screening to later, more stable decision-making [12], [13].

Despite the operational needs of EWS and the growing body of predictive studies, three gaps remain. First, there is limited adoption of course-disjoint (group-based) validation to test out-of-course generalization, even though it better approximates deployment, where models are applied to courses not seen during training [2], [7]. Second, while many studies report discrimination metrics (e.g., accuracy or F1), systematic reporting of probability quality using calibration evidence (e.g., NLL/LogLoss, Brier, ECE/MCE, and reliability diagrams) and explicit policy translation via thresholding, risk bands, and capacity simulations remains comparatively rare [10], [14]. Third, in multiclass formulations such as {Dropout, Enrolled, Graduate} used in widely adopted benchmarks, the Enrolled class often behaves as a heterogeneous intermediate state and becomes a major source of ambiguity, yet its error structure is rarely analyzed explicitly to inform policy design for borderline or transitional cases [6], [12], [13].

To address these gaps, this study makes six contributions. (1) Phased S0/S1/S2 baselines are replicated using popular models and leakage-safe SMOTE within training folds, reporting Macro-F1 and Balanced Accuracy. (2) Deployment-oriented evaluation is conducted using StratifiedGroupKFold by Course and the generalization gap relative to Random-CV is quantified. (3) Calibration evidence (reliability diagrams and calibration metrics) is linked to capacity-aware risk banding (Top-k% triage) for intervention planning under limited resources. (4) Selective prediction via abstention is incorporated to restrict recommendations to high-confidence cases and to report coverage–performance trade-offs [15]. (5) Split conformal prediction sets are optionally reported to communicate multiclass uncertainty through empirical coverage and average set size [16]. (6) Enrolled error decomposition is reported across phases to substantiate its intermediate nature and clarify policy implications for triage in ambiguous cases. Overall, the central message is that deployable EWS require realistic generalization assessment, reliable probabilities, and capacity-aware decision policies so predictions can be translated into defensible actions under practical constraints [10].

2. Related Works

2.1. Student Success and Dropout Prediction

Dropout/success prediction in higher education is commonly formulated as supervised classification using a combination of demographic/administrative variables and academic performance signals, where semester-level academic evidence frequently provides the strongest predictive value [12], [17]. In the widely used “Predict Students’ Dropout and Academic Success (Dropout/Enrolled/Graduate)” dataset, the feature set comprising curricular units from semesters 1–2 serves as the strongest signal for predicting terminal outcomes; consequently, many studies use the ‘early’ (pre-enrollment or early in the program) versus “late” (after the semester has begun) scenarios as the primary comparison [18], [19], [20], [21]. Recent applied studies report competitive performance of tree-based boosting and ensembles (e.g., XGBoost/LightGBM/CatBoost) on mixed-type educational data, particularly when class imbalance is addressed through resampling or cost-sensitive strategies [17], [22]. Beyond accuracy, a growing line of work emphasizes operational relevance, namely, how predictive models can support early intervention and institutional decision processes, rather than serving purely as retrospective benchmarks [12], [13].

2.2. Evaluation Protocols, Domain Shift, and Leakage Risk

Most dropout prediction studies still employ random splits or stratified cross-validation, which can yield stable estimates but may be optimistically biased when the data contain structured heterogeneity across courses or programs

[23]. In deployment, such heterogeneity acts as a domain shift: models trained on some courses may underperform on unseen courses due to differences in cohorts, assessment practices, or learning dynamics [24], [25]. Accordingly, group-based evaluation (e.g., course-disjoint validation) has been advocated as a more deployment-aligned protocol to assess out-of-course generalization [7]. In addition, leakage risks, especially when resampling is not isolated within the training fold, can inflate reported performance and should be explicitly mitigated in experimental design [13], [26].

2.3. Probability Calibration for Decision Support

High discrimination (e.g., Macro-F1) does not guarantee that predicted probabilities are reliable for decision making [27], [6]. In decision-support settings, probabilities are used to rank risk, set action thresholds, and allocate limited resources; therefore, calibration evidence (e.g., reliability diagrams and metrics such as NLL/LogLoss, Brier score, and ECE/MCE) is essential to ensure decision-grade probability outputs [28], [29]. Practical evaluations further require leakage-safe calibration procedures (e.g., fold-internal proper-train/calibration splits) so that probability quality reflects out-of-sample performance rather than test-time overfitting [30].

2.4. Capacity-Aware Intervention, Abstention, and Uncertainty Communication

A robust applied data science pipeline should translate model outputs into actionable policy mechanisms under capacity constraints, such as Top-k% risk bands for triage and workload planning [11],[26]. Selective prediction (reject option/abstention) provides a safety mechanism by allowing the system to withhold recommendations for low-confidence cases, reporting explicit coverage–performance trade-offs that are directly interpretable for operational deployment [15]. Complementarily, conformal prediction sets provide distribution-free uncertainty communication by returning label sets with empirical coverage guarantees, which is useful for multiclass settings with borderline or transitional cases [16]. Together, capacity-aware risk banding, abstention, and conformal sets form a decision-support framework that bridges predictive modeling to policy translation under uncertainty and limited capacity [15],[16].

3. Dataset and Problem Formulation

3.1. Dataset

This study uses the widely adopted benchmark dataset *Predict Students' Dropout and Academic Success* from higher education, containing N = 4,424 records and 34 predictors (mixed demographic, administrative, macroeconomic, and semester-level academic features) with a three-class outcome: Dropout, Enrolled, and Graduate [20]. The dataset includes a Course attribute (17 groups) that naturally partitions students into distinct academic contexts and is used in this work to evaluate deployment-relevant generalization across unseen courses. Class distribution is imbalanced: Graduate is the majority class, followed by Dropout, while Enrolled is the smallest, motivating imbalance-aware evaluation and probability reporting in later sections. Given that courses define distinct academic contexts, table 1 summarizes course heterogeneity in terms of outcome composition and coarse academic profiles (semester grades and approvals).

Table 1. Course heterogeneity check

Course	N	Pct Dropout	Pct Enrolled	Pct Graduate	Sem1 Grade Mean	Sem1 Grade Median	Sem2 Grade Mean	Sem2 Grade Median	Sem1 Approved Mean	Sem2 Approved Mean
12	766	15.40	13.05	71.54	12.46	13.26	12.35	13.25	6.23	6.50
9	380	35.26	28.42	36.32	10.12	11.33	9.70	11.50	3.77	3.46
10	355	18.31	11.83	69.86	11.24	12.00	11.12	12.00	5.19	4.92
6	337	26.71	22.26	51.04	11.86	13.00	11.16	12.67	4.42	4.03
15	331	30.51	10.27	59.21	11.69	12.67	11.04	12.50	4.85	4.55
14	268	35.45	17.91	46.64	11.33	12.40	10.68	12.63	4.24	3.96
17	268	50.75	20.15	29.10	9.10	11.00	9.22	11.40	3.85	3.72
11	252	38.10	16.27	45.63	10.46	11.67	9.93	11.80	4.50	3.72
5	226	22.57	18.58	58.85	12.08	12.67	11.20	12.45	5.60	5.26
2	215	38.14	17.21	44.65	2.07	0.00	2.08	0.00	1.98	1.80

3	215	33.02	9.77	57.21	10.57	12.00	9.80	12.00	5.44	4.67
4	210	40.95	17.62	41.43	10.38	12.05	9.48	12.00	5.47	4.75
16	192	44.27	26.04	29.69	11.47	12.40	10.57	11.89	5.16	4.61
7	170	54.12	37.65	8.24	8.57	11.60	8.79	11.33	2.44	2.16
8	141	55.32	14.89	29.79	10.08	12.00	9.68	12.40	3.43	3.11
13	86	38.37	19.77	41.86	10.33	12.16	9.77	12.02	4.66	4.77
1	12	66.67	25.00	8.33	9.64	12.18	9.00	11.57	6.75	4.00

Dropout proportions vary substantially across courses, ranging from 15.40% (Course 12, n=766) to 66.67% (Course 1, n=12); excluding very small courses, the range remains wide (15.40% to 55.32%). Academic indicators also differ across courses (e.g., sem1_grade_mean and sem1_approved_mean), supporting Course as a practical proxy for contextual shift and justifying course-disjoint evaluation.

3.2. Task definition

The primary predictive task is multiclass classification with labels $y \in$ {Dropout,Enrolled,Graduate}, reflecting terminal outcomes (Dropout/Graduate) and a transitional state (Enrolled). From a decision-support standpoint, the multiclass output can additionally be mapped to an operational triage view, where At-risk = {Dropout,Enrolled} and Safe = {Graduate}, enabling capacity-aware intervention planning without replacing the main multiclass evaluation [11].

3.3. Phased prediction setup (S0–S2)

To mirror realistic evidence availability over time, a phased prediction design is adopted a phased prediction design with three information levels. S0 (pre-enrollment) uses only pre-entry attributes (demographic/administrative and macroeconomic variables) and excludes semester-level academic performance. S1 (+ semester-1 evidence) augments S0 with the full set of first-semester academic indicators (*curricular units 1st sem*). S2 (+ semester-2 evidence) further adds second-semester indicators (*curricular units 2nd sem*). This design operationalizes early screening versus later, more stable decision stages and supports systematic comparison under increasing evidence. For reproducibility and policy interpretation, the phase configuration (S0–S2), feature blocks, class distribution, and Course grouping used for course-disjoint evaluation are summarized in table 2.

