Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 5191~5204
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.12300

Calibration-aware decision-support policy for student dropout early warning

Wijiyanto[1], Aris Marjuni[2], Ahmad Zainul Fanani[2], Ruri Suko Basuki[2], Nurmalitasari[1]
[1]Faculty of Computer Science, Universitas Duta Bangsa (UDB), Surakarta, Indonesia
[2]Department of Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University (UDINUS), Semarang, Indonesia

Article Info

Article history:

Received Jan 12, 2026
Revised Sep 10, 2026
Accepted Oct 4, 2026

Keywords:

Decision support
Early warning system
Probability calibration
Risk stratification
Threshold optimization

ABSTRACT

Given the limited intervention capacity, dropout early warning systems must be operationalized as auditable decision-support policies rather than standalone predictive models. This study developed and evaluated a calibrated, recall-constrained early-warning policy using data from 329 students at a private university in Central Java, Indonesia, with academic, socioeconomic, and perceptual/relational predictors, to prioritize dropout risk for intervention triage. The workflow integrated: i) predictive modeling evaluated under nested cross-validation, ii) probability calibration to obtain reliable risk scores, and iii) constraint-based decision thresholding (minimum recall target), combined with risk band stratification (low/medium/high/very high), to map probabilities to actionable service levels. Performance was assessed across imbalance-sensitive utility area under the precision-recall curve (AUPRC), discrimination area under the receiver operating characteristic curve (ROC-AUC), and probability reliability (Brier score), complemented by significance testing and ablation/importance analyses for policy interpretability. Results indicated that calibration and recall-based threshold tuning improved decision usability while preserving competitive predictive performance. The primary contribution was an explicit decision-support layer (calibration, threshold tuning, and risk bands) with empirical validation and factor-driven guidance for implementation. Accordingly, the resulting calibration-aware decision-support policy provided an auditable basis for translating student dropout risk estimates into prioritized intervention actions under limited institutional capacity.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Wijiyanto
Faculty of Computer Science, Universitas Duta Bangsa
Bhayangkara street 55, Surakarta, Indonesia
Email: wijiyanto@udb.ac.id

1. INTRODUCTION

Student attrition is a significant issue for both individuals and institutions: students forfeit opportunities for skill enhancement and social advancement, while universities encounter academic, operational, and financial repercussions (e.g., increased service demands, retention objectives, and the efficacy of support programs). In recent years, learning analytics and educational data mining have been increasingly employed to identify at-risk students promptly, enabling interventions before the risk becomes irreparable [1], [2]. Recent systematic reviews and meta-analytic evidence, together with empirical dropout-prediction studies, demonstrate the growing use of learning analytics for risk identification and intervention support in higher education [3]–[5].

The primary issue in the subject lies not merely in achieving high predictive performance but in transforming forecasts into implementable intervention options. Because early-warning interventions require institutional resources and support capacity, operational systems need explicit triage rules for determining which students should receive intervention and at what level [6], [7]. The early warning research indicates that intervention practices are comparatively infrequent in relation to prediction studies, and intervention decisions frequently stay at the discretion of teachers rather than being based on measurable and auditable policies [5], [8]. This underscores the necessity for research that prioritizes decision policy as the principal outcome, rather than merely "highly accurate models" [9]. Recent dropout early-warning research can be organized into three complementary streams. Prediction-oriented studies primarily optimize discrimination and classification performance using machine-learning or ensemble models [5], [10]. Intervention-oriented frameworks move beyond prediction by linking identified risk to teacher or institutional actions; examples include dynamic dropout-risk identification and intervention mechanisms [8] and the student performance prediction and action framework [7], [9]. A smaller calibration-aware stream has emphasized that classifier scores should not automatically be interpreted as probabilities. Nagy and Molontay [11] compared Platt/sigmoid and isotonic calibration in an interpretable dropout-prediction framework, while Carballo-Mendívil *et al.* [12] incorporated isotonic probability calibration into a recent pre-enrollment dropout early-warning system. These studies demonstrate increasing attention to reliable probabilities and actionable predictions, but calibration, operating-point selection, and intervention prioritization are still commonly treated as separate design components.

Dropout models typically produce a risk score that may inform counseling, nudges, or support prioritization. Three methodological gaps remain for operational decision support. First, probability quality is often underexamined: discrimination and calibration are distinct properties, so competitive area under the receiver operating characteristic curve (ROC-AUC) or area under the precision-recall curve (AUPRC) does not guarantee reliable probabilities [13], [14]. Institutional use should also preserve human oversight [15] and monitor cohort drift that can alter predictive behavior over time [16]. Second, risk scores are often converted into decisions without an explicit operating objective; in resource-constrained settings, the threshold should reflect a policy target such as minimum dropout recall rather than a fixed default [17]. Third, risk estimates require an interpretable mapping to intervention priority and workload [18]. These gaps motivate an integrated policy combining probability calibration, recall-constrained threshold selection, and operational risk stratification. The remaining gap is therefore not the absence of accurate dropout classifiers, calibrated models, or intervention frameworks individually, but the lack of an integrated and auditable policy layer that jointly controls probability reliability, the operational decision threshold, and intervention priority under an explicit institutional objective. Existing systems provide complementary elements such as dynamic risk identification and intervention [8], prediction-to-action infrastructure [7], [9], or calibrated and explainable risk estimation [11], [12]. In contrast, the present study evaluates the complete chain of training-side probability calibration, calibration-method sensitivity, recall-constrained threshold optimization, predefined risk-band stratification, and reproducible policy evaluation within the same held-out validation workflow. A practical engineering challenge is therefore to adapt the operating threshold under class imbalance without converting the early-warning model into an autonomous decision maker. In the present study, this policy is evaluated offline and is intended as a human-in-the-loop triage mechanism in which calibrated probabilities and the operating threshold support, rather than replace, institutional judgment.

Accordingly, the contribution of this study is a reproducible calibration-aware decision-support policy rather than a new classification algorithm. The contribution is operationalized through four measurable components: i) nested cross-validation with confidence intervals and paired statistical/effect-size analysis; ii) empirical comparison of sigmoid and isotonic calibration, followed by training-side threshold selection under $\text{Recall}_{\text{dropout}} \geq 0.80$; iii) interpretable low/medium/high/very high risk stratification for intervention prioritization; and iv) an auditable implementation procedure supported by publicly available reproducibility resources.

To ensure transparent state-of-the-art positioning, the literature used in this study was selected through a targeted review of recent peer-reviewed studies, primarily published during 2023–2026. Studies were retained when they directly addressed at least one of four themes relevant to the proposed framework: student dropout/at-risk prediction, probability calibration, decision-threshold or risk-based decision making, and learning analytics intervention or explainability. Methodological studies outside education were included only when they provided direct foundations for calibration, threshold selection, or validation. Non-peer-reviewed sources and studies without a direct connection to these themes were excluded. This procedure was used for literature positioning rather than as a systematic review. From an engineering perspective, the framework separates model discrimination, probability calibration, operating-point selection, and service-level mapping into auditable modules that can be independently validated or updated. From an educational perspective, this modular policy converts model outputs into transparent intervention priorities, allowing institutions to align missed-risk tolerance with available advising, counseling, and academic-support capacity.

