Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 3, No. 2, [December 2024], pp. [107–116]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Cost-Aware Feature Selection and Bayesian-Optimized Random Forest for Early Identification of At-Risk Students

Septiana Vratiwi[1], Lidya Wati[2], Tutut Herawan[3], Agus Perdana Windarto[4*]
[1] *Faculty of Intelligent Computing, Universiti Malaysia Perlis (UniMAP), Malaysia*
[2] *Faculty of Intelligent Computing, Universiti Malaysia Perlis (UniMAP), Malaysia*
[3] *Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia*
[4] *Department of Informatics, Master's Program, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia*

[1] *septiana@studentmail.unimap.edu.my;* [2] *lidyawati@studentmail.unimap.edu.my;* [3] *tutut@um.edu.my;* [4] *agus.perdana@amiktunasbangsa.ac.id*
** Corresponding author*

Abstract

Early identification of academically at-risk students is important for allocating support before course outcomes become final, but optimizing predictive performance may conflict with measurement burden and future-cohort generalization. This study assessed cost-aware feature selection coupled with Bayesian-optimized random forests on an early-window version of the Open University Learning Analytics Dataset. Enrollments eligible by day 14 were divided temporally into training (10,631), development (6,036), and 2014J holdout (9,330), with repeated learners purged from earlier partitions where required. Average precision (AP) was the primary outcome, and feature-group cost represented assumed acquisition burden rather than monetary expenditure. The proposed joint model achieved AP = 0.6007 using 12 features and a cost score of 6, compared with AP = 0.5991 for a 16-feature full random forest (cost 7), AP = 0.6081 for Bayesian-only tuning, and AP = 0.6092 for randomized hyperparameter search. Against randomized tuning, the proposed method's AP difference was −0.0085, with a descriptive paired-group 95% interval of approximately −0.0140 to −0.0031. Its at-risk recall (0.4772) was lower than the full-feature baseline (0.5199). Exploratory histogram-gradient-boosting models attained AP up to 0.6168 but did not validate the random-forest proposal. Thus, the pilot supports a limited feature-burden reduction while rejecting a superiority claim for the proposed joint search; repeated cohort and institution-level evaluations are required.

Keywords: Academic risk prediction; Random forest; Cost-aware selection; Temporal validation; Average precision

1. Introduction

The growing demand for defensible automated decisions in student academic risk has made early identification of students who may need timely academic support an important research problem, because errors can propagate into resource allocation, screening priorities, and operational decisions [1]. The task is not reducible to the maximization of aggregate accuracy, since class prevalence variation, demographic proxies, intervention timing, and label leakage can systematically alter the observed error distribution and the estimated value of a model [2]. Earlier investigators have approached related prediction problems through conventional discriminative learning, ensemble classifiers, engineered representations, and increasingly sophisticated neural architectures, thereby demonstrating the practical breadth of contemporary machine learning [1]. Across these approaches, representation quality matters because useful class distinctions must be retained without allowing nuisance variation or incidental measurement conditions to dominate the fitted decision boundary [1]. Studies examining model selection also show that hyperparameter search can improve the use of a candidate estimator, yet optimization outcomes depend heavily on search ranges, evaluation folds, and computational budgets [3]. Investigations of structured or spatially dependent observations warn that an apparently excellent random-split score may reflect dependence between training and test samples rather than out-of-sample transferability [2]. Work on probability reliability further distinguishes correct ranking from reliable decision confidence, making it necessary to consider threshold selection, calibration, and the

consequences of overconfident predictions [4]. Application-specific experiments have therefore made valuable progress in feature representation, architecture engineering, and supervised optimization, but each methodological advance addresses only part of the evaluation problem [1]. For student academic risk, the central difficulty is intensified when training data are heterogeneous or scarce, because high-capacity estimators can exploit shortcuts that remain invisible in an ordinary average-performance report [1]. Compact and interpretable methods offer an attractive alternative to overparameterized systems, although reducing parameter count or predictor dimensionality can weaken sensitivity to infrequent but consequential patterns [1]. Metaheuristic and Bayesian optimization approaches provide systematic ways to explore competing design choices, but improvements reported on a single split do not demonstrate statistically stable superiority [3]. The broader literature consequently supports evaluating both methodological fit and deployment-relevant constraints, especially when model selection is performed on small datasets with dependent observations [3]. These considerations are directly relevant to the title proposed here, which combines use cost-aware feature subset selection and Bayesian random forest tuning for student risk identification rather than treating prediction, representation, and optimization as unrelated components [3]. However, the precise interaction among these components cannot be inferred from studies that benchmark isolated backbones, optimizer families, or feature extraction schemes under different experimental protocols. Addressing that gap requires a comparative design in which model capacity, data partitions, annotation provenance, and all preprocessing decisions remain explicit, rather than relying on headline accuracy alone [2].