Table 2. Dataset and phase configuration (S0–S2) with Course grouping for deployment-oriented evaluation.

Component	S0 (pre-enrollment)	S1 (+ semester 1 evidence)	S2 (+ semester 2 evidence)
Samples (N)	4,424	4,424	4,424
Target classes	{Dropout, Enrolled, Graduate}	{Dropout, Enrolled, Graduate}	{Dropout, Enrolled, Graduate}
Class distribution	Graduate 49.93%, Dropout 32.12%, Enrolled 17.95%	Graduate 49.93%, Dropout 32.12%, Enrolled 17.95%	Graduate 49.93%, Dropout 32.12%, Enrolled 17.95%
Feature blocks	Pre-entry + macroeconomic	S0 + semester-1 academic indicators	S1 + semester-2 academic indicators
Predictors (raw, pre-OHE)	22	28	34
Group variable for Group-CV	Course (17 groups)	Course (17 groups)	Course (17 groups)

4. Method

4.1. Leakage-Safe Preprocessing and Imbalance Handling

All preprocessing steps were designed to be leakage-safe by fitting transformations only on the training portion of each split and then applying them to the corresponding test portion. The pipeline includes missing-value imputation, one-hot encoding for categorical variables, and scaling for numeric variables to stabilize model training and probability estimates. To address class imbalance, SMOTE is applied strictly within the training fold after preprocessing has been

fit on the training data, preventing synthetic examples from leaking information into evaluation folds. This design aligns the reported performance and probability metrics with out-of-sample behavior expected under deployment [31], [32], [33]. SMOTE is applied only within the training fold after the preprocessing pipeline has been fit on the training data. The study uses SMOTE with k_neighbors=5 (default), sampling_strategy='not majority', and a fixed random seed for reproducibility. Because categorical variables are one-hot encoded prior to resampling, SMOTE operates in the resulting numeric feature space, and synthetic samples are generated only from the training fold to prevent leakage into evaluation folds.

4.2. Model Zoo

A compact model zoo is evaluated that is commonly used in applied classification and educational data mining: multinomial logistic regression (LR), random forest (RF), and tree-based boosting models (XGBoost, LightGBM, CatBoost). Boosting models serve as strong nonlinear baselines for mixed-type data, while LR provides a robust linear reference that often generalizes well under distribution shift [7]. Model selection is performed separately for each phase (S0–S2) to reflect phase-dependent evidence availability.

4.3. Validation Protocols: Baseline Vs Deployment-Oriented Evaluation

Results are reported under two complementary evaluation protocols. First, Baseline (benchmark replication): Random repeated stratified cross-validation (Random-CV) provides a conventional estimate of discrimination performance under i.i.d.-like splits and is used primarily for comparability with common practice. Second, Main (deployment-oriented): StratifiedGroupKFold by Course enforces course-disjoint folds and evaluates out-of-course generalization, approximating deployment where the model is applied to courses not observed during training. This protocol is critical when course-level heterogeneity acts as a domain shift and can materially affect decision quality in operational use [31].

4.4. Evaluation Metrics

Macro-F1 and Balanced Accuracy are reported because the task is multiclass and class-imbalanced, and the minority class (Enrolled) is operationally important. Macro-F1 weights each class equally and therefore prevents majority-class dominance in the overall score, while Balanced Accuracy averages class-wise recall and directly reflects sensitivity to each outcome category. Weighted-F1 can mask poor minority-class performance when the majority class is dominant, which is undesirable for early-warning decision support. Multiclass AUROC is not used as the primary metric because it requires multiple one-vs-rest aggregations and can be less interpretable for capacity-constrained triage; however, AUROC can be reported as supplementary evidence if needed [10].

4.5. Calibration Procedure

For each outer fold, calibration is performed strictly within the training portion to avoid leakage. (1) Split the training fold into a proper-training subset and a calibration subset using a stratified split. (2) Fit the base classifier on the proper-training subset. (3) Obtain uncalibrated class scores on the calibration subset and fit a calibrator using either sigmoid scaling or isotonic regression in a one-vs-rest manner. (4) Apply the fitted calibrator to the test fold scores to obtain calibrated probabilities. (5) Evaluate probability quality on the held-out fold using LogLoss, Brier score, ECE/MCE, and reliability diagrams. This procedure ensures that calibration metrics reflect out-of-sample probability quality.

4.6. Decision Layer: Risk Score, Thresholding, and Capacity-Aware Risk Bands

To translate predictions into actionable decisions, a dropout risk score $r(x) = \hat{P}(\text{Dropout} \mid x)$ and use it for ranking and triage. Capacity-aware intervention is operationalized via Top-k% risk bands, where the highest-risk k%of non-abstained cases are selected for intervention under a limited-resource assumption. Policy performance is reported using precision, recall, and lift relative to prevalence, enabling decision-makers to interpret how much better the triage is compared with random selection [11].

4.7. Selective Prediction (Abstention) and Coverage–Performance Reporting

Selective prediction is incorporated as a safety mechanism for operational deployment. Confidence is defined as

$$\gamma(x) = \max_j \hat{P}(y = j \mid x) \quad (1)$$

 ISSN 2723-6471

The system abstains when $\gamma(x) < \tau$, withholding recommendations for ambiguous cases. Coverage–performance curves are reported by varying τ to quantify the trade-off between the fraction of decided cases (coverage) and decision quality (e.g., Macro-F1 and policy metrics) [15],[34].

4.8. Split Conformal Prediction Sets (Optional)

To communicate uncertainty in multiclass predictions, split conformal prediction sets are optionally used. Data are partitioned into proper training, calibration, and test sets. Nonconformity scores from the calibration set determine thresholds for producing prediction sets with empirical coverage close to $1 - \alpha$. Empirical coverage and average set size are reported across α values to characterize the sharpness–reliability trade-off [14], [35].

Figure 1 summarizes the proposed deployment-oriented EWS pipeline, including leakage-safe preprocessing and within-fold imbalance handling, phase-wise evaluation (S0–S2) under Random-CV and StratifiedGroupKFold by Course, calibration and reliability assessment, and the decision-support layer (Top-k% risk bands and abstention), with optional split conformal prediction sets for uncertainty communication.

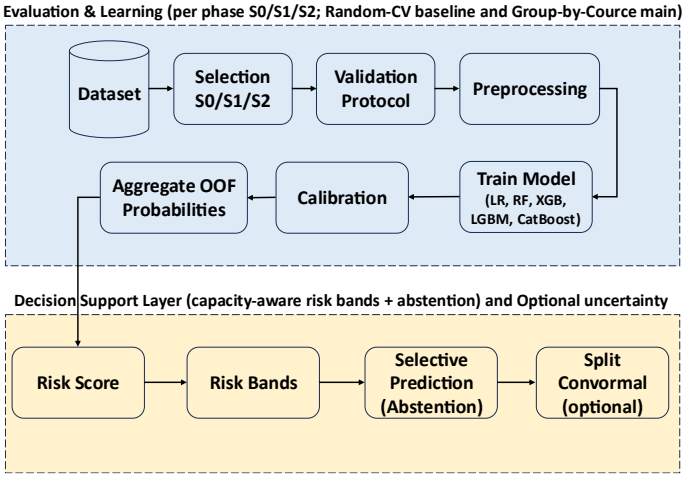

Figure 1. Methods flow diagram

In operational use, set-valued outputs can be mapped to actions as follows: if the set contains only {Dropout}, the case is eligible for high-priority triage; if the set contains {Dropout, Enrolled} or has size > 1, the case is treated as uncertain and routed to human review, additional evidence collection, or deferred decision until a later phase; if the set is {Graduate}, the case is deprioritized. This mapping connects conformal uncertainty to the same capacity-aware workflow used for risk bands and abstention.

4.9. Algorithm summary

The full pipeline cost scales with the number of phases, candidate models, calibration options, and cross-validation folds. Calibration and conformal layers add overhead compared to base training, but remain tractable because training is performed offline and operational use requires only scoring with the selected model and (optionally) applying the fitted calibrator or conformal set rule.

Algorithm 1. Capacity-aware calibrated risk banding with abstention.

Input: $D = \{(x_i, y_i, g_i)\}_{i=1}^{N}$with $y \in$ {Dropout,Enrolled,Graduate}, group g =Course; phases $p \in \{S0, S1, S2\}$with features F_p; model M_p; splitter $\mathcal{K}$(StratifiedGroupKFold by Course); preprocessing ϕ(impute+OHE+scale); SMOTE; calibration $c \in$ {none,sigmoid,isotonic}; capacity set $K\% = \{5,10,20\}$; abstain threshold τ.
Output: OOF probabilities P_p, abstention mask a_p, risk bands $B_{k,p}$, and metrics.