2. PROPOSED DECISION-SUPPORT PROCEDURE

This section specifies a decision-support procedure that converts calibrated dropout probabilities into auditable intervention decisions under explicit operational constraints. The approach aims to provide a dependable risk score (dropout probability) via calibration, establish a clear decision threshold (e.g., target recall ≥0.80), and thereafter categorize risks into risk bands for triage and intervention. The scientific basis relies on current work regarding: i) early warning and intervention systems [8], ii) the significance of probability calibration and probabilistic assessment [13], and iii) optimization of objective/cost-based decision thresholds [17].

2.1. Definition of terms

The risk score represents the likelihood of dropout as predicted by the model for each student. Calibration is a process that renders the risk score "meaningful as a probability" (reliable), enabling its application in decision-making and risk reporting [13]. The decision threshold (t*) is a criterion that converts the calibrated risk score into an intervention decision under a policy objective (e.g., recall ≥0.80). Risk bands are discrete strata of the calibrated risk score: low ($p_{\text{cal}} \leq 0.20$), medium ($0.20 < p_{\text{cal}} \leq 0.50$), high ($0.50 < p_{\text{cal}} \leq 0.80$), and very high ($p_{\text{cal}} > 0.80$); these bands are used for triage, prioritization, and capacity planning.

2.2. Policy procedures

To operationalize the calibrated risk score as an intervention policy, the proposed procedure consists of five sequential stages. These stages describe the conceptual decision logic that connects prediction, probability reliability, threshold selection, risk stratification, and intervention mapping. The implementation of these stages within the experimental validation framework is described in subsection 3.3.

a. Prediction

A dropout prediction model is developed using academic, demographic, and/or learning-related variables to produce an initial dropout probability for each student. Although integrating multiple sources of academic and learning information may improve predictive performance, the model must be rigorously evaluated before its outputs are used for intervention decisions [1].

b. Calibration

Because raw model scores may be overconfident or underconfident, probability calibration is applied before the scores are interpreted as intervention risks. Sigmoid/Platt or isotonic calibration can be used within an appropriate validation framework, and calibration quality can be assessed using reliability-oriented metrics such as the Brier score [13], [14], [19].

c. Policy-target threshold selection

A policy objective specified in this study is a minimum dropout recall of 0.80 to reduce the number of at-risk students who remain undetected. The decision threshold is selected from calibrated scores under this recall constraint while considering a secondary operational criterion when multiple feasible thresholds are available. This makes the decision rule dependent on explicit objectives and operating consequences rather than on a conventional fixed threshold or global discrimination metric alone [17], [20]–[22].

d. Risk-band stratification

The calibrated probability P_{cal} is subsequently mapped into four operational risk bands: low (P_{cal}≤0.20), medium (0.20<P_{cal}≤0.50), high (0.50<P_{cal}≤0.80), and very high (P_{cal}>0.80). These author-defined operational bands provide an interpretable mechanism for triage and capacity planning, consistent with the broader use of risk stratification to prioritize students for preventive support [23].

e. Intervention mapping

Each risk band is associated with a proportional level of support. Low-risk students receive preventive information and self-help resources; medium-risk students receive monitoring and targeted nudges; high-risk students receive proactive academic or counselling support; and very high-risk students receive the highest-priority, multi-party intervention when institutional resources permit [7], [24], [25].

2.3. Decision-support system perspective

The proposed early-warning policy can be implemented as a decision-support layer that sits on top of an institutional data pipeline and converts calibrated risk scores into auditable intervention actions. The operational flow begins with data ingestion from academic information systems and the learning platforms, followed by preprocessing/feature construction consistent with the experimental setup in section 3. The predictive model produces an initial dropout risk estimate $\hat{p}(y = \text{dropout} \mid x)$, which is subsequently transformed by a calibration module to obtain reliable probabilities p_{cal} that are suitable for policy decisions. A policy engine then applies the recall-constrained decision threshold t* and risk-band stratification rules to

map p_{cal} into discrete service levels (low/medium/high/very high), enabling triage under limited intervention capacity.

In deployment, the policy engine outputs: i) a prioritized intervention queue (e.g., high and very high bands) and ii) recommended actions aligned with institutional resources (e.g., advising, counseling, academic support), and an audit log capturing model version, calibration method, threshold t*, and decision rationale for each case. The system can operate in batch mode (e.g., weekly) or near-real-time depending on data availability. Finally, continuous monitoring is required to detect performance drift and calibration degradation, triggering periodic recalibration or retraining and re-estimation of t* when necessary. Because the resulting risk classifications inform student-support decisions, institutional deployment should preserve human oversight and auditable decision processes rather than treat model outputs as fully automated determinations [15].

3. RESEARCH METHOD

3.1. Dataset and variables

This study used a real-world higher-education dataset of n=329 students originally collected and reported in [26], from a private university in Central Java, Indonesia. The dataset was selected because it provides a verified binary academic outcome (Dropout versus Graduate) together with academic, socioeconomic, and perceptual/relational variables relevant to intervention-oriented risk analysis. Dropout (1) was the minority class, and Graduate (0) was the majority class. The eight predictors were parent income, individual income, cumulative grade point average (CGPA), marital status, employment status, program-study interest, relationship with lecturer/supervisor, and lecturer-quality satisfaction. No additional synthetic predictors were engineered; the analysis retained the variables documented in the source dataset. Preprocessing was implemented within the modeling pipeline: marital status was one-hot encoded, while the remaining ordinal/numeric variables were standardized using StandardScaler. Training-dependent transformations were fitted using training-side data only within the cross-validation procedure to avoid information leakage.

For M2_RF_SMOTENC, synthetic minority over-sampling technique for nominal and continuous (SMOTE-NC) was implemented as the first step of an imbalanced-learn pipeline, followed by fold-specific preprocessing and random forest (RF) estimation. Marital status and employment status were designated as categorical variables during SMOTE-NC resampling; after resampling, preprocessing proceeded according to the feature transformations described above. Because RandomizedSearchCV received the complete pipeline as its estimator, SMOTE-NC, scaling, and classifier fitting were re-executed independently within each inner-training fold. An explicit inner-fold audit confirmed zero training-validation index overlap; the three audited inner-training partitions contained 175–176 original observations and were balanced to 278–280 observations only after fold-specific SMOTE-NC. Hereafter, balanced RF is denoted as M1_BRF, while RF combined with SMOTE-NC is denoted as M2_RF_SMOTENC.