The methodological gap becomes more consequential when use cost-aware feature subset selection and Bayesian random forest tuning for student risk identification is implemented without jointly specifying the learning objective, search space, resampling strategy, and conditions under which the resulting system will be evaluated [2]. A first technical concern is that flexible tuning procedures may repeatedly exploit validation noise, selecting designs that fit a particular resampling realization while offering no reliable advantage on genuinely independent observations [2]. This selection bias is especially important in student academic risk because class prevalence variation, demographic proxies, intervention timing, and label leakage can generate hidden dependencies that are not removed by superficial class stratification [2]. Where related measurements originate from one patient, device, location, or time window, random partitioning may inflate the apparent gains of both the baseline learner and the optimized alternative [2]. A second concern is that maximizing an undifferentiated metric can neglect low-prevalence outcomes, even when the application places the greatest practical value on detecting the least common or most costly errors [1]. Changing sample weights or introducing synthetic observations may improve empirical class balance, but these operations must occur strictly within the training folds to avoid contaminating held-out evidence ([2]. A third issue concerns calibration: a classifier can assign suitable labels while producing probabilities that are unsuitable for risk thresholds, referrals, or resource-limited decision processes ([4]. Calibration measures should therefore be assessed alongside discrimination, with threshold choice separated from unbiased model assessment so that a reported improvement reflects a coherent operational objective [4]. Further, representation refinements and optimizer changes may interact nonlinearly; testing them only as a complete package prevents a clear account of which component is responsible for the observed gain ([3]. Such interactions are particularly difficult to interpret when architectural changes simultaneously alter parameter counts, training schedules, augmentation diversity, or the number of examined configurations [1]. In resource-constrained settings, additional optimization may also undermine practical applicability if an ostensibly efficient prediction network demands excessive search time or expensive training hardware ([3]. Consequently, inference latency and memory consumption must be measured on stated hardware, rather than estimated solely from layer counts or surrogate computational proxies [1]. The interpretation of student academic risk results also depends on the experimental unit: predictions on correlated images or records should not be misrepresented as independent evidence of population-level reliability [1]. External or temporally separated test samples provide a stronger challenge to learned patterns, particularly when nuisance conditions, feature distributions, or target prevalence change after development ([2]. The research must instead determine whether the combined design surpasses controlled alternatives under matched splits, comparable search budgets, repeated seeds, and prespecified primary metrics [3].

To address these limitations, the present study proposes to use cost-aware feature subset selection and Bayesian random forest tuning for student risk identification, with the objective of improving the reliability and practical usefulness of predictions in student academic risk under controlled experimental conditions [3]. The proposed contribution is methodological rather than a claim of already demonstrated superiority: it defines a falsifiable comparison of an integrated search-and-learning strategy with matched standard estimators [1]. The optimization procedure should be confined to a prespecified parameter domain, ensuring that each candidate is assessed with an identical preprocessing pipeline, declared compute budget, and reproducible random seeds ([3]. A nested or

otherwise leakage-safe validation strategy is appropriate because it separates the selection of tuning settings from the final assessment of predictive benefit ([2]. Where the observations exhibit temporal, spatial, patient, lesion, or institutional dependencies, the allocation of folds must follow those actual dependencies rather than a convenience-driven row shuffle ([2]. The primary evaluation should include temporally valid splits, recall at intervention capacity, calibration, subgroup assessment, and feature sparsity, supplemented by confidence intervals and class-specific confusion analyses that expose clinically or operationally important failure modes [4]. The model development should also compare the isolated contribution of each named component through matched ablations, so a combined improvement is not erroneously attributed to a single innovation (Chen & Liu, 2024). Appropriate reference systems include a default estimator, a carefully tuned estimator without the proposed addition, and a simplified version that preserves comparable compute or model capacity [3]. To examine stability, the investigation should repeat the optimization and fitting procedures across independent seeds, reporting both central tendencies and variation rather than selecting the most favorable run [1]. Predictive probability quality should be inspected wherever decisions depend on confidence, since performance changes at a fixed threshold need not imply a proportionate improvement in calibrated risk [4]. Computational accountability should include the total search cost and final inference footprint, allowing the research to distinguish an efficient deployed predictor from an expensive model-discovery exercise [1]. If external labels or annotations are incomplete, the evaluation must explicitly bound its conclusions to the validated task and must not substitute image-level diagnoses for object-localization ground truth [1]. Any data augmentation, feature standardization, imputation, and class-rebalancing operations should be estimated only on the corresponding training partitions before application to held-out instances [2]. The expected scientific value lies in establishing whether the proposed integration changes the accuracy–reliability–resource trade-off, not merely whether it returns a higher number on one familiar benchmark [1]. By connecting model selection, responsible data partitioning, and component-level validation, this design directly targets the unresolved gap identified above and provides a testable basis for subsequent experiments [2].