1. For each phase p: initialize $P \leftarrow 0_{N\times 3}$, $a \leftarrow 0_N$.
2. For each fold $(\mathcal{I}_{tr}, \mathcal{I}_{te}) \sim \mathcal{K}(D; \text{groups} = g)$:
3. Fit ϕon $X_{tr}[F_p]$; transform $Z_{tr} = \phi(X_{tr})$, $Z_{te} = \phi(X_{te})$.
4. Apply SMOTE on $(Z_{tr}, y_{tr}) \rightarrow (Z'_{tr}, y'_{tr})$(train only).

 ISSN 2723-6471

5. If $c =$ none: fit $M_p$on $(Z'_{tr}, y'_{tr})$and compute $\tilde{P} = \text{proba}(M_p, Z_{te})$.
6. Else: split $(Z'_{tr}, y'_{tr}) \rightarrow (Z_{prop}, y_{prop})$ and (Z_{cal}, y_{cal}); fit $M_p$on (Z_{prop}, y_{prop}); fit calibrator $\psi_c$on (Z_{cal}, y_{cal}); set $\tilde{P} = \psi_c(\text{scores}(M_p, Z_{te}))$.
7. Store OOF: $P[\mathcal{I}_{te}] \leftarrow \tilde{P}$(align missing classes if needed).
8. Compute confidence $\gamma_i = \max_j P_{ij}$and abstain $a_i = \mathbb{1}[\gamma_i < \tau]$.
9. For non-abstained $S = \{i: a_i = 0\}$, define risk $r_i = P_{i,\text{Dropout}}$.
10. For each $k \in K\%$, set risk band $B_{k,p}$as Top-k% of $S$by r_i(descending).
11. Report OOF discrimination (Macro-F1, Balanced Accuracy), calibration (LogLoss, Brier, ECE/MCE, reliability), and policy metrics on $B_{k,p}$(precision, recall, lift, coverage).
12. Return $\{P_p, a_p, B_{k,p}\}$and metrics for all phases.

5. Results

5.1. Benchmark Replication Under Random-CV (Phased S0–S2)

Benchmark replication results are first reported benchmark replication results under Random repeated stratified cross-validation with leakage-safe preprocessing and SMOTE applied within training folds (table 3). Performance improves monotonically with evidence availability. In S0 (pre-enrollment), the best baseline is XGBoost with Macro-F1 = 0.575 ± 0.020 and Balanced Accuracy = 0.570 ± 0.017. In S1 (+ semester 1 evidence), CatBoost achieves the best discrimination with Macro-F1 = 0.690 ± 0.025 and Balanced Accuracy = 0.683 ± 0.024. In S2 (+ semester 2 evidence), performance further increases, and XGBoost reaches Macro-F1 = 0.719 ± 0.017 and Balanced Accuracy = 0.711 ± 0.017. These results establish credible baselines while highlighting that later-phase academic signals substantially enhance discriminative performance.

Table 3. Baseline Random-CV: Macro-F1 and Balanced Accuracy (mean±std) per fase S0–S2

Scenario	Model	F1 macro Mean	F1 macro Std	Bal acc Mean	Bal acc Std	Logloss Mean	Logloss Std	ECE Uniform Mean	ECE Uniform std
S0_pre_enrollment	XGB	0.575	0.020	0.570	0.017	0.818	0.037	0.067	0.016
S0_pre_enrollment	CatBoost	0.567	0.020	0.564	0.017	0.818	0.033	0.060	0.015
S0_pre_enrollment	LR	0.561	0.016	0.570	0.017	0.889	0.033	0.056	0.015
S0_pre_enrollment	LGBM	0.557	0.018	0.552	0.017	1.210	0.095	0.215	0.019
S0_pre_enrollment	RF	0.547	0.030	0.547	0.024	0.817	0.030	0.053	0.017
S1_plus_sem1	CatBoost	0.690	0.025	0.683	0.024	0.617	0.033	0.050	0.011
S1_plus_sem1	XGB	0.684	0.023	0.676	0.021	0.629	0.039	0.065	0.016
S1_plus_sem1	LR	0.679	0.026	0.686	0.027	0.686	0.040	0.048	0.009
S1_plus_sem1	LGBM	0.676	0.027	0.668	0.026	1.083	0.111	0.178	0.017
S1_plus_sem1	RF	0.664	0.024	0.656	0.022	0.655	0.022	0.085	0.019
S2_plus_sem2	XGB	0.719	0.017	0.711	0.017	0.573	0.043	0.070	0.014
S2_plus_sem2	LGBM	0.719	0.017	0.709	0.018	1.054	0.119	0.162	0.015
S2_plus_sem2	CatBoost	0.716	0.018	0.708	0.017	0.557	0.035	0.043	0.013
S2_plus_sem2	LR	0.713	0.017	0.721	0.019	0.619	0.035	0.039	0.014
S2_plus_sem2	RF	0.710	0.022	0.700	0.021	0.584	0.021	0.080	0.016

5.2. Deployment-Oriented Evaluation: Stratifiedgroupkfold By Course

Out-of-course generalization is then evaluated out-of-course generalization using StratifiedGroupKFold by Course (table 4). Compared to Random-CV, discrimination consistently drops, confirming that course-disjoint evaluation is stricter and more deployment-aligned. Best-performing models also differ by phase: S0: XGBoost (Macro-F1 = 0.502 ± 0.020; BalAcc = 0.504 ± 0.021), S1: multinomial LR (Macro-F1 = 0.630 ± 0.048; BalAcc = 0.641 ± 0.041), and S2: CatBoost (Macro-F1 = 0.667 ± 0.075; BalAcc = 0.669 ± 0.064). Notably, LR becomes the most robust choice in S1 under course-disjoint evaluation, suggesting that simpler models may generalize better when course-level heterogeneity induces domain shift.

Table 4. Results of StratifiedGroupKFold by Course: Macro-F1, Balanced Accuracy, and calibration metrics.

Scenario	Model	F1 Macro mean	F1 Macro std	Bal Acc mean	Bal Acc std	Logloss mean	Logloss std	ECE Uniform mean	ECE Uniform std
S0_pre_enrollment	XGB	0.502	0.020	0.504	0.021	0.919	0.044	0.087	0.022
S0_pre_enrollment	CatBoost	0.496	0.028	0.498	0.030	0.925	0.042	0.084	0.032
S0_pre_enrollment	LR	0.495	0.030	0.518	0.031	0.991	0.023	0.091	0.036
S0_pre_enrollment	LGBM	0.490	0.009	0.492	0.013	1.429	0.068	0.249	0.010
S0_pre_enrollment	RF	0.482	0.026	0.490	0.027	0.888	0.049	0.043	0.008
S1_plus_sem1	LR	0.630	0.048	0.641	0.041	0.987	0.569	0.061	0.051
S1_plus_sem1	CatBoost	0.628	0.060	0.630	0.048	0.823	0.337	0.086	0.085
S1_plus_sem1	XGB	0.619	0.078	0.625	0.061	0.946	0.572	0.123	0.096
S1_plus_sem1	LGBM	0.612	0.071	0.615	0.054	1.830	1.244	0.233	0.078
S1_plus_sem1	RF	0.606	0.067	0.607	0.053	0.760	0.130	0.086	0.027
S2_plus_sem2	CatBoost	0.667	0.075	0.669	0.064	0.821	0.499	0.085	0.095
S2_plus_sem2	XGB	0.665	0.084	0.668	0.070	0.960	0.751	0.119	0.099
S2_plus_sem2	LGBM	0.664	0.070	0.666	0.058	1.906	1.577	0.211	0.077
S2_plus_sem2	RF	0.660	0.086	0.660	0.069	0.735	0.264	0.110	0.045
S2_plus_sem2	LR	0.657	0.067	0.668	0.070	0.919	0.547	0.075	0.067

5.3. Generalization gap: Random-CV vs course-disjoint evaluation

To quantify the optimism of Random-CV, the study compute the generalization gap as the difference between Random-CV and Group-by-Course mean Macro-F1 per model and phase (table 5; figure 2). The average Macro-F1 gaps across models are 0.0684 (S0), 0.0596 (S1), and 0.0530 (S2), indicating persistent overestimation under random splits. For example, the S0 best baseline (XGBoost) drops from 0.575 (Random-CV) to 0.502 (Group-by-Course), while S2 drops from 0.719 to 0.667 for the phase-best models. Overall, these results show that deployment-oriented evaluation is necessary to avoid overly optimistic claims when model usage requires transfer across courses.

Table 5. Generalization gap (F1, BalAcc) Random-CV → Group-CV.