3.2. Evaluation protocol

The main evaluation used nested cross-validation to reduce model-selection bias [27]–[29]: outer RepeatedStratifiedKFold 5×2 (10 test folds) and inner StratifiedKFold with three folds for RandomizedSearchCV (n_iter=15, scoring=average precision). For M1_BRF, the search space included n_estimators ∈{150, 250, 350, 500}; for M2_RF_SMOTENC, n_estimators ∈{200, 400, 600}. For both candidates, max_depth ∈{None, 3, 5, 8, 12}, min_samples_split ∈{2, 5, 10}, min_samples_leaf ∈{1, 2, 4}, and max_features ∈{sqrt, log2, 0.6, 0.8}. AUPRC (average precision) was used as the primary ranking metric because dropout was the minority class, while ROC-AUC was retained as a complementary discrimination measure [30]. Probabilistic reliability was assessed using the Brier score, log-loss, and expected calibration error (ECE) [14], [31]. ECE was computed using both 10-bin uniform and 10-bin quantile partitions as $ECE = \sum_{m=1}^{M}(|B_m|/n)|\bar{y}_m - \bar{p}_m|$, where $\bar{y}_m$ is the observed dropout frequency and $\bar{p}_m$ is the mean predicted probability in bin m. Performance was reported as mean±standard deviation (SD) and 95% confidence intervals across the 10 outer test folds.

$$\mathrm{CI}_{95\%} = \bar{x} \pm t_{0.975,\,n-1} \cdot \frac{s}{\sqrt{n}} \quad (1)$$

where $\bar{x}$ is the mean score, s is the SD, n=10 is the number of outer test folds, and $t_{0.975,\,n-1}$ is the 97.5th percentile of Student's t-distribution with n-1 degrees of freedom. Comparisons of the primary candidates were conducted using the Wilcoxon signed-rank test on paired outer-fold scores [32]. Shapiro-Wilk diagnostics of the paired differences did not indicate significant departures from normality for AUPRC

(W=0.921, p=0.362) or recall (W=0.924, p=0.388). Nevertheless, the Wilcoxon test was retained as the prespecified conservative paired comparison because only 10 paired outer-fold observations were available. All hyperparameter selection was confined to the inner cross-validation procedure, whereas the outer folds were used only for held-out performance estimation.

3.3. Policy layer and implementation algorithm

The policy layer was implemented within each outer cross-validation fold to maintain strict separation between model development and held-out evaluation. After hyperparameter tuning on the outer-training partition, sigmoid calibration was fitted using only the outer-training data with three-fold internal cross-validation. For a base-model score s, sigmoid/Platt calibration used the mapping $p_{cal} = 1/[1 + \exp(As + B)]$, where A and B were estimated by minimizing binary log-loss, $-\sum_i[y_i\ log\ (p_i) + (1 - y_i) \log\ (1 - p_i)]$. Isotonic calibration instead estimated a non-decreasing mapping g by minimizing $\sum_i[y_i - g(s_i)]^2$. Three internal folds were used to balance calibration-sample size and fold independence for the relatively small dataset (n=329), while maintaining complete separation from the outer-test fold and consistency with the inner model-selection structure. To assess sensitivity to the choice of calibration method, the complete calibration-policy procedure was additionally repeated using isotonic calibration. The comparison held the outer and inner cross-validation splits, tuned base-model parameters, recall constraint, and threshold-selection rule fixed, so that the calibration method was the only experimental factor varied. Calibration methods were compared using AUPRC, ROC-AUC, Brier score, log-loss, and policy-oriented precision, recall, and F1 across the same outer folds. Cross-validated calibrated probabilities were subsequently generated from the outer-training partition and used for policy-threshold optimization. No information from the corresponding outer-test fold was used for model fitting, calibration, or threshold selection.

For threshold optimization, candidate thresholds were derived from the unique calibrated training-side cross-validated probabilities rounded to six decimal places; therefore, the primary analysis did not impose an arbitrary fixed threshold step size. Thresholds satisfying the predefined policy constraint, $\text{Recall}_{\text{dropout}} \geq 0.80$, were considered feasible. Among the feasible candidates, the decision threshold t* was selected by maximizing precision; ties were resolved sequentially using the F1-score and then the larger threshold value. If no candidate threshold satisfied the recall constraint, the threshold maximizing F1 was used as a fallback. Sensitivity analysis repeated threshold selection using fixed candidate-grid steps of 0.001, 0.005, and 0.010 and two alternative secondary selection rules: F1-first and largest-feasible-threshold-first. Once selected, t* was fixed and applied to the calibrated probabilities of the corresponding held-out outer-test fold. Consequently, the outer-test data were not involved in model tuning, calibration, or policy-threshold selection.

The calibrated test-fold probabilities were additionally categorized into low, medium, high, and very high risk bands to support intervention triage and prioritization [23]. This separation between model estimation, calibration, threshold selection, and held-out evaluation was maintained to prevent information from the evaluation fold from influencing the decision policy. Algorithm 1 summarizes the implementation procedure.

Algorithm 1. Calibration-aware decision-support policy implementation

Input: training data $\mathcal{D}_{train}$, held-out test data $\mathcal{D}_{test}$, minimum recall target $\mathfrak{r}_0 = 0.80$, and predefined risk-band boundaries.

Output: calibrated dropout probabilities, intervention decisions, and risk-band assignments.

1. Tune the predictive model using only $\mathcal{D}_{train}$.
2. Fit the selected model and generate the training-side risk scores $\hat{p}$.
3. Fit the selected calibration procedure using training-side data and obtain calibrated scores p_{cal}; use sigmoid for the primary analysis and isotonic for the calibration-sensitivity analysis.
4. Evaluate candidate decision thresholds on the calibrated training-side scores.
5. Select t* among feasible thresholds by maximizing precision; break ties sequentially using the F1-score and then the larger threshold value. If no threshold satisfies $\text{Recall}_{\text{dropout}} \geq 0.80$, select the threshold maximizing F1 as a fallback.
6. Apply the trained model and calibration mapping to $\mathcal{D}_{test}$ to obtain $p_{\text{cal, test}}$.
7. Apply the fixed threshold t* to obtain the intervention decision.
8. Assign each $p_{\text{cal, test}}$ to the predefined low, medium, high, or very high risk band.
9. Evaluate predictive, probabilistic, and policy-oriented performance on $\mathcal{D}_{test}$.
10. Repeat the procedure across all outer folds and aggregate the results.

For the calibration-method sensitivity analysis, the calibration step was repeated using sigmoid and isotonic calibration while keeping the outer/inner data splits, tuned model parameters, recall constraint, and threshold-selection rule unchanged.

3.4. Explainability and insight factors

To support policy interpretation in line with explainable student-dropout modeling research [33], [34], two complementary analyses were conducted: permutation importance aggregated across outer folds to quantify global feature contribution, and ablation experiments to assess dependence on selected feature groups.