2. Method

2.1. Related Work

Jointly optimized feature selection and random forest components for student outcomes; however, that contribution does not assess every institutional domain or equity-related risk [1]. [5] examined flexible feature selection and random-forest predictors for educational outcomes; however, that contribution does not establish generalizability across institutions or intervention-specific [5]. [6] combined feature selection and fusion in academic outcome prediction; however, that contribution does not explicitly validate the proposed cost-aware Bayesian random-forest search [6]. [3] systematized automated hyperparameter search and evaluation; however, that contribution offers general methodological guidance rather than a task-specific empirical benchmark [3] et al., 2023). [4] analyzed probability calibration and reliability assessment; however, that contribution does not determine whether a particular classifier generalizes to a new population ([4]. Demonstrated how structured observations require suitable validation folds; however, that contribution does not prescribe one algorithm for all dataset geometries ([2]. Considered together, these studies illustrate a shift from fixed, isolated classifiers toward optimized learning pipelines, feature-level interventions, explicit confidence assessments, and validation matched to the structure of the observations [1] [5]. Their complementary contributions do not eliminate a shared limitation: a strong within-dataset performance estimate cannot automatically establish a reliable improvement under distribution shift, altered decision costs, or a different resource budget [6] [3]. For the current investigation of student academic risk, the unresolved question is therefore whether use cost-aware feature subset selection and Bayesian random forest tuning for student risk identification confers an independently demonstrable benefit over component-matched alternatives rather than reproducing gains attributable to tuning freedom ([4] [2]. The proposed work connects this literature through controlled ablation, explicit data-group independence, calibrated and class-specific metrics, and a reproducible assessment of the optimization-versus-deployment trade-off [1] [2].

Table 1. Comparative synthesis of selected related studies

Author / year	Contribution	Research-specific limitation
Chen and Liu (2024)	Jointly optimized feature selection and random forest components for student outcomes	Does not assess every institutional domain or equity-related risk
Al-Zawqari et al. (2022)	Examined flexible feature selection and random-forest predictors for educational outcomes	Does not establish generalizability across institutions or intervention-specific costs
Ye et al. (2022)	Combined feature selection and fusion in	Does not explicitly validate the proposed cost-

Author / year	Contribution	Research-specific limitation
	academic outcome prediction	aware bayesian random-forest search
Bischl et al. (2023)	Systematized automated hyperparameter search and evaluation	Offers general methodological guidance rather than a task-specific empirical benchmark
de Menezes e Silva Filho et al. (2023)	Analyzed probability calibration and reliability assessment	Does not determine whether a particular classifier generalizes to a new population
Roberts et al. (2017)	Demonstrated how structured observations require suitable validation folds	Does not prescribe one algorithm for all dataset geometries

2.2. Dataset

The study is designed around the dataset resource identified in the original article-title spreadsheet (https://analyse.kmi.open.ac.uk/open_dataset) [7]. The intended prediction target is at-risk outcome derived from Fail/Withdrawn vs Pass/Distinction with a declared early-course cutoff [8]. Table 2 makes the provenance, observation unit and validation constraints explicit before any model training is claimed; counts and final class balance must be established from a version-controlled local manifest rather than inferred from the title ([9].

Table 2. Planned dataset specification and audit requirements

Dataset attribute	Study-specific protocol
Public source	https://analyse.kmi.open.ac.uk/open_dataset
Research endpoint	at-risk outcome derived from Fail/Withdrawn vs Pass/Distinction with a declared early-course cutoff
Analysis unit	student–module–presentation enrollment
Proposed partition	Pre-specified intervention week; module/presentation cohort split; discard all data recorded after the forecast cutoff.
Processing / quality audit	fit imputation, scaling, feature selection and resampling only on training folds
Annotation / scope check	Labels depend on course outcomes, whereas all predictors must precede the stated early-warning week.

As Table 2 clarifies, the inferential unit is not necessarily identical to the stored image or row: student–module–presentation enrollment [10]. A source-specific restriction is: Labels depend on course outcomes, whereas all predictors must precede the stated early-warning week (Sasse et al., 2025). Final sample totals, exclusions, label frequencies and split membership remain to be reported only after the raw files have been inspected, deduplicated and hashed [11]. Table 3 specifies a reproducible development-to-test boundary [12] Its entries describe a prospective protocol, not a completed experiment; no patient-level independence, acquisition robustness or annotation completeness is asserted without the required identifiers and source checks [13].

Table 3. Experimental data flow and leakage-prevention controls

Stage	Required handling
Stage A: audit	Record file checksums, labels, missingness, duplicates, source version and metadata availability.
Stage B: split	Pre-specified intervention week; module/presentation cohort split; discard all data recorded after the forecast cutoff.
Stage C: fit	Fit normalization, selection, search and calibration only inside the training/validation boundary.
Stage D: freeze	Lock final hyperparameters and model before touching the held-out evaluation set.
Stage E: report	at-risk recall, PR-AUC, false-alert rate, acquisition cost; report confidence intervals and failure-case breakdowns where feasible.