Scenario	Model	F1 macro Mean random	Bal acc Mean random	F1 macro Mean group	Bal acc Mean group	F1_gap	Bal acc gap
S0_pre_enrollment	XGB	0.575	0.569	0.502	0.504	0.073	0.065
S0_pre_enrollment	CatBoost	0.567	0.564	0.496	0.498	0.071	0.066
S0_pre_enrollment	LGBM	0.557	0.552	0.490	0.492	0.067	0.060
S0_pre_enrollment	LR	0.561	0.570	0.495	0.518	0.066	0.052
S0_pre_enrollment	RF	0.547	0.547	0.482	0.490	0.065	0.056
S1_plus_sem1	XGB	0.684	0.676	0.619	0.625	0.065	0.051
S1_plus_sem1	LGBM	0.676	0.668	0.612	0.615	0.064	0.052
S1_plus_sem1	CatBoost	0.690	0.683	0.628	0.630	0.062	0.053
S1_plus_sem1	RF	0.664	0.656	0.606	0.607	0.058	0.049
S1_plus_sem1	LR	0.679	0.686	0.630	0.641	0.049	0.044
S2_plus_sem2	LR	0.713	0.721	0.657	0.668	0.056	0.053

ISSN 2723-6471

S2_plus_sem2	XGB	0.719	0.711	0.665	0.668	0.055	0.043
S2_plus_sem2	LGBM	0.718	0.709	0.664	0.666	0.054	0.043
S2_plus_sem2	RF	0.710	0.700	0.660	0.660	0.051	0.040
S2_plus_sem2	CatBoost	0.716	0.708	0.667	0.669	0.049	0.039

Figure 2 compares Random-CV and Group-by-Course Macro-F1 across models and phases, illustrating consistent downward shifts under course-disjoint evaluation.

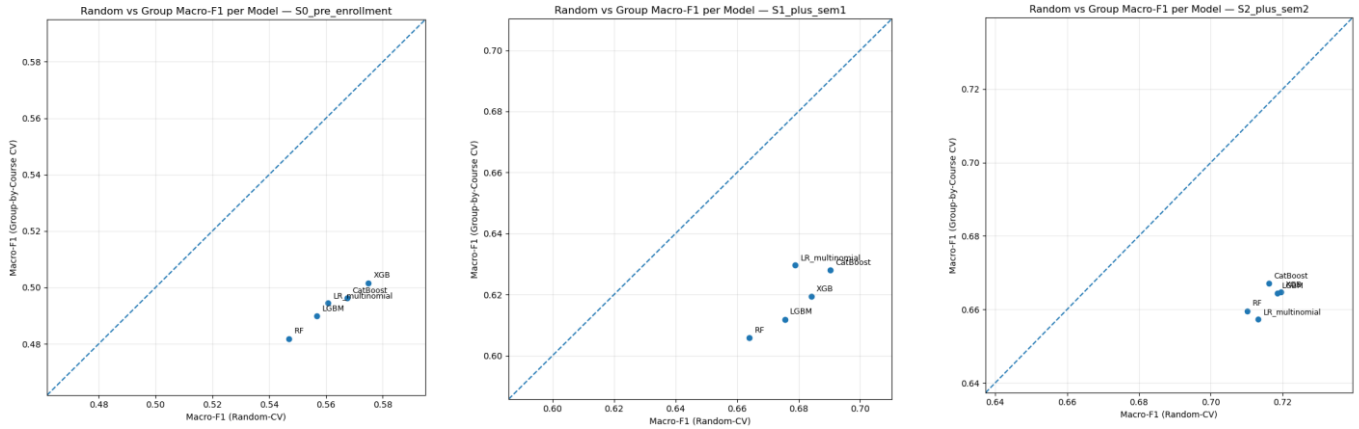

Figure 2. Plot F1 Random vs Group per model

The study further assesses whether the observed generalization gap exceeds repeat-level variability by performing paired Wilcoxon signed-rank tests on repeat-wise mean Macro-F1 (paired by repeat seed; n_splits=5, n_pairs=10) between Random-CV and course-disjoint Group-by-Course CV (table 6). Random-CV consistently outperforms Group-by-Course in all phases (win rate = 1.0), and all paired tests are significant (p = 0.001953), indicating that the performance drop reflects a systematic generalization gap rather than fold noise.

Table 6. Paired significance test for generalization gap (Random-CV vs Group-by-Course).

Scenario	Model	N Splits	N Pairs	Random Mean	Group Mean	Delta Mean	Delta Median	Win Rate Random Gt Group	Wilcoxon Stat	P_Value
S0_pre_enrollment	XGB	5	10	0.5674	0.5059	0.0615	0.0596	1.0	0.0	0.001953
S1_plus_sem1	LR_multinomial	5	10	0.6729	0.6438	0.0291	0.0286	1.0	0.0	0.001953
S2_plus_sem2	CatBoost	5	10	0.7164	0.6774	0.0390	0.0388	1.0	0.0	0.001953

5.4. Error Analysis: OOF Confusion Matrices Under Group-By-Course

Pooled out-of-fold (OOF) predictions are analyzed for the best Group-by-Course model in each phase (figure 3). While table 4 reports mean ± std across folds, OOF confusion matrices summarize aggregate behavior across all held-out observations. In S0 (XGB), Enrolled is the most challenging class, with low recall (0.170), and misclassification is dominated by Enrolled → Graduate. In S1 (LR_multinomial), Enrolled recognition improves markedly (recall 0.523), but confusion increases between Enrolled and the terminal classes, consistent with its transitional status. In S2 (CatBoost), terminal classes become more stable (Dropout recall 0.771, Graduate recall 0.871), yet Enrolled remains the main bottleneck (recall 0.401). This pattern motivates Enrolled-specific error decomposition and supports policy-aware handling of intermediate trajectories under course-disjoint evaluation.

 ISSN 2723-6471

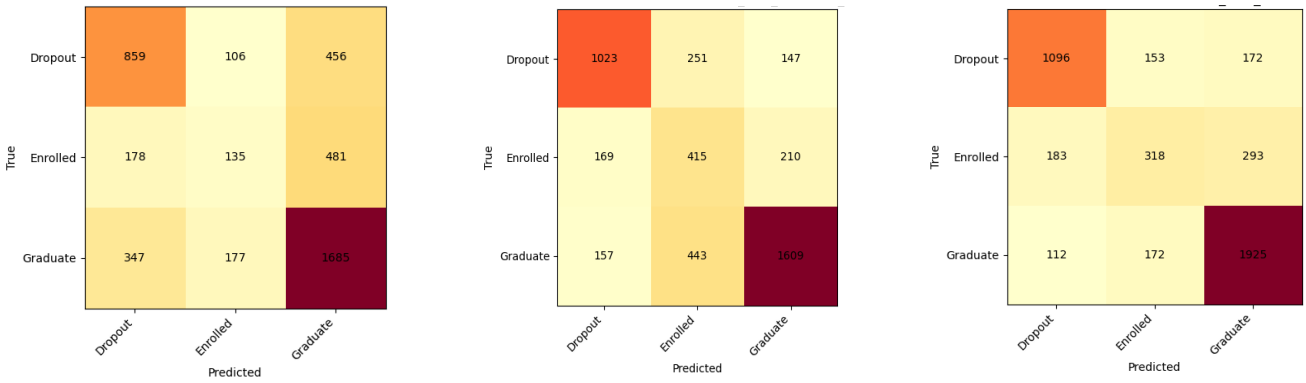

Figure 3. OOF confusion matrix for the Group-by-Course analysis of phase S0, S1, and S2.

5.5. Enrolled error decomposition (intermediate-class evidence)

To substantiate the "intermediate class" hypothesis, errors for true Enrolled are decomposed across phases (table 7; figure 4). In S0, only 17.00% of Enrolled cases are correctly classified (135/794), while 60.58% are predicted as Graduate (481/794) and 22.42% as Dropout (178/794). With first-semester evidence in S1, correct Enrolled increases to 52.27% (415/794), and misclassifications become more balanced toward Graduate (26.45%) and Dropout (21.28%). In S2, correct enrollment is 40.05% (318/794), with a substantial split toward Graduate (36.90%) and Dropout (23.05%). This phase-dependent redistribution supports the interpretation that Enrolled aggregates heterogeneous, evolving trajectories rather than representing a stable terminal outcome, and therefore should be handled using policy-aware decision layers rather than relying solely on hard labels.

Table 7. Enrolled error decomposition: Enrolled {Dropout, Enrolled, Graduate} per phase.

Phase	N Enrolled	To Dropout	Correct Enrolled	To Graduate	Count To Dropout	Count Correct	Count To Graduate
S0_pre_enrollment (XGB)	794	22.42	17.00	60.58	178	135	481
S1_plus_sem1 (LR_multinomial)	794	21.28	52.27	26.45	169	415	210
S2_plus_sem2 (CatBoost)	794	23.05	40.05	36.90	183	318	293

Enrolled error decomposition for the OOF Group-by-Course dataset: the stacked bar chart shows the distribution of predictions for true Enrolled (Pred Dropout, Pred Enrolled/correct, Pred Graduate) for S0–S2.