3.5. Reproducibility resources

The experimental workflow was implemented with random_state=42 and parallel execution (n_jobs=-1) using Python 3.11.7, NumPy 1.26.4, pandas 2.1.4, scikit-learn 1.7.1, imbalanced-learn 0.14.0, SciPy 1.11.4, matplotlib 3.8.0, and Jupyter Notebook 7.0.8 (JupyterLab 4.0.11; ipykernel 6.28.0). The evaluation workstation used Windows 11 build 10.0.26200, an Intel Core i7 processor with 10 physical and 12 logical cores, and 15.63 GB RAM. Fold-wise separation of model tuning, calibration, threshold selection, and held-out evaluation was maintained throughout the workflow. Algorithm 1 provides implementation-level pseudocode for the calibration and decision-policy procedure. To support independent replication, the experiment notebook, software requirements, calibration and threshold-selection routines, and derived non-identifiable result files are publicly available through the reproducibility repository specified in the data availability statement. The original student-level dataset is not redistributed by this study; its provenance and collection details are provided in the source publication [26].

4. RESULTS AND DISCUSSION

Performance is reported as mean±SD and 95% CI across the 10 outer test folds (5×2 repeated stratified CV); dropout represented 20.36% of the n=329 cohort.

4.1. Model performance (nested tuned) at default threshold 0.5

In the main nested tuned evaluation, the top two candidates showed almost identical performance in terms of AUPRC and ROC-AUC (Table 1). At the default threshold, M1_BRF achieved higher dropout recall, whereas M2_RF_SMOTENC produced a lower Brier score and slightly higher F1 and MCC. Because the proposed policy prioritizes missed-risk reduction, both models were subsequently evaluated under the recall-constrained threshold rule.

Table 1. Model performance after nested tuning at the default threshold of 0.5

Model ID	Setting	AUPRC	ROC-AUC	Recall	F1	MCC	Brier
M1_BRF	NestedTuned @0.5	0.7025±0.1024 (CI[0.6292,0.7758])	0.8750±0.0469 (CI[0.8415,0.9085])	0.7060±0.1092 (CI[0.6279,0.7842])	0.6022±0.0659 (CI[0.5550,0.6493])	0.4932±0.0850 (CI[0.4325,0.5540])	0.1340±0.0168 (CI[0.1220,0.1460])
M2_RF_SMOTENC	NestedTuned @0.5	0.7019±0.1046 (CI[0.6271,0.7767])	0.8758±0.0510 (CI[0.8393,0.9123])	0.6264±0.1190 (CI[0.5412,0.7115])	0.6148±0.0812 (CI[0.5567,0.6729])	0.5201±0.0988 (CI[0.4494,0.5908])	0.1112±0.0214 (CI[0.0958,0.1265])

AUPRC; ROC-AUC; MCC: Matthews correlation coefficient; Brier: Brier score; CI: 95% confidence interval.

4.2. Threshold tuning policy with recall constraint ≥0.80 (without calibration)

The decision threshold is set through threshold tuning to meet the policy target of recall dropout ≥0.80 (Table 2).

Table 2. Performance after recall-constrained threshold tuning without calibration

Model ID	Setting	AUPRC	ROC-AUC	Recall	F1	MCC	Brier	Threshold
M1_BRF	Threshold Tuned (Recall≥0.80)	0.7025±0.1024 (CI[0.6292, 0.7758])	0.8750±0.0469 (CI[0.8415, 0.9085])	0.8264±0.1100 (CI[0.7477, 0.9051])	0.5977±0.0709 (CI[0.5469, 0.6484])	0.4940±0.0988 (CI[0.4233, 0.5646])	0.1340±0.0168 (CI[0.1220, 0.1460])	0.4169±0.0423 (CI[0.3866, 0.4471])
M2_RF_SMOTENC	Threshold Tuned (Recall ≥0.80)	0.7019±0.1046 (CI[0.6271, 0.7767])	0.8758±0.0510 (CI[0.8393, 0.9123])	0.8115±0.0981 (CI[0.7414, 0.8817])	0.5937±0.0832 (CI[0.5342, 0.6532])	0.4857±0.1140 (CI[0.4042, 0.5673])	0.1112±0.0214 (CI[0.0958, 0.1265])	0.2967±0.0502 (CI[0.2608, 0.3327])

Threshold: selected decision threshold.

Lowering the decision threshold increased sensitivity at the cost of intervention volume. Across the 10 outer test folds, corresponding to 658 repeated held-out prediction instances, M1_BRF changed from the true-negative/false-positive/false-negative/true-positive counts (TN/FP/FN/TP)=439/85/39/95 at the default threshold to 398/126/23/111 under the recall-constrained policy; M2_RF_SMOTENC changed from 470/54/50/84 to 397/127/25/109. For M1_BRF, policy tuning therefore captured 16 additional dropout prediction instances while generating 41 additional false-positive flags. The number of flagged instances increased from 180 to 237, corresponding to approximately 1.89 versus 2.14 intervention flags per detected dropout. Because the evaluation used 5×2 repeated cross-validation, these totals represent repeated held-out prediction instances rather than 658 unique students.

4.3. Calibration-method sensitivity and recall-constrained policy

Table 3 compares sigmoid and isotonic calibration for M1_BRF under the same recall-constrained policy (minimum dropout recall ≥0.80). Both calibration methods were evaluated using identical outer/inner data splits, tuned model parameters, and threshold-selection rules to isolate the effect of the calibration procedure [13], [14], [19].

Table 3. Comparison of calibration methods under recall-constrained threshold tuning

Calibration	AUPRC	Brier	Log loss	Precision	Recall	F1	Threshold
Sigmoid	**0.6952± 0.1080**	**0.1040± 0.0108**	**0.3379± 0.0276**	**0.4630± 0.0722**	0.8341± 0.1011	**0.5926± 0.0726**	0.1841± 0.0269
Isotonic	0.6880± 0.0937	0.1041± 0.0163	0.3854± 0.1424	0.4323± 0.0736	**0.8797± 0.0960**	0.5757± 0.0729	0.1570± 0.0369

AUPRC; F1: harmonic mean of precision and recall; and threshold: mean selected t* across outer folds.

The calibration-method sensitivity analysis showed similar ranking and Brier score performance for sigmoid and isotonic calibration. Sigmoid achieved AUPRC=0.6952, Brier=0.1040, and log-loss=0.3379±0.0276, whereas isotonic achieved 0.6880, 0.1041, and 0.3854±0.1424, respectively. Both methods exceeded the predefined minimum recall target of 0.80; isotonic yielded higher recall (0.8797 vs. 0.8341), whereas sigmoid provided higher precision (0.4630 vs. 0.4323) and F1 (0.5926 vs. 0.5757). Paired comparisons did not indicate significant differences in the principal predictive or reliability metrics $p > 0.050$; sigmoid was therefore retained because it provided a more balanced policy trade-off and substantially more stable log-loss. Relative to the uncalibrated M1_BRF, sigmoid calibration reduced the mean Brier score from 0.1340 to 0.1040 (22.4%), log-loss from 0.4210 to 0.3379 (19.8%), uniform-bin ECE from 0.1741 to 0.0856 (50.8%), and quantile-bin ECE from 0.1692 to 0.0933 (44.8%). The calibrated uniform-bin ECE was 0.0856±0.0171 (95% CI: 0.0734–0.0978). Threshold-sensitivity analysis further produced mean t* values between 0.1782 and 0.1861, held-out recall between 0.8264 and 0.8341, F1 between 0.5856 and 0.5940, and 99.09–99.85% decision agreement with the primary policy. Thus, downstream intervention decisions were largely insensitive to reasonable variations in candidate-threshold granularity and secondary tie-breaking rules. Figure 1 visually confirms the improvement in probability alignment after sigmoid calibration. The diagonal indicates perfect probability calibration.