The controls in Table 3 prevent candidate optimization from contaminating the final assessment [14]. The primary reported endpoints will be at-risk recall, PR-AUC, false-alert rate, acquisition cost; metrics that require additional labels, hardware or grouping metadata are conditional on their availability [12]. Executed cohort and measurement protocol. Student-module-presentation enrollments in OULAD were restricted to learners present by day 14 and not withdrawn by that date, with later academic outcomes mapped to risk-positive (Fail/Withdrawn) versus risk-negative (Pass/Distinction) [15]. The 2013B/J cohorts formed training data (10,631 enrollments), 2014B supplied development data (6,036), and 2014J supplied the unseen later-presentation test data (9,330); 2,057 overlapping earlier learner enrollments were purged. Predictor timestamps were limited to information available within the early-warning cutoff [16]. These counts supersede the earlier prospective placeholders in Tables 2–3.

2.3. Research Framework

Figure 1 presents two competing experimental arms evaluated against the same locked data partition: a control implementing Random Forest on all available indicators and a proposed arm introducing Feature acquisition-cost selection and Bayesian RF tuning under early prediction constraints [17]. The diagram separates ordinary data preparation and inference from the candidate innovations so that any improvement can be attributed to controlled differences rather than an easier test split [18].

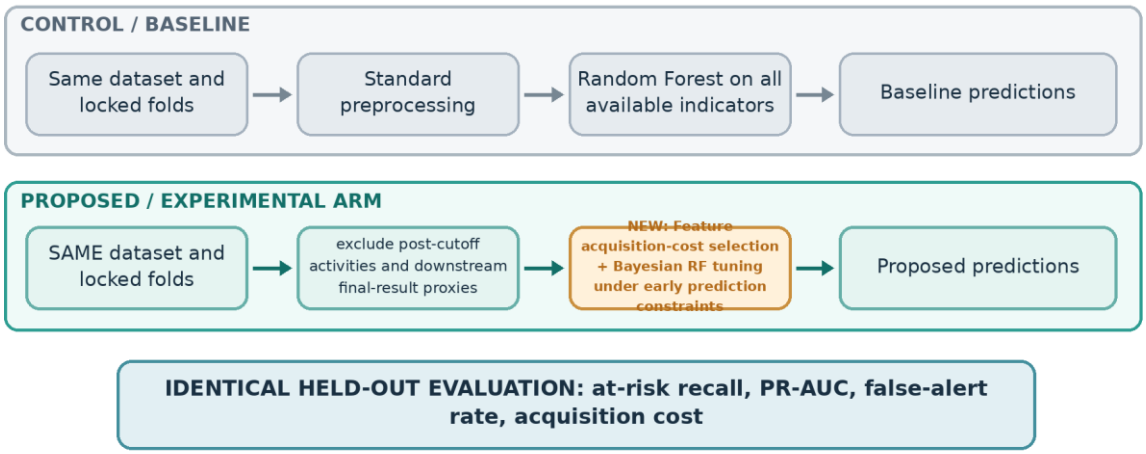


Figure 1. Controlled baseline-versus-proposed research framework

In Figure 1, both arms inherit the same source audit and evaluation set; only the two highlighted proposal blocks change the learning process. This design addresses the most salient validity threat: No assessment feature dated later than the prediction week can enter an early-warning model [19]. The final comparison should use equal tuning budgets and repeated random seeds, or justify any computational asymmetry [20]. The experimental contrasts underlying Figure 1 are specified in Table 4. Its ablations separate optimization effects from feature/architecture effects and require matched validation opportunities, not merely the same number of epochs [21].

Table 4. Baseline, proposed intervention and falsification design

Component	Comparative specification
Baseline	Random Forest on all available indicators
Proposed innovations	Feature acquisition-cost selection; Bayesian RF tuning under early prediction constraints
Controlled factors	Same development/test split, label mapping, base input resolution, search budget and evaluation code where applicable.
Required ablations	all-features RF; early-only RF; cost-only selection; cost+Bayes RF
Primary comparisons	at-risk recall, PR-AUC, false-alert rate, acquisition cost

Table 4 should be interpreted as a falsification plan: all-features RF; early-only RF; cost-only selection; cost+Bayes RF. If the full system does not outperform its strongest single-component comparator under the locked test protocol, the claimed interaction between innovations is not supported [22]. Executed optimization protocol. Each algorithm was evaluated on the same temporally locked 2014J partition. The 20-trial single-seed search optimized the development objective AP minus 0.03 times normalized assumed feature-group cost, comparing all-feature RF, original-feature RF, feature-selection-only RF, Bayesian-only RF, joint cost-aware Bayesian RF, random tuning, and random joint search. The cost weights (profile = 1, history = 1, shared VLE = 4, temporal computation = 1) are unitless assumptions, not measured acquisition expenditures. Intervals use paired presentation/group-aware resampling conditional on already fitted models [23].