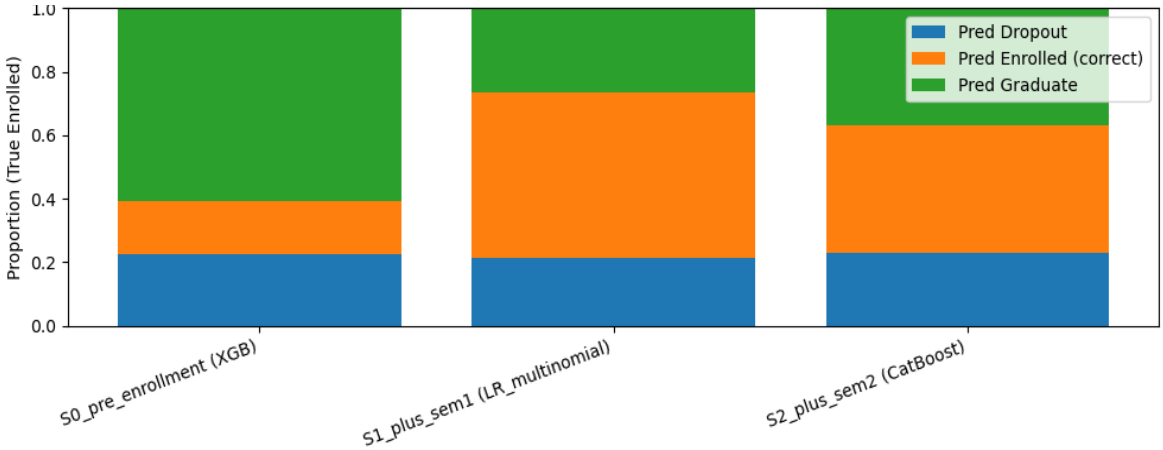

Figure 4. Stacked bar Enrolled decomposition across phases

5.6. Calibration Evidence (None Vs Sigmoid Vs Isotonic)

Probability quality is evaluated for the phase-best Group-by-Course models using within-fold calibration (table 8; figure 5). Results show phase-specific trade-offs between discrimination and calibration metrics. S0 (XGB): sigmoid reduces worst-case calibration error (MCE 0.172 vs 0.186 for none) but slightly reduces Macro-F1 (0.488 vs 0.502). Isotonic degrades NLL substantially (LogLoss 1.056). S1 (LR): sigmoid improves NLL (LogLoss 0.929 vs 0.987 for none) with marginal Macro-F1 change (0.632 vs 0.630), but worst-case calibration worsens (MCE 0.172 vs 0.102 for none). S2 (CatBoost): none provides the best NLL (LogLoss 0.821), while sigmoid reduces MCE (0.260 vs 0.304) with slightly worse NLL (0.857). Isotonic becomes unstable in NLL (LogLoss 1.867).

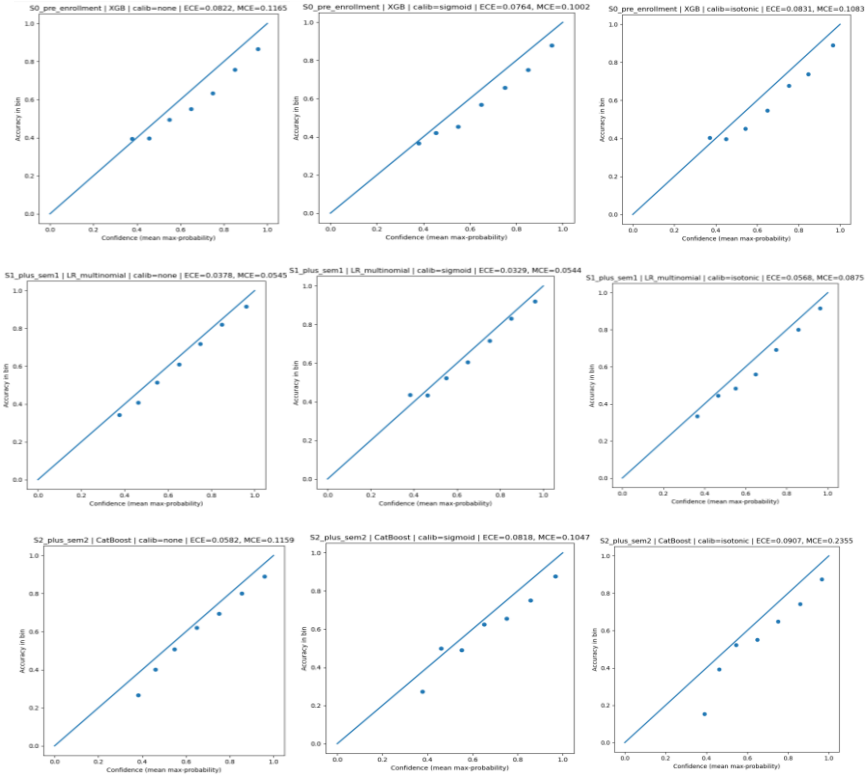

Figure 5. Reliability diagrams per phase (none vs sigmoid vs isotonic).

In reliability diagrams, the diagonal line indicates perfect calibration; points above the diagonal indicate under-confidence, while points below indicate over-confidence. ECE summarizes average miscalibration across bins, whereas MCE captures the worst-case bin deviation; therefore, MCE is more relevant when overconfidence in a small subset is operationally costly. Both are reported to support decision-driven calibration selection. Reliability diagrams (figure 5) mirror these trade-offs, indicating that calibration choices should be decision-driven (e.g., minimizing NLL for ranking vs controlling MCE for high-stakes triage).

Table 8. Calibration metrics (NLL, Brier, ECE, MCE) per phase & method calibration.

Scenario	Model	Calibration	F1 Macro Mean	F1 Macro Std	Bal Acc Mean	Bal Acc Std	Logloss Mean	Logloss Std	Brier Mean	Brier Std	ECE Uniform Mean	ECE Uniform Std	MCE Uniform Mean	MCE Uniform Std
S0_pre_enrollment	XGB	none	0.502	0.020	0.504	0.021	0.919	0.044	0.531	0.026	0.087	0.022	0.186	0.072
S0_pre_enrollment	XGB	sigmoid	0.488	0.024	0.492	0.023	0.920	0.054	0.534	0.031	0.089	0.037	0.172	0.079
S0_pre_enrollment	XGB	isotonic	0.464	0.015	0.482	0.015	1.056	0.148	0.533	0.032	0.094	0.035	0.207	0.082
S1_plus_sem1	LR	sigmoid	0.632	0.046	0.644	0.040	0.929	0.454	0.450	0.076	0.064	0.044	0.172	0.076
S1_plus_sem1	LR	none	0.630	0.048	0.641	0.041	0.987	0.569	0.452	0.078	0.061	0.051	0.102	0.056
S1_plus_sem1	LR	isotonic	0.623	0.045	0.635	0.042	1.038	0.322	0.459	0.073	0.089	0.045	0.180	0.059
S2_plus_sem2	CatBoost	none	0.667	0.075	0.669	0.064	0.821	0.499	0.396	0.142	0.085	0.095	0.304	0.209

 ISSN 2723-6471

S2_plus_sem2	CatBoost	sigmoid	0.670	0.079	0.673	0.065	0.857	0.505	0.406	0.146	0.109	0.094	0.260	0.100
S2_plus_sem2	CatBoost	isotonic	0.656	0.088	0.659	0.069	1.867	2.247	0.410	0.145	0.121	0.094	0.249	0.122

Group-by-Course reliability diagrams for phases S0, S1, and S2 (best model) with three calibration options (none, sigmoid, isotonic). The diagonal line represents ideal calibration; deviations indicate over- or under-confidence.

5.7. Capacity-Aware Risk Bands (Top-K% Triage)

Capacity-limited interventions are operationalized using Top-k% risk bands based on $P(\text{Dropout} \mid x)$(table 9; figure 6). With dropout prevalence 0.321, the Top-10% band yields high precision in all phases: S0 (XGB, none) precision 0.916 with recall 0.285 (lift 2.853), S1 (LR, sigmoid) precision 0.971 with recall 0.302 (lift 3.022), and S2 (CatBoost, none) precision 0.914 with recall 0.284 (lift 2.846). Increasing capacity to Top-20% improves recall substantially (e.g., S2 recall 0.538 with precision 0.863), supporting practical planning of intervention workload under different resource scenarios. These results demonstrate that the proposed pipeline yields decision-ready outputs that translate directly into defensible triage policies.

Table 9. Risk bands (top-k%): precision/recall/lift for dropout.

Scenario	Model	Calibration	Topk	Coverage	Precision dropout	Recall dropout	Lift	Prevalence dropout
S0_pre_enrollment	XGB	none	0.05	0.050	0.946	0.147	2.944	0.321
S0_pre_enrollment	XGB	none	0.1	0.100	0.916	0.285	2.853	0.321
S0_pre_enrollment	XGB	none	0.2	0.200	0.764	0.476	2.378	0.321
S0_pre_enrollment	XGB	none	0.3	0.300	0.627	0.586	1.952	0.321
S1_plus_sem1	LR	sigmoid	0.05	0.050	0.991	0.154	3.085	0.321
S1_plus_sem1	LR	sigmoid	0.1	0.100	0.971	0.302	3.022	0.321
S1_plus_sem1	LR	sigmoid	0.2	0.200	0.893	0.556	2.779	0.321
S1_plus_sem1	LR	sigmoid	0.3	0.300	0.774	0.723	2.409	0.321
S2_plus_sem2	CatBoost	none	0.05	0.050	0.950	0.148	2.958	0.321
S2_plus_sem2	CatBoost	none	0.1	0.100	0.914	0.284	2.846	0.321
S2_plus_sem2	CatBoost	none	0.2	0.200	0.863	0.538	2.688	0.321
S2_plus_sem2	CatBoost	none	0.3	0.300	0.808	0.754	2.515	0.321

Capacity-aware risk curves for Group-by-Course: changes in precision and recall of Dropout as intervention capacity increases (Top-k%). Lift relative to prevalence is shown as a summary of the policy's benefits.