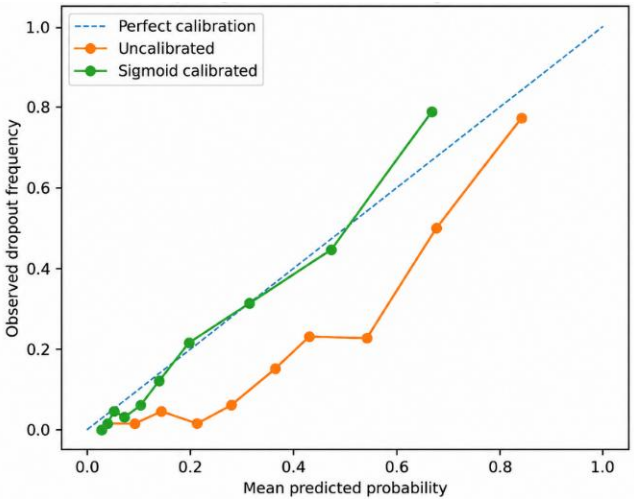


Figure 1. Reliability diagram of M1_BRF before and after sigmoid calibration using 10 quantile-based bins

4.4. Illustrative risk-band mapping from the screening baseline

To illustrate how fixed probability bands can be translated into an operational triage scheme, Table 4 presents one-pass out-of-fold probabilities from the best screening baseline RF with SMOTE-NC (RF+SMOTE-NC). This analysis is provided as an illustrative risk-band mapping and is distinct from the calibration-aware policy evaluation reported in subsection 4.3. Accordingly, Table 4 is intended to demonstrate the ordering and operational interpretation of the predefined risk bands rather than to re-estimate the performance of the calibrated policy.

Table 4. Distribution and observed dropout rate across illustrative screening-baseline risk bands

Risk band	N	Dropout rate (%)
Low (≤0.2)	192	5.21
Medium (0.2-0.5)	62	16.13
High (0.5-0.8)	39	51.28
Very high (>0.8)	36	75.00

A consistent pattern of escalating dropout rates throughout bands is shown, rendering risk bands an appropriate triage system. Operationally, a pragmatic action mapping may be delineated as follows: low=passive prevention; medium=monitoring/nudges; high=proactive outreach; and very high=extensive multi-stakeholder engagement. Within the framework of this article, risk bands function as a conduit between risk scores and service priorities, as well as a mechanism for managing intervention burden. The observed dropout prevalence increased monotonically from the low to the very high band (5.21%, 16.13%, 51.28%, and 75.00%, respectively), supporting the use of the four categories as an interpretable triage mechanism. These numerical boundaries should, however, be regarded as operational cut-offs rather than universal constants and should be revalidated when transferred to other cohorts or institutions.

4.5. Ablation study (M1_BRF)

The ablation study evaluated the sensitivity of model performance to key feature groups. Table 5 summarizes AUPRC, ROC-AUC, recall, F1, MCC, and Brier score for the full and ablated feature configurations.

Table 5. Ablation analysis of feature-group contributions

Ablation	AUPRC	ROC-AUC	Recall	F1	MCC	Brier
Full feature set	0.7037±0.1000 (CI[0.6322,0.7752])	0.8729±0.0475 (CI[0.8390,0.9069])	0.7066±0.0950 (CI[0.6386,0.7746])	0.6007±0.0518 (CI[0.5636,0.6377])	0.4904±0.0688 (CI[0.4411,0.5396])	0.1357±0.0186 (CI[0.1224,0.1490])
Without CGPA	0.6266±0.0867 (CI[0.5646,0.6887])	0.8272±0.0437 (CI[0.7959,0.8584])	0.6390±0.1256 (CI[0.5492,0.7289])	0.5408±0.0552 (CI[0.5013,0.5803])	0.4136±0.0753 (CI[0.3598,0.4675])	0.1624±0.0137 (CI[0.1526,0.1722])
Without perceptual/ relational features	0.5513±0.0908 (CI[0.4863,0.6162])	0.7880±0.0515 (CI[0.7511,0.8248])	0.6225±0.1560 (CI[0.5110,0.7341])	0.4877±0.0659 (CI[0.4406,0.5349])	0.3442±0.0799 (CI[0.2870,0.4014])	0.1834±0.0180 (CI[0.1706,0.1963])
CGPA only	0.4264±0.0843 (CI[0.3661,0.4867])	0.7495±0.0455 (CI[0.7169,0.7820])	0.8412±0.2197 (CI[0.6840,0.9984])	0.4271±0.0593 (CI[0.3847,0.4695])	0.2748±0.0779 (CI[0.2191,0.3306])	0.1916±0.0232 (CI[0.1751,0.2082])
Perceptual/ relational features only	0.5942±0.0903 (CI[0.5296,0.6589])	0.7768±0.0627 (CI[0.7320,0.8217])	0.6549±0.1514 (CI[0.5466,0.7633])	0.5452±0.0809 (CI[0.4873,0.6031])	0.4207±0.1028 (CI[0.3472,0.4943])	0.1789±0.0077 (CI[0.1735,0.1844])

AUPRC; ROC-AUC; MCC: Matthews correlation coefficient; brier: brier score; and CI: 95% confidence interval.

The ablation analysis also reveals an operational workload trade-off. At the default 0.50 threshold, CGPA alone achieved high recall (0.8412) but mean precision of only 0.3056, corresponding to a false-discovery share of 69.4% among positive flags. In comparison, the full multidomain feature set achieved a precision of 0.5247 and reduced the false-discovery rate to 47.5%, a reduction of 21.9 percentage points. Thus, combining academic and non-academic information sharpened intervention targeting and may reduce unnecessary advisor outreach while retaining broader risk information. Advisor workload or burnout, however, was not directly measured in this retrospective study.

4.6. Significance test

To determine whether the observed performance differences between M1_BRF and M2_RF_SMOTENC were statistically supported across the outer folds, a Wilcoxon signed-rank test was applied to the paired per-fold AUPRC and recall scores. The results are summarized in Table 6.

Table 6. Wilcoxon signed-rank comparison with paired effect sizes

Metric	Model 1	Model 2	W	p-value	Rank-biserial r_{rb}
AUPRC	M1_BRF	M2_RF_SMOTENC	25.0	0.846	0.091
Recall	M1_BRF	M2_RF_SMOTENC	11.5	0.131	0.582

W: Wilcoxon signed-rank statistic and r_{rb}: paired rank-biserial correlation.