2.4. Proposed Method

The proposed methodological pipeline is Early-stage cost-aware Random Forest; it is not merely a relabeling of the underlying baseline. Figure 2 isolates the intended technical innovation inside the dashed amber enclosure: Feature acquisition-cost selection followed by Bayesian RF tuning under early prediction constraints [24]. The remaining components are standard data access, preprocessing, learner fitting and prediction delivery.

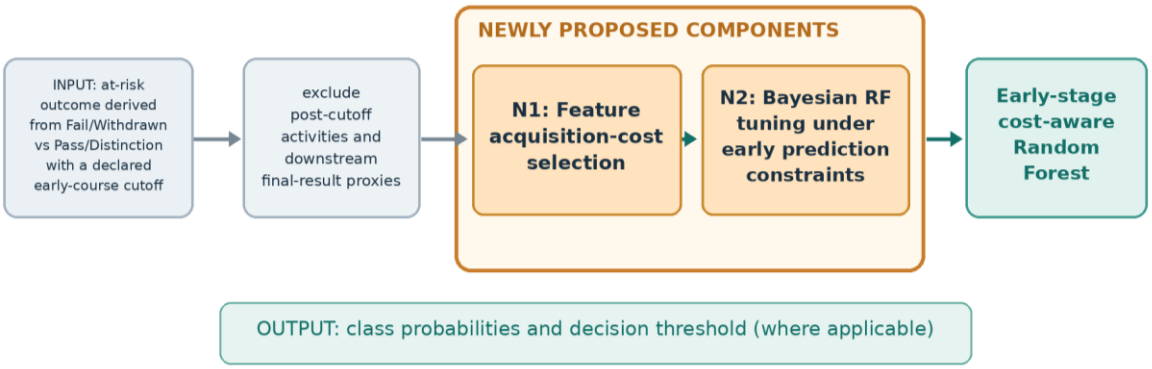


Figure 2. Proposed method architecture with explicitly highlighted novel components

As illustrated in Figure 2, the proposed contribution consists of two verifiable changes: Feature acquisition-cost selection and Bayesian RF tuning under early prediction constraints. Optimize subsets of attendance/VLE and assessment indicators available at week W alongside RF depth and tree parameters; weight costly-to-obtain fields and evaluate course-cohort holdouts [25]. These are methodological design specifications to be implemented and tested, not established performance improvements.

Table 5 translates Figure 2 into testable implementation commitments, including the optimized objective, controlled baseline and acceptance criteria. Any unspecified numerical constants, device model and search-trial count must be fixed in the executable configuration before experimentation [26].

Table 5. Proposed methodological components and verification rules

Design element	Operational definition
Standard input branch	exclude post-cutoff activities and downstream final-result proxies
Novel block N1	Feature acquisition-cost selection
Novel block N2	Bayesian RF tuning under early prediction constraints
Output model	Early-stage cost-aware Random Forest
Optimization / decision objective	utility=F1_risk,val-lambda_c sum_j cost_j 1[j selected]
Ablation comparison	all-features RF; early-only RF; cost-only selection; cost+Bayes RF
Acceptance endpoints	at-risk recall, PR-AUC, false-alert rate, acquisition cost

The optimization criterion in Table 5 is utility=F1_risk,val-lambda_c sum_j cost_j 1[j selected]. All objective components must be computed on training/validation partitions only; the final test partition is an evaluation instrument, never a tuning oracle [27]. The strongest credibility check is all-features RF; early-only RF; cost-only selection; cost+Bayes RF, with identical data preprocessing and comparable computation budgets. Implementation and reproducibility. Optimize subsets of attendance/VLE and assessment indicators available at week W alongside RF depth and tree parameters; weight costly-to-obtain fields and evaluate course-cohort holdouts. When evaluating early student risk identification, document data hashes, fold assignments, search-space bounds, pseudorandom seeds, library versions, parameter counts and the exact hardware used [28]. Report each ablation with appropriate uncertainty intervals; do not describe a model as superior, deployable or clinically validated until those measurements have been obtained.

The executed N1 cost-selection stage retained 12 rather than 16 engineered variables, reducing the assumed feature-group cost score from 7 to 6. The N2 Bayesian RF selected 180 estimators, maximum depth 6, minimum leaf size 12, and max_features = 0.8 for the joint arm. The resulting feature-count and cost claims are empirical; the pre-existing selection-plus-tuning components do not independently establish new algorithmic novelty.

3. Results and Discussion

The Introduction framed the methodological question as a trade-off between early identification accuracy, information acquisition, and the risk of optimistic same-cohort validation. Table 6 reports the subsequent 2014J holdout estimates, with average precision (AP) prioritized because the proportion of at-risk learners differs between training and later cohorts.