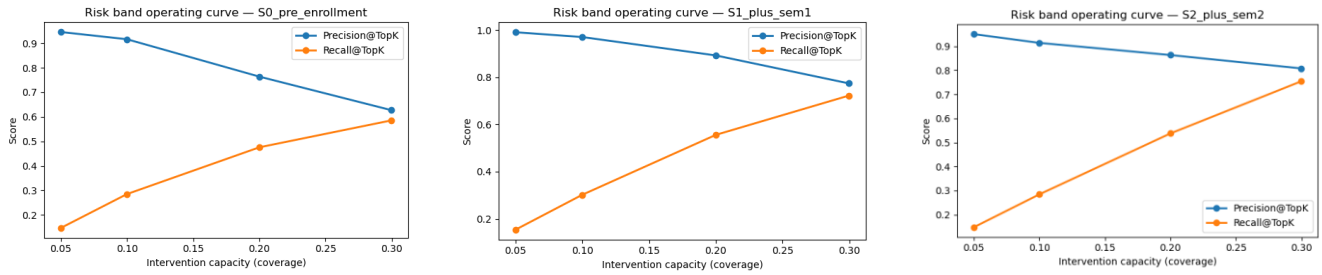

Figure 6. Capacity curve (coverage vs precision/recall).

To contextualize Top-k policies, the capacity fraction k% is translated into approximate case volumes per term, and example staffing assumptions are provided (e.g., number of advisors and cases per advisor). This illustrates how institutional capacity maps to k% and supports realistic triage planning.

5.8. Selective Prediction (Abstention): Coverage–Performance Trade-Off

The study evaluates abstention by thresholding prediction confidence and reporting coverage versus performance trade-offs (table 10; figure 7). At full coverage, Macro-F1 is 0.511 (S0), 0.649 (S1), and 0.686 (S2). At an operationally reasonable coverage level near 0.8–0.9, performance improves: for S0, threshold 0.50 yields coverage 0.819 and Macro-F1 0.531; for S1, threshold 0.50 yields coverage 0.883 and Macro-F1 0.674; for S2, threshold 0.60 yields coverage 0.838 and Macro-F1 0.717. Policy-relevant dropout precision/recall also improves at these operating points (e.g., S2 dropout precision 0.826; dropout recall 0.839). This confirms abstention as a practical "safety valve" to avoid consuming intervention capacity on low-confidence cases.

Table 10. Selective prediction: threshold vs coverage vs macro-F1/balance accuracy/dropout recall.

Threshold	Coverage	Macro_f1	Bal_acc	Dropout precision	Dropout recall	Scenario	Model	Calibration
0.30	1.000	0.511	0.512	0.621	0.605	S0_pre_enrollment	XGB	none
0.35	0.998	0.511	0.513	0.621	0.604	S0_pre_enrollment	XGB	none
0.40	0.972	0.514	0.515	0.631	0.610	S0_pre_enrollment	XGB	none
0.45	0.910	0.518	0.520	0.642	0.624	S0_pre_enrollment	XGB	none
0.50	0.819	0.531	0.533	0.678	0.652	S0_pre_enrollment	XGB	none
0.55	0.718	0.544	0.547	0.717	0.682	S0_pre_enrollment	XGB	none
0.60	0.620	0.548	0.554	0.748	0.718	S0_pre_enrollment	XGB	none
0.65	0.541	0.560	0.566	0.781	0.752	S0_pre_enrollment	XGB	none
0.70	0.461	0.574	0.581	0.802	0.791	S0_pre_enrollment	XGB	none
0.75	0.384	0.588	0.596	0.819	0.839	S0_pre_enrollment	XGB	none
0.80	0.312	0.601	0.607	0.842	0.873	S0_pre_enrollment	XGB	none
0.85	0.238	0.610	0.615	0.868	0.913	S0_pre_enrollment	XGB	none
0.90	0.159	0.621	0.619	0.900	0.957	S0_pre_enrollment	XGB	none
0.95	0.091	0.548	0.519	0.934	0.997	S0_pre_enrollment	XGB	none
0.30	1.000	0.649	0.658	0.769	0.721	S1_plus_sem1	LR	sigmoid
0.35	0.999	0.649	0.658	0.769	0.721	S1_plus_sem1	LR	sigmoid
0.40	0.988	0.652	0.661	0.771	0.721	S1_plus_sem1	LR	sigmoid
0.45	0.954	0.661	0.671	0.781	0.730	S1_plus_sem1	LR	sigmoid
0.50	0.883	0.674	0.683	0.791	0.752	S1_plus_sem1	LR	sigmoid
0.55	0.797	0.690	0.698	0.809	0.783	S1_plus_sem1	LR	sigmoid
0.60	0.706	0.710	0.716	0.825	0.812	S1_plus_sem1	LR	sigmoid
0.65	0.612	0.728	0.733	0.857	0.840	S1_plus_sem1	LR	sigmoid
0.70	0.524	0.741	0.741	0.875	0.874	S1_plus_sem1	LR	sigmoid
0.75	0.443	0.754	0.748	0.894	0.901	S1_plus_sem1	LR	sigmoid
0.80	0.357	0.756	0.741	0.912	0.932	S1_plus_sem1	LR	sigmoid
0.85	0.280	0.755	0.739	0.938	0.947	S1_plus_sem1	LR	sigmoid
0.90	0.201	0.746	0.722	0.946	0.968	S1_plus_sem1	LR	sigmoid
0.95	0.127	0.714	0.683	0.964	0.991	S1_plus_sem1	LR	sigmoid
0.30	1.000	0.686	0.681	0.788	0.771	S2_plus_sem2	CatBoost	none

0.35	1.000	0.686	0.681	0.788	0.771	S2_plus_sem2	CatBoost	none
0.40	0.993	0.690	0.684	0.791	0.778	S2_plus_sem2	CatBoost	none
0.45	0.973	0.693	0.687	0.798	0.784	S2_plus_sem2	CatBoost	none
0.50	0.935	0.704	0.697	0.807	0.799	S2_plus_sem2	CatBoost	none
0.55	0.882	0.713	0.704	0.819	0.823	S2_plus_sem2	CatBoost	none
0.60	0.838	0.717	0.705	0.826	0.839	S2_plus_sem2	CatBoost	none
0.65	0.788	0.719	0.705	0.834	0.855	S2_plus_sem2	CatBoost	none
0.70	0.739	0.713	0.696	0.840	0.868	S2_plus_sem2	CatBoost	none
0.75	0.678	0.705	0.689	0.845	0.884	S2_plus_sem2	CatBoost	none
0.80	0.620	0.684	0.670	0.848	0.898	S2_plus_sem2	CatBoost	none
0.85	0.550	0.665	0.657	0.851	0.918	S2_plus_sem2	CatBoost	none
0.90	0.450	0.632	0.637	0.862	0.942	S2_plus_sem2	CatBoost	none
0.95	0.290	0.620	0.634	0.885	0.972	S2_plus_sem2	CatBoost	none

Coverage–performance curves for selective prediction: changes in Macro-F1 (and/or precision/recall dropout) as a function of coverage when the system applies threshold-based abstention

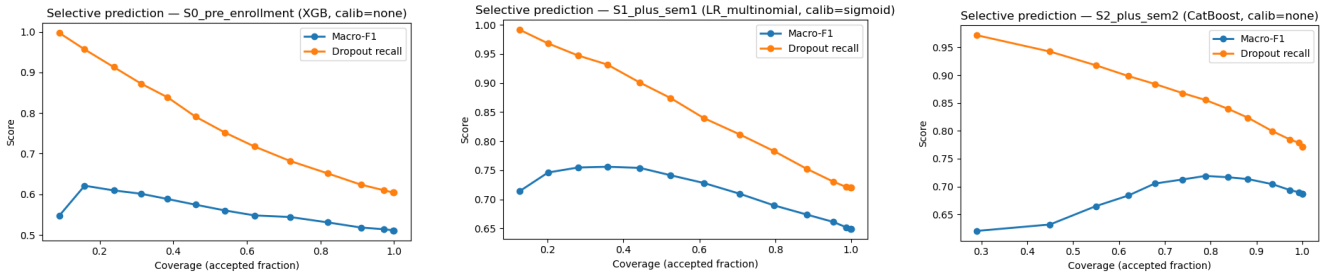

Figure 7. Coverage–performance curves.

Coverage directly implies the fraction of cases receiving an automatic recommendation, while $1-$coverage corresponds to cases routed to manual review or deferred decisions. For example, at coverage 0.838 on N = 4,424, approximately 717 students (16.2%) would be deferred from automated triage and require an alternative workflow.