As shown in Table 6, neither the AUPRC difference (p=0.846) nor the recall difference (p=0.131) reached statistical significance at α=0.05. The paired rank-biserial effect sizes, however, provide additional information about the magnitude and direction of these differences. The AUPRC effect was very small (r_{rb}=0.091), consistent with the nearly identical mean AUPRC values of the two models. In contrast, recall showed a larger positive paired effect (r_{rb}=0.582) in favor of M1_BRF, despite the non-significant p-value. Because missed-risk sensitivity is central to the predefined recall-constrained policy, this effect-size evidence further supports retaining M1_BRF as the reference model for calibration, ablation, and explainability analyses.

4.7. The most important factors

Because the dataset contains only eight input features, Table 7 reports the outer-fold aggregated permutation importance for all predictors.

Table 7. Permutation importance of the eight input features for the selected model

Feature	Mean importance	SD importance
Program-study interest	0.2131	0.0794
CGPA	0.1769	0.0742
Individual income	0.0367	0.0320
Lecturer-quality satisfaction	0.0217	0.0375
Relationship lecturer supervisor	0.0188	0.0408
Marital status	0.0167	0.0627
Employment status	0.0105	0.0166
Parent income	0.0014	0.0205

The consistency with the ablation analysis indicates that program-study interest and CGPA are the dominant factors, whereas lecturer-relationship and lecturer-quality indicators provide complementary information. Because program-study interest and lecturer-related satisfaction are perceptual or self-reported variables, their institutional collection requires additional safeguards. Institutions should communicate the supportive purpose of data collection, use low-stakes and non-punitive survey procedures, minimize unnecessary identifiable information, and monitor response and non-response patterns. These indicators should complement rather than independently determine intervention decisions because selective participation, socially desirable responses, or student distrust may introduce survey-response bias.

4.8. Technical comparison and policy implications

From a technical perspective, the results separate three properties that are often conflated in dropout prediction: discrimination, probability reliability, and operating-point selection. M1_BRF and M2_RF_SMOTENC produced nearly identical ranking performance (AUPRC≈0.70; ROC-AUC≈0.875), with no significant per-fold difference. This contrasts with prediction-oriented studies that primarily emphasize ranking or classification performance [5], [10]. For M1_BRF, sigmoid calibration reduced the Brier score from 0.1340 to 0.1040, while recall-constrained threshold tuning achieved a dropout recall of 0.8341 at an average calibrated threshold of 0.1841. These results support treating discrimination, calibration, and threshold selection as separate technical components rather than treating 0.50 as an intrinsic operating point.

The policy layer extends model outputs toward operational triage. Early-warning intervention studies link risk identification to student support [7], [8], while explainable dropout studies emphasize interpretable model behavior [33], [34]. In this study, the screening-baseline bands showed an ordered increase in dropout prevalence from low to very high, while ablation and permutation analyses indicated that

the policy signal was not attributable to CGPA alone. Thus, the contribution is not a claim of universal numerical superiority across datasets, but an evaluation of the complete chain from predictive ranking to calibrated risk, constrained threshold selection, risk stratification, and interpretable support.

Table 8 shows that the distinction of the proposed framework lies primarily in policy integration rather than cross-dataset predictive superiority. Existing systems provide important complementary capabilities, including intervention delivery, explainable personalized guidance, or calibrated probabilities. The present framework specifically combines calibration-method evaluation with an explicit training-side recall constraint, transparent threshold-selection logic, operational risk-band mapping, and reproducibility resources within one validated policy pipeline. Because the compared studies differ in dataset composition, prediction horizon, and validation design, their reported predictive scores are not interpreted as directly comparable numerical benchmarks. The comparison is therefore made at the decision-support architecture and methodological-control level.

Table 8. Comparative taxonomy of recent dropout early-warning and decision-support frameworks

Study/framework	Main decision-support focus	Probability calibration	Explicit operating policy	Risk/intervention mapping	Validation/reproducibility orientation
Bañeres *et al.* [8]	Dynamic course-dropout identification and intervention	Not a central reported component	Dynamic temporal-window logic	Risk levels linked to intervention messages	Course-specific validation and intervention pilot
Nagy and Molontay [11]	Calibrated, explainable dropout prediction and personalized intervention	Platt/sigmoid and isotonic	No explicit recall-constrained policy reported	XAI-supported personalized guidance	Temporal cohort-based evaluation
Alalawi *et al.* [9]	Prediction-to-action learning analytics framework	Not a central reported component	Course-specific risk identification	Knowledge-gap and personalized intervention support	Applied intervention framework; open-source implementation
Carballo-Mendívil *et al.* [12]	Calibrated pre-enrollment dropout early warning	Isotonic	False-negative/recall-sensitive operational orientation	Risk identification for early support	Large historical institutional dataset
Present study	Calibration-aware, capacity-oriented decision policy	Sigmoid vs. isotonic sensitivity; sigmoid retained	Training-side $Recall_{dropout} \geq 0.80$ constraint with explicit tie-breaking	Four risk bands linked to service priority	Nested CV, CI, effect size, public reproducibility package

XAI: explainable artificial intelligence; CV: cross-validation; and CI: confidence interval.

The engineering significance of the framework arises from the separation of four controllable modules: prediction, probability calibration, operating-threshold optimization, and service-level assignment. This separation improves auditability because an institution can identify whether a change in intervention volume originates from model discrimination, probability remapping, the policy threshold, or the risk-band definition. It also improves maintainability: calibration and t* can be re-estimated without redesigning the underlying intervention structure, while the predictive model can be replaced without changing the policy interface. The public implementation resources further make these design choices inspectable and reproducible.

Operationally, the framework should function as a human-supported decision mechanism. Institutions should document the model, calibration mapping, threshold, and intervention rules, while maintaining appropriate human oversight and accountability when high-risk classifications are translated into student-support actions [15], [35]. For institutional deployment, the framework can be implemented as a batch decision-support layer over periodically updated student records, with the calibrated probability, decision threshold, risk band, and recommended service level stored as auditable outputs. The measured runtime indicates that the current workflow is computationally manageable for offline model development at the evaluated scale; however, scalability to substantially larger, multi-program, or multi-institutional data streams has not yet been empirically established. Such deployment should therefore be accompanied by monitoring of intervention volume, calibration quality, processing time, and changes in cohort composition. Accordingly, the practical value of the framework lies in providing institutions with a transparent mechanism for aligning predictive risk, intervention capacity, and accountable student-support decisions rather than using model outputs as automated determinations.

The framework's computational burden is concentrated in offline model development. On the evaluation environment described in subsection 3.5, nested model selection required 330.54 s (5.51 min),

while sigmoid calibration and recall-constrained policy construction required 94.15 s (1.57 min), totaling approximately 7.08 min for the two core offline stages. These times characterize offline evaluation and policy construction rather than per-student inference latency.

4.9. Limitations and threats to validity

The primary constraint of this study pertains to external validity. The dataset was limited in size (n=329) and sourced from a singular institutional environment; hence, the identified risk patterns may be indicative of local demographic traits, curricular frameworks, and academic methodologies. Moreover, the inclusion of both perceptual/relational indicators and ordinal categories introduces the possibility of measurement bias (e.g., variations in respondents' scale ratings) and selection bias (e.g., respondents who failed to complete the questionnaire possess different characteristics compared to those who did). Consequently, the findings should be interpreted as proof-of-concept evidence for the proposed decision-support procedure rather than as evidence of uniform effectiveness across other institutional settings.