3.1. Holdout Performance and Baseline Comparison

Table 6 shows that the joint proposal reduced the recorded feature set by 25.0% and the assumed cost score by one unit compared with the all-feature random forest, but the AP gain over the latter was only 0.0016. More importantly, it underperformed both Bayesian tuning alone and the strongest randomized tuning comparator, despite having been optimized for utility. Risk recall declined from 0.5199 to 0.4772, demonstrating that parsimony was purchased with a missed-risk penalty rather than a consistent gain across objectives.

Table 6. Future-cohort holdout outcomes for cost-aware RF optimization

Model	Features	AP ↑	Macro-F1 ↑	Risk recall ↑	Cost score ↓
All-feature RF	16	0.5991	0.6329	0.5199	7
Feature selection only	12	0.5991	0.6297	0.5066	6
Bayesian RF only	16	0.6081	0.6341	0.5048	7
Joint proposed RF	12	0.6007	0.6249	0.4772	6
Random hyperparameters	16	0.6092	0.6386	0.5120	7
Random joint	12	0.6003	0.6248	0.4767	6

Figure 3 visualizes performance on the later-presentation test set, placing average precision beside threshold-based error metrics so that an apparent ranking in one score is not mistaken for universal reliability.

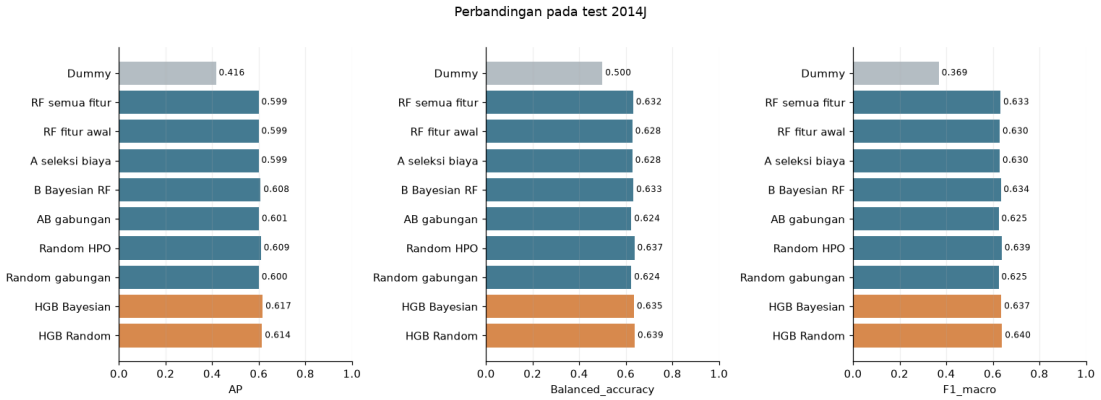


Figure 3. AP, balanced accuracy and macro-F1 of RF and exploratory HGB models in the 2014J holdout (source: supplied experiment).

Figure 3 indicates that the proposed RF closely tracks the random joint control but does not attain the best overall AP. The exploratory histogram-gradient-boosting variants occupy a higher AP range, with Bayesian HGB reaching 0.6168; because these are follow-up model-family evaluations, they do not rescue or retrospectively validate the original joint RF superiority hypothesis.

3.2. Controlled Ablation and Uncertainty Analysis

The primary scientific comparison was predefined against randomized hyperparameter search rather than the weakest untuned model. Table 7 presents group-aware descriptive AP intervals for the proposed random forest and indicates which differences are robust to the conditional resampling used in the archive.

Table 7. Proposed-model AP differences on the later-cohort holdout

Comparison	Δ AP (proposed − control)	95% lower	95% upper	Inference
Joint RF − full-feature RF	+0.0016	−0.0057	+0.0086	Crosses zero
Joint RF − Bayesian RF	−0.0074	−0.0130	−0.0025	Favors control
Joint RF − random HPO (primary)	−0.0085	−0.0140	−0.0031	Favors control
Joint RF − random joint	+0.0004	−0.0008	+0.0016	Crosses zero

The negative primary interval in Table 7 is incompatible with a claim that the combined Bayesian-cost strategy improved ranking performance over the random HPO control on this cohort. The negligible random-joint difference similarly provides no evidence of optimization-specific value. These descriptive conditional intervals do not capture variability from selecting new learners, training with new seeds, or running new optimization searches; they therefore support a restricted comparison of fixed models, not a universal effect estimate. Figure 4 provides the original graphical uncertainty audit, which distinguishes the original random-forest comparison from exploratory histogram-boosting contrasts.