5.9. Split Conformal Prediction Sets (Optional Uncertainty Communication)

Finally, split conformal prediction sets are reported as an optional uncertainty communication layer (table 11). For S0 (XGB), $\alpha = 0.10$ yields empirical coverage 0.892 with an average set size of 1.962, while $\alpha = 0.05$ increases coverage to 0.940 at the cost of a larger set (2.270). For S1 (LR), $\alpha = 0.10$ achieves coverage 0.882 with set size 1.592, and $\alpha = 0.05$ reaches 0.958 with set size 2.015. These results quantify the sharpness–reliability trade-off and provide a transparent mechanism for communicating multiclass uncertainty, especially valuable for borderline cases around the Enrolled class.

Table 11. Split conformal: alpha vs empirical coverage vs avg set size.

Scenario	Model	Alpha	Empirical coverage	Avg_set_size
S0_pre_enrollment	XGB	0.05	0.940	2.270
S0_pre_enrollment	XGB	0.10	0.892	1.962
S0_pre_enrollment	XGB	0.15	0.814	1.628
S0_pre_enrollment	XGB	0.20	0.767	1.427
S1_plus_sem1	LR	0.05	0.958	2.015

S1_plus_sem1	LR	0.10	0.882	1.592
S1_plus_sem1	LR	0.15	0.824	1.349
S1_plus_sem1	LR	0.20	0.766	1.183

6. Discussion and Implications

Course-disjoint evaluation provides a more deployment-aligned view of EWS than conventional random splits. The results show a consistent generalization gap under StratifiedGroupKFold by Course, indicating that Random-CV can overestimate deployable performance when course-level heterogeneity induces domain shift. In applied decision analytics terms, this matters because institutions that plan intervention capacity based on optimistic estimates risk misallocating limited support resources (e.g., advising slots, remediation capacity) and may underdeliver timely assistance to the highest-need cases [11].

The phased design (S0→S1→S2) clarifies how evidence availability affects decision stability. As semester-level academic indicators become available, discrimination improves for terminal outcomes (Dropout and Graduate), while Enrolled remains structurally challenging. Enrolled error decomposition supports interpreting Enrolled as a heterogeneous intermediate state whose errors split toward both terminal outcomes and vary by phase [19]. This implies that policy design should not treat Enrolled as a stable terminal label. Instead, institutions should handle Enrolled through probability-based triage and uncertainty-aware workflows, recognizing that borderline cases may require monitoring, additional evidence, or human review rather than immediate high-cost interventions.

Decision support requires decision-grade probabilities, not only high Macro-F1. Probabilities are used to rank students, set action thresholds, and allocate scarce resources; therefore, calibration evidence (NLL/Brier and ECE/MCE, supported by reliability diagrams) is essential for governable decision making. The findings show phase-specific trade-offs, for example, improving worst-case miscalibration may slightly reduce discrimination, reinforcing that calibration should be chosen based on the decision objective (ranking vs high-stakes triage). Risk bands operationalize these concepts into auditable policy: Top-k% triage converts probabilistic scores into a fixed workload, and reports expected yield via precision/recall/lift, enabling transparent capacity planning and accountability. Selective prediction adds a safety valve: abstention withholds recommendations for low-confidence cases and supports a two-tier governance workflow (automated triage for confident cases; human review or deferred decisions for ambiguous cases) [15]. Optional conformal prediction sets complement abstention by providing set-valued outputs with empirical coverage guarantees, which can be particularly useful for multiclass borderline decisions around Enrolled. Together, risk bands, abstention, and conformal sets form a coherent applied decision framework that bridges predictive modeling to simulation-based policy translation under uncertainty and limited capacity.

This work is intended as a decision-support framework rather than an automated decision maker. Recommended interventions should remain under human oversight, especially for uncertain cases (abstention or set-valued predictions). Institutions should also assess fairness and potential disparate impact across demographic groups (e.g., gender, nationality, financial status proxies) before operational use, and maintain audit trails documenting thresholds, capacity rules, and review outcomes.

Threats to validity include the use of the course as a proxy for domain shift and potential prior distortions from SMOTE; both are mitigated through course-disjoint evaluation, leakage-safe resampling, and calibration reporting [33]. Calibration metrics are also sensitive to binning and confidence definitions; reporting multiple metrics alongside reliability diagrams reduces single-metric bias and supports transparent governance. The study does not include deep sequence-based models in the current benchmark; while such models may capture temporal dynamics, the present dataset is tabular, and the main contribution focuses on evaluation realism and decision policy translation. Adding sequence-based baselines and longitudinal trajectory models is an important direction for future work. These limitations motivate future work that strengthens external validity (e.g., cohort/time splits or cross-institution evaluation), explores alternative formulations better aligned with an intermediate class (e.g., ordinal or multi-task learning), and optimizes decision layers end-to-end for capacity-constrained objectives rather than selecting thresholds post hoc [31]. Future deployments can further integrate abstention and conformal sets into explicit routing protocols (triage vs human review)

and monitor calibration drift over time. This work does not report an institutional pilot or real-world deployment. Therefore, claims are limited to deployment-oriented evaluation and simulation-based policy translation under capacity constraints. External validation across institutions and prospective studies is required before operational adoption. The study does not perform longitudinal trajectory modeling; therefore, the 'intermediate' interpretation is operational, based on error decomposition and uncertainty patterns rather than latent trajectory inference. All experiments are conducted on a single public benchmark dataset; therefore, external validity across institutions, cohorts, and policy environments is untested. Cross-institution evaluation and prospective validation are necessary before claims of operational generalizability.

These findings translate into two practical implications for applied settings and reporting standards. Practically, institutions can adopt Top-10% risk band triage as a defensible workload planning rule: it fixes intervention capacity and provides expected yield (precision/lift), supporting auditable allocation of limited resources [11]. A phased policy can align intervention intensity with evidence availability: S0 supports early screening and monitoring, S1 enables stronger triage after semester-1 evidence becomes available, and S2 supports more stable targeting decisions. At the same time, abstention routes ambiguous cases to human review or deferred decisions, and conformal sets can be used when explicit uncertainty communication is required. [32]. Methodologically, applied EWS studies should (i) report Random-CV alongside group-disjoint protocols and quantify the generalization gap, (ii) report probability quality using calibration evidence (NLL/Brier plus ECE/MCE and reliability diagrams), and (iii) report policy translation via capacity-aware triage metrics (Top-k% risk bands) and coverage–performance under abstention. This package of evidence makes performance claims more credible and directly actionable for institutions operating under limited intervention capacity.

7. Conclusion

This study demonstrates that deployable dropout EWS require an applied decision analytics perspective: realistic generalization assessment, decision-grade probabilities, and capacity-aware policies. Course-disjoint evaluation exposes a consistent generalization gap relative to Random-CV, indicating that random splits can overestimate deployable performance under course heterogeneity. The phased design improves stability for terminal outcomes as evidence grows, while Enrolled behaves as a heterogeneous intermediate state, motivating probability-based triage and uncertainty-aware handling rather than hard-label decisions. By integrating leakage-safe preprocessing and within-fold calibration with Top-k% risk bands and abstention, the proposed pipeline translates predictive modeling into defensible triage and workload planning under limited resources. Optional conformal prediction sets further support transparent uncertainty communication. Overall, the contribution is not only predictive performance but a practical blueprint that connects model outputs to institutional action in a governable and resource-aware manner.

8. Declarations

8.1 Author Contributions

Conceptualization: W.J. and A.M.; Methodology: W.J.; Software: W.J.; Validation: W.J., A.M., A.Z., and R.S.; Formal Analysis: W.J. and A.M.; Investigation: W.J.; Data Curation: W.J.; Writing Original Draft Preparation: W.J.; Writing Review and Editing: W.J., A.M., A.Z., and R.S.; Visualization: W.J.; Supervision: A.M., A.Z., and R.S. All authors have read and agreed to the published version of the manuscript.

8.2 Data Availability Statement

The dataset used in this study is publicly available from the UCI Machine Learning Repository [20].

8.3 Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

8.4 Institutional Review Board Statement

Not applicable.

8.5 Informed Consent Statement

Not applicable.

8.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Boudjehem and Y. Lafifi, "An early warning system to predict dropouts inside e-learning environments," *Educ Inf Technol*, vol. 29, no. 13, pp. 16365–16385, Sep. 2024, doi: 10.1007/s10639-024-12498-1.

[2] R. M. Santos and R. Henriques, "Accurate, timely, and portable: Course-agnostic early prediction of student performance from LMS logs," *Computers and Education: Artificial Intelligence*, vol. 5, no. pp. 1-15, 2023, doi: 10.1016/j.caeai.2023.100175.

[3] D. Bañeres, M. E. Rodríguez-González, A.-E. Guerrero-Roldán, and P. Cortadas, "An early warning system to identify and intervene online dropout learners," *Int J Educ Technol High Educ*, vol. 20, no. 1, pp. 1-13, Jan. 2023, doi: 10.1186/s41239-022-00371-5.

[4] M. Sailer, M. Ninaus, S. E. Huber, E. Bauer, and S. Greiff, "The End is the Beginning is the End: The closed-loop learning analytics framework," *Computers in Human Behavior*, vol. 158, no. Sep., pp. 1-15, Sep. 2024, doi: 10.1016/j.chb.2024.108305.