Regarding internal validity and statistical conclusion validity, bias may arise if validation procedures do not prevent information leakage. This study mitigated this risk through nested cross-validation and strict separation between model development and held-out evaluation. Nonetheless, implementations beyond research must adhere to the same principle, with resampling, calibration, and other training-dependent operations performed exclusively within the training data. The definition of the label (Dropout vs. Graduate) and the observation period affect the interpretation of the policy; if the dropout label is assigned at varying time horizons, the threshold and risk bands may fluctuate. Consequently, future studies should perform temporal validation (e.g., training on an initial cohort/semester and testing on a subsequent period) and external validation across institutions, while assessing the stability of the calibration and threshold policy under drift [36].

Post-deployment governance should include predefined institutional monitoring intervals, for example once per academic term. Monitoring should track rolling Brier score, log-loss/ECE, intervention volume, and feature-distribution drift using an indicator such as the population stability index (PSI). Material deterioration relative to an institution-defined reference period should trigger review of the calibration mapping and re-estimation of t^*; persistent deterioration in discrimination or feature distributions may justify full model retraining. Subgroup calibration and intervention burden should also be audited to identify uneven operational effects. Future multi-institution studies should test whether calibration mappings, threshold policies, and risk-band workloads remain transferable across institutions. Where raw student-level data cannot be centrally pooled, privacy-preserving or federated-learning approaches should be evaluated. These monitoring and transfer mechanisms constitute proposed future deployment procedures and were not empirically evaluated in the present retrospective study.

5. CONCLUSION

This study demonstrates that dropout prediction becomes more operationally useful when predictive discrimination, probability reliability, decision-threshold selection, and intervention priority are treated as separate and auditable components. The principal reusable contribution is a human-supported policy layer that translates calibrated dropout probabilities into capacity-aware intervention priorities under an explicit minimum-recall objective. The resulting procedure is model-agnostic and can be locally recalibrated, re-thresholded, and revalidated without treating model outputs as automated institutional determinations.

The findings remain constrained by the small single-institution sample, retrospective design, and locally defined predictors and risk-band boundaries. Future research should therefore address three concrete questions: i) how stable are calibration mappings, operating thresholds, and intervention workloads across cohorts and institutions; ii) which privacy-preserving multi-institutional strategies can preserve calibration and policy utility without transferring raw student records; and iii) which post-deployment drift signals should trigger recalibration, threshold re-estimation, or complete model retraining? Addressing these questions is necessary to move the framework from retrospective proof-of-concept evaluation toward sustainable and accountable institutional deployment.

ACKNOWLEDGMENTS

The authors acknowledge that this study was conducted without a specific research grant or contract from any funding agency in the public, commercial, or not-for-profit sectors.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Wijiyanto	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Aris Marjuni	✓	✓		✓		✓			✓	✓		✓		
Ahmad Zainul Fanani		✓	✓	✓			✓			✓	✓	✓		
Ruri Suko Basuki			✓	✓		✓	✓			✓	✓	✓		
Nurmalitasari					✓	✓	✓	✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The dataset analyzed in this study was originally collected and reported in [26]; the study context, data-collection procedure, and source information are provided in the original publication. The experiment notebook, software requirements, calibration and threshold-selection routines, and derived non-identifiable result files required to reproduce the analyses are publicly available in the accompanying reproducibility repository: https://github.com/wiji-sys/BEEI_12300_reproducibility-. The original student-level dataset is not redistributed by the present study.

REFERENCES

[1] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technology in Society*, vol. 76, pp. 1–10, Mar. 2024, doi: 10.1016/j.techsoc.2024.102474.

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

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

[4] Y. Liu, W. Wang, and E. Xu, "The Effectiveness of Learning Analytics-Based Interventions in Enhancing Students' Learning Effect: A Meta-Analysis of Empirical Studies," *SAGE Open*, vol. 15, no. 2, p. 21582440251336707, Apr. 2025, doi: 10.1177/21582440251336707.

[5] J. A. Talamás-Carvajal and H. G. Ceballos, "A stacking ensemble machine learning method for early identification of students at risk of dropout," *Education and Information Technologies*, vol. 28, no. 9, pp. 12169–12189, Sep. 2023, doi: 10.1007/s10639-023-11682-z.

[6] B. B. Alkan, S. Kuzucuk, N. Alkan, and A. Sinan, "Using machine learning to predict student outcomes for early intervention and formative assessment," *Scientific Reports*, vol. 15, no. 1, p. 39797, Nov. 2025, doi: 10.1038/s41598-025-23409-w.

[7] K. Alalawi, R. Athauda, R. Chiong, and I. Renner, "Evaluating the student performance prediction and action framework through a learning analytics intervention study," *Education and Information Technologies*, vol. 30, no. 3, pp. 2887–2916, Feb. 2025, doi: 10.1007/s10639-024-12923-5.

[8] D. Bañeres, M. E. Rodríguez-González, A.-E. Guerrero-Roldán, and P. Cortadas, "An early warning system to identify and intervene online dropout learners," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, pp. 1–25, Jan. 2023, doi: 10.1186/s41239-022-00371-5.

[9] K. Alalawi, R. Athauda, and R. Chiong, "An Extended Learning Analytics Framework Integrating Machine Learning and Pedagogical Approaches for Student Performance Prediction and Intervention," *International Journal of Artificial Intelligence in Education*, vol. 35, no. 3, pp. 1239–1287, Sep. 2025, doi: 10.1007/s40593-024-00429-7.

[10] M. R. Marcolino *et al.*, "Student dropout prediction through machine learning optimization: insights from moodle log data," *Scientific Reports*, vol. 15, no. 1, p. 9840, Mar. 2025, doi: 10.1038/s41598-025-93918-1.

[11] M. Nagy and R. Molontay, "Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention," *International Journal of Artificial Intelligence in Education*, vol. 34, no. 2, pp. 274–300, Jun. 2024, doi: 10.1007/s40593-023-00331-8.

[12] B. Carballo-Mendívil, A. J. Pérez-Morales, A. Arellano-González, M. D. P. Lizardi-Duarte, and N. J. Ríos-Vázquez, "Predicting Student Dropout from Pre-Enrollment Data in Mexican Higher Education: A Theoretically Grounded and Calibrated Machine Learning Approach," *Education Sciences*, vol. 16, no. 8, pp. 1–27, Aug. 2026, doi: 10.3390/educsci16081216.

[13] T. S. 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," *Machine Learning*, vol. 112, no. 9, pp. 3211–3260, Sep. 2023, doi: 10.1007/s10994-023-06336-7.