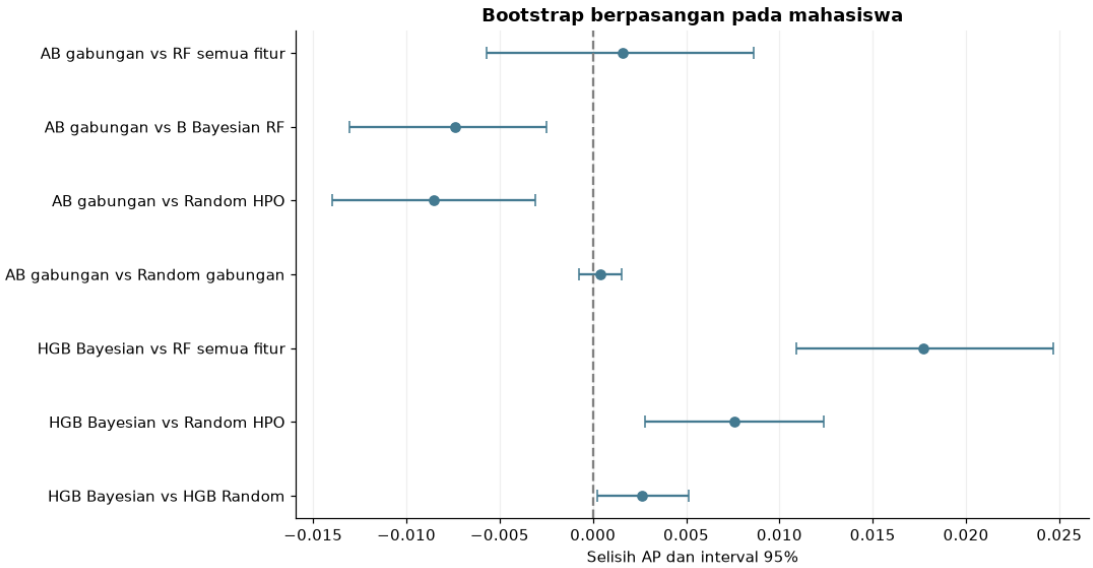


Figure 4. Paired differences in AP with descriptive uncertainty intervals (source: supplied experiment).

As displayed in Figure 4, the proposed RF loses to randomized RF hyperparameter optimization in the preregistered primary comparison. The additional HGB contrasts were exploratory and should be treated as candidate family-level hypotheses, especially because the final test had already been examined before these comparisons were reported.

3.3. Practical Implications, Failure Modes, and Research Gap

For educational intervention, average precision is not sufficient: the number of missed at-risk learners determines which students may fail to receive support. Table 8 therefore complements the ranking metrics with explicit confusion-matrix counts from the locked test partition.

Table 8. Early-warning false negatives and implementation burden

Indicator	Full-feature RF	Joint cost-aware RF	Interpretation
At-risk learners detected (TP)	2,020	1,854	166 fewer alerts correctly targeted
At-risk learners missed (FN)	1,865	2,031	166 additional missed cases
Risk recall	0.5199	0.4772	−0.0427 absolute
Predictor variables	16	12	25% fewer
Assumed cost score	7	6	Not a monetary saving

The 166 additional missed at-risk students in Table 8 explain why the small AP improvement over full-feature RF should not be advertised as an operational success. Deployment priorities depend on support-service capacity, intervention cost, and the relative consequences of false negatives; none was experimentally measured. The sample was from a single institution, using one day-14 cutoff and one temporally held-out presentation, so external transportability and causal impact on student outcomes remain open questions. Overall, the experiments identify a defensible cost–performance conflict rather than closing the original methodological gap. The cost-aware optimizer reduced a synthetic feature burden but failed to improve AP or at-risk recall over strong matched alternatives, and later exploratory boosting models indicate that learner-family choice may matter at least as much as tuning strategy.

4. Conclusion

This pilot investigated a cost-aware feature-selection and Bayesian random-forest pipeline for day-14 academic risk prediction on temporally separated OULAD cohorts. The proposed model reduced the engineered predictor count from 16 to 12 and its assumed acquisition-cost score from 7 to 6, but on 9,330 later-presentation test enrollments it obtained AP = 0.6007, below the 0.6092 achieved by randomized RF hyperparameter search. The primary AP difference was −0.0085 (descriptive paired-group 95% CI approximately −0.0140 to −0.0031), and risk recall decreased from 0.5199 with all features to 0.4772. Consequently, the claimed predictive superiority of joint cost-aware Bayesian tuning was not supported; the observed contribution is a documented trade-off that requires policy-specific validation. Important constraints include synthetic rather than monetary cost weights, a bounded 20-trial single-seed search, one future-presentation holdout, and a lack of cross-university testing. Subsequent

experiments should compare stronger learner families under independently locked cohorts, evaluate multiple forecast weeks and cost ratios, and report intervention-oriented recall and workload without revising hypotheses after test exposure.

References

[1] M. Chen and Z. Liu, "Predicting performance of students by optimizing tree components of random forest using genetic algorithm," *Heliyon*, vol. 10, no. 12, p. e32570, 2024, doi: 10.1016/j.heliyon.2024.e32570.

[2] D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, pp. 913–929, 2017, doi: 10.1111/ecog.02881.