[5] J. E. Raffaghelli, M. E. Rodríguez, A.-E. Guerrero-Roldán, and D. Bañeres, "Applying the UTAUT model to explain the students' acceptance of an early warning system in Higher Education," *Computers & Education*, vol. 182, no. Jun., pp. 1-18, Jun. 2022, doi: 10.1016/j.compedu.2022.104468.

[6] J. Mandi, J. Kotary, S. Berden, M. Mulamba, V. Bucarey, T. Guns, and F. Fioretto, "Decision-Focused Learning: Foundations, State of the Art, Benchmark and Future Opportunities," *jair*, vol. 80, no. pp. 1623–1701, Aug. 2024, doi: 10.1613/jair.1.15320.

[7] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain Generalization: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 2022, No. 1, pp. 1–20, 2022, doi: 10.1109/TPAMI.2022.3195549.

[8] C. Kumar, G. Walton, P. Santi, and C. Luza, "Random Cross-Validation Produces Biased Assessment of Machine Learning Performance in Regional Landslide Susceptibility Prediction," *Remote Sensing*, vol. 17, no. 2, pp. 213-228, Jan. 2025, doi: 10.3390/rs17020213.

[9] A. Palanci, R. M. Yılmaz, and Z. Turan, "Learning analytics in distance education: A systematic review study," *Educ Inf Technol*, vol. 29, no. 17, pp. 22629–22650, Dec. 2024, doi: 10.1007/s10639-024-12737-5.

[10] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Mach Learn*, vol. 112, no. 9, pp. 3211–3260, Sep. 2023, doi: 10.1007/s10994-023-06336-7.

[11] U. Sadana, A. Chenreddy, E. Delage, A. Forel, E. Frejinger, and T. Vidal, "A survey of contextual optimization methods for decision-making under uncertainty," *European Journal of Operational Research*, vol. 320, no. 2, pp. 271–289, Jan. 2025, doi: 10.1016/j.ejor.2024.03.020.

[12] V. Christou, I. Tsoulos, V. Loupas, A. T. Tzallas, C. Gogos, P. S. Karvelis, N. Antoniadis, E. Glavas, and N. Giannakeas, "Performance and early drop prediction for higher education students using machine learning," *Expert Systems with Applications*, vol. 225, no. Sep., pp. 1-19, Sep. 2023, doi: 10.1016/j.eswa.2023.120079.

[13] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, "Balancing Performance and Explainability in Academic Dropout Prediction," *IEEE Trans. Learning Technol.*, vol. 17, no. 1, pp. 2086–2099, 2024, doi: 10.1109/TLT.2024.3425959.

[14] J. Dong, Z. Jiang, D. Pan, Z. Chen, Q. Guan, H. Zhang, G. Gui, and W. Gui, "A Survey on Confidence Calibration of Deep Learning-Based Classification Models Under Class Imbalance Data," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 36, no. 9, pp. 15664–15684, Sep. 2025, doi: 10.1109/TNNLS.2025.3565159.

[15] K. Hendrickx, L. Perini, D. Van Der Plas, W. Meert, and J. Davis, "Machine learning with a reject option: a survey," *Mach Learn*, vol. 113, no. 5, pp. 3073–3110, May 2024, doi: 10.1007/s10994-024-06534-x.

[16] X. Zhou, B. Chen, Y. Gui, and L. Cheng, "Conformal Prediction: A Data Perspective," *ACM Comput. Surv.*, vol. 58, no. 2, pp. 1–37, Jan. 2026, doi: 10.1145/3736575.

[17] J. G. C. Krüger, A. D. S. Britto, and J. P. Barddal, "An explainable machine learning approach for student dropout prediction," *Expert Systems with Applications*, vol. 233, no. Dec., pp. 1-13, Dec. 2023, doi: 10.1016/j.eswa.2023.120933.

[18] Z. Azizah, T. Ohyama, X. Zhao, Y. Ohkawa, and T. Mitsuishi, "Predicting at-risk students in the early stage of a blended learning course via machine learning using limited data," *Computers and Education: Artificial Intelligence*, vol. 7, no. Dec., pp. 1-21, Dec. 2024, doi: 10.1016/j.caeai.2024.100261.

[19] L. Herrmann and J. Weigert, "AI-based prediction of academic success: Support for many, disadvantage for some?," *Computers and Education: Artificial Intelligence*, vol. 7, no. Dec., pp. 1-13, Dec. 2024, doi: 10.1016/j.caeai.2024.100303.

[20] M. V. M. Valentim Realinho, "Predict Students' Dropout and Academic Success." *UCI Machine Learning Repository,* 2021. doi: 10.24432/C5MC89.

[21] H. Waheed, S.-U. Hassan, R. Nawaz, N. R. Aljohani, G. Chen, and D. Gasevic, "Early prediction of learners at risk in self-paced education: A neural network approach," *Expert Systems with Applications*, vol. 213, no. mar., pp. 1-18, Mar. 2023, doi: 10.1016/j.eswa.2022.118868.

[22] H. El Aouifi, M. El Hajji, and Y. Es-Saady, "A hybrid approach for early-identification of at-risk dropout students using LSTM-DNN networks," *Educ Inf Technol*, vol. 29, no. 14, pp. 18839–18857, Oct. 2024, doi: 10.1007/s10639-024-12588-0.

[23] G. I. Austin, I. Pe'er, and T. Korem, "Distributional bias compromises leave-one-out cross-validation," *Sci. Adv.*, vol. 11, no. 48, pp. 1-16, Nov. 2025, doi: 10.1126/sciadv.adx6976.

[24] A. Apicella, F. Isgrò, and R. Prevete, "Don't push the button! Exploring data leakage risks in machine learning and transfer learning," *Artif Intell Rev*, vol. 58, no. 11, pp. 339-352, Aug. 2025, doi: 10.1007/s10462-025-11326-3.

[25] J. Allgaier and R. Pryss, "Practical approaches in evaluating validation and biases of machine learning applied to mobile health studies," *Commun Med*, vol. 4, no. 1, pp. 76-93, Apr. 2024, doi: 10.1038/s43856-024-00468-0.

[26] M. Phan, A. De Caigny, and K. Coussement, "A decision support framework to incorporate textual data for early student dropout prediction in higher education," *Decision Support Systems*, vol. 168, no. pp. 1-20, May 2023, doi: 10.1016/j.dss.2023.113940.

[27] T. Dawood, C. Chen, B. S. Sidhu, B. Ruijsink, J. Gould, B. Porter, M. K. Elliott, V. Mehta, C. A. Rinaldi, E. Puyol-Antón, R. Razavi, and A. P. King, "Uncertainty aware training to improve deep learning model calibration for classification of cardiac MR images," *Medical Image Analysis*, vol. 88, no. Aug., pp. 1-21, Aug. 2023, doi: 10.1016/j.media.2023.102861.

[28] W. Huang, G. Cao, J. Xia, J. Chen, H. Wang, and J. Zhang, "H-Calibration: Rethinking Classifier Recalibration With Probabilistic Error-Bounded Objective," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 10, pp. 9023–9042, Oct. 2025, doi: 10.1109/TPAMI.2025.3582796.

[29] T. Dimitriadis, T. Gneiting, A. I. Jordan, and P. Vogel, "Evaluating probabilistic classifiers: The triptych," *International Journal of Forecasting*, vol. 40, no. 3, pp. 1101–1122, Jul. 2024, doi: 10.1016/j.ijforecast.2023.09.007.

[30] M. Kängsepp, K. Valk, and M. Kull, "On the usefulness of the fit-on-test view on evaluating calibration of classifiers," *Mach Learn*, vol. 114, no. 4, pp. 105-119, Apr. 2025, doi: 10.1007/s10994-024-06652-6.

[31] B. Sonnleitner, T. Madou, M. Deceuninck, F. Theodosiou, and Y. R. Sagaert, "Evaluation of early student performance prediction given concept drift," *Computers and Education: Artificial Intelligence*, vol. 8, no. Jun., pp. 1-19, Jun. 2025, doi: 10.1016/j.caeai.2025.100369.

[32] D. Hooshyar, G. Šír, Y. Yang, E. Kikas, R. Hämäläinen, T. Kärkkäinen, D. Gašević, and R. Azevedo, "Towards responsible AI for education: Hybrid human-AI to confront the elephant in the room," *Computers and Education: Artificial Intelligence*, vol. 9, no. Dec., pp. 1-24, Dec. 2025, doi: 10.1016/j.caeai.2025.100524.

[33] A. Demircioğlu, "Applying oversampling before cross-validation will lead to high bias in radiomics," *Sci Rep*, vol. 14, no. 1, pp. 1-13, May 2024, doi: 10.1038/s41598-024-62585-z.

[34] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif Intell Rev*, vol. 57, no. 6, pp. 137-149, May 2024, doi: 10.1007/s10462-024-10759-6.

[35] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *Ann. Statist.*, vol. 51, no. 2, pp. 816–845, Apr. 2023, doi: 10.1214/23-AOS2276.