[14] 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.
[15] R. Alfredo *et al.*, "Human-centred learning analytics and AI in education: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 6, pp. 1–23, Jun. 2024, doi: 10.1016/j.caeai.2024.100215.
[16] 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, pp. 1–11, Jun. 2025, doi: 10.1016/j.caeai.2025.100369.
[17] J. C-Rella, R. Cao, and J. M. Vilar, "Cost-sensitive thresholding over a two-dimensional decision region for fraud detection," *Information Sciences*, vol. 657, pp. 1–15, Feb. 2024, doi: 10.1016/j.ins.2023.119956.
[18] F. Cabezas, L. Burgos, J. Darrigol, and M. Zúñiga, "Implementation of an early alert system in quizzes of a high complexity subject in higher education: Improvement of student performance and teacher perception," *Education and Information Technologies*, vol. 29, no. 15, pp. 19321–19341, Oct. 2024, doi: 10.1007/s10639-024-12610-5.
[19] M. Yang and X. Bi, "Cost-Aware Calibration of Classifiers," *INFORMS Journal on Data Science*, vol. 4, no. 2, pp. 101–113, Apr. 2025, doi: 10.1287/ijds.2024.0038.
[20] K. Chalkou, A. J. Vickers, F. Pellegrini, A. Manca, and G. Salanti, "Decision Curve Analysis for Personalized Treatment Choice between Multiple Options," *Medical Decision Making*, vol. 43, no. 3, pp. 337–349, Apr. 2023, doi: 10.1177/0272989X221143058.
[21] I. Hozo, G. Guyatt, and B. Djulbegovic, "Decision curve analysis based on summary data," *Journal of Evaluation in Clinical Practice*, vol. 30, no. 2, pp. 281–289, Mar. 2024, doi: 10.1111/jep.13945.
[22] N. K. Motlagh, J. Davis, T. Anderson, and J. Gwinnup, "Naturally constrained reject option classification," *Machine Vision and Applications*, vol. 36, no. 1, p. 9, Jan. 2025, doi: 10.1007/s00138-024-01620-5.
[23] P. Putrik, I. Kant, H. Hoofs, R. Reijs, and M. Jansen, "Prediction of School Dropout Outside School Setting: Potential for Early risk Stratification by Youth Health Care Services in the Netherlands. Results from a Retrospective Cohort Study," *Child & Youth Care Forum*, vol. 53, no. 2, pp. 349–365, Apr. 2024, doi: 10.1007/s10566-023-09757-6.
[24] W. Dai *et al.*, "Learning Analytics for Early Identification of At-Risk Students and Feedback Intervention," *Journal of Learning Analytics*, vol. 12, no. 3, pp. 102–125, Nov. 2025, doi: 10.18608/jla.2025.8735.
[25] J. Chen, "Effects of learning analytics-based feedback on students' self-regulated learning and academic achievement in a blended EFL course," *System*, vol. 124, p. 103388, Aug. 2024, doi: 10.1016/j.system.2024.103388.
[26] Nurmalitasari, Z. A. Long, and M. F. M. Noor, "Factors Influencing Dropout Students in Higher Education," *Education Research International*, vol. 2023, pp. 1–13, Feb. 2023, doi: 10.1155/2023/7704142.
[27] G. I. Austin, I. Pe'er, and T. Korem, "Distributional bias compromises leave-one-out cross-validation," *Science Advances*, vol. 11, no. 48, p. eadx6976, Nov. 2025, doi: 10.1126/sciadv.adx6976.
[28] D. Wilimitis and C. G. Walsh, "Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial," *JMIR AI*, vol. 2, p. e49023, Dec. 2023, doi: 10.2196/49023.
[29] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiology: Artificial Intelligence*, vol. 5, no. 4, p. e220232, Jul. 2023, doi: 10.1148/ryai.220232.
[30] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, pp. 1–12, Jun. 2024, doi: 10.1016/j.patter.2024.100994.
[31] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*, 1st ed. in Adaptive Computation and Machine Learning Series. Cambridge: MIT Press, 2023.
[32] B. R. Sziklai, M. Baranyi, and K. Héberger, "Does cross-validation work in telling rankings apart?," *Central European Journal of Operations Research*, vol. 33, no. 4, pp. 1503–1528, Dec. 2025, doi: 10.1007/s10100-024-00932-1.
[33] 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, p. 120933, Dec. 2023, doi: 10.1016/j.eswa.2023.120933.
[34] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, "Balancing Performance and Explainability in Academic Dropout Prediction," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 2086–2099, 2024, doi: 10.1109/TLT.2024.3425959.
[35] B. Memarian and T. Doleck, "Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100152, 2023, doi: 10.1016/j.caeai.2023.100152.
[36] E. Tiukhova *et al.*, "Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI," *Decision Support Systems*, vol. 182, p. 114229, Jul. 2024, doi: 10.1016/j.dss.2024.114229.

BIOGRAPHIES OF AUTHORS



Wijiyanto is a lecturer at the Faculty of Computer Science at the Universitas Duta Bangsa (UDB), Surakarta, Indonesia, and a doctoral graduate in computer science from Dian Nuswantoro University (UDINUS) in Semarang, Indonesia. His research interests include applied machine learning, artificial intelligence, data mining, educational data analytics, and robust model evaluation. He can be contacted at email: wijiyanto@udb.ac.id.



Aris Marjuni is a full professor at Dian Nuswantoro University (UDINUS), Semarang, Indonesia. He serves as the head of the doctoral program in Computer Science (S3) at UDINUS. His expertise and research interests include image processing/computer vision, data mining, and software engineering (including soft computing and information systems). He is affiliated with the information systems program and actively contributes to teaching and research in the above fields. He can be contacted at email: aris.marjuni@dsn.dinus.ac.id.



Ahmad Zainul Fanani received a bachelor's and master's degree in 1999 and 2001, respectively, from the Department of Informatics Technology, STTIBI, Jakarta, Indonesia. He is a doctoral graduate from the Department of Electrical Engineering of Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia. He works as a lecturer and full professor in the Faculty of Computer Science at Dian Nuswantoro University (UDINUS), Semarang, Indonesia. His research interests include computer vision, data mining, game, and multimedia design. He can be contacted at email: a.zainul.fanani@dsn.dinus.ac.id.



Ruri Suko Basuki is a senior lecturer (lektor kepala) at Dian Nuswantoro University (UDINUS), Semarang, Indonesia. He currently serves as head of the D-4 Film and Television Study Program at UDINUS. His research interests include computer vision, artificial intelligence, and soft computing. He teaches courses such as artificial intelligence and other research/technology-related subjects. He can be contacted at email: ruri.basuki@dsn.dinus.ac.id.



Nurmalitasari is a lecturer in the Information Systems discipline at Universitas Duta Bangsa Surakarta, Indonesia. She completed her doctoral (S3) studies in the Doctor of Philosophy (Ph.D.) program at Universiti Kuala Lumpur (UniKL), Malaysia. Her research interests cover machine learning, forecasting, modeling, and predictive analytics, with applications in educational, machine learning, and institutional analytics. She has contributed to peer-reviewed work on predictive learning analytics and factors contributing to student dropout in higher education. She can be contacted at email: nurmalitasari@udb.ac.id.