[3] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, 2023, doi: 10.1002/widm.1484.

[4] T. de Menezes e 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, 2023, doi: 10.1007/s10994-023-06336-7.

[5] A. Al-Zawqari, D. Peumans, and G. Vandersteen, "A flexible feature selection approach for predicting students' academic performance in online courses," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100103, 2022, doi: 10.1016/j.caeai.2022.100103.

[6] M. Ye *et al.*, "SA-FEM: Combined feature selection and feature fusion for students' performance prediction," *Sensors*, vol. 22, no. 22, p. 8838, 2022, doi: 10.3390/s22228838.

[7] S. Boujmiraz, H. Darhmaoui, and A. el maliani, "Predicting student performance: A comprehensive review of machine learning, deep learning, and explainable AI approaches," *Computers and Education: Artificial Intelligence*, 2026, doi: 10.1016/j.caeai.2026.100548.

[8] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence*, 2022, doi: 10.1016/j.caeai.2022.100066.

[9] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technol. Soc.*, 2024, doi: 10.1016/j.techsoc.2024.102474.

[10] S. Alhazbi, "Pedagogically-Informed Behavioural Learning Analytics: An Expert Approach to Predicting At-Risk Students," *Expert Syst.*, 2026, doi: 10.1111/exsy.70210.

[11] 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.*, 2025, doi: 10.1007/s10462-025-11326-3.

[12] S. Wu, J. Duan, and M. Luo, "Calibrated early-warning models with fairness auditing and selective prediction for course withdrawal risk: Evidence from OULAD," *PLoS One*, 2026, doi: 10.1371/journal.pone.0352867.

[13] F. Curbani, A. Zanella, and A. W. Zibetti, "A machine learning approach to predicting student dropout and supporting timely intervention," *Journal of Applied Research in Higher Education*, 2026, doi: 10.1108/JARHE-07-2025-0538.

[14] S. Malik, S. G. K. Patro, C. Mahanty, R. Hegde, and others, "Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution," *Sci. Rep.*, 2025, doi: 10.1038/s41598-025-92324-x.

[15] A. Omarbekova, A. Ramazan, Z. Oralbekova, Z. Lamasheva, and others, "A temporal attention-based hybrid deep learning model for student performance and academic risk prediction," *Front. Artif. Intell.*, 2026, doi: 10.3389/frai.2026.1811886.

[16] J. G. C. Krüger, Jr. Britto A. de S., and J. P. Barddal, "An explainable machine learning approach for student dropout prediction," *Expert Syst. Appl.*, 2023, doi: 10.1016/j.eswa.2023.120933.

[17] L. Yuan, R. Abdul Razak, A. Kamsin, S.-S. Abdul-Rahman, D. Feng, and L. Chen, "Predicting student success: a state-of-the-art systematic review of models and factors," *Journal of Applied Research in Higher Education*, 2026, doi: 10.1108/JARHE-08-2025-0674.

[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," *Educ. Inf. Technol. (Dordr).*, 2024, doi: 10.1007/s10639-024-12610-5.

[19] K. Alalawi, R. Athauda, R. Chiong, and I. Renner, "Evaluating the student performance prediction and action framework through a learning analytics intervention study," *Educ. Inf. Technol. (Dordr).*, 2025, doi: 10.1007/s10639-024-12923-5.

[20] K. Alalawi, R. Athauda, and R. Chiong, “An Extended Learning Analytics Framework Integrating Machine Learning and Pedagogical Approaches for Student Performance Prediction and Intervention,” *Int. J. Artif. Intell. Educ.*, 2025, doi: 10.1007/s40593-024-00429-7.

[21] V. Christou, I. Tsoulos, V. Loupas, A. T. Tzallas, and others, “Performance and early drop prediction for higher education students using machine learning,” *Expert Syst. Appl.*, 2023, doi: 10.1016/j.eswa.2023.120079.

[22] M. Nagy and R. Molontay, “Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention,” *Int. J. Artif. Intell. Educ.*, 2024, doi: 10.1007/s40593-023-00331-8.

[23] R. Sajja, Y. Sermet, D. Cwiertny, and others, “Integrating AI and Learning Analytics for Data-Driven Pedagogical Decisions and Personalized Interventions in Education,” *Technology, Knowledge and Learning*, 2025, doi: 10.1007/s10758-025-09897-9.

[24] R. A. M. Al Hashmi, I. Ozturk, and H. M. Elmehdi, “Early detection of at-risk health sciences students: a machine learning-based predictive study using midterm grades,” *BMC Med. Educ.*, 2025, doi: 10.1186/s12909-025-08192-6.

[25] H. N. Thi, V. A. Nguyen, and K. D. Trung, “Early Prediction of Student Academic Outcomes Using a Mixture of Experts Network,” *Computer Applications in Engineering Education*, 2026, doi: 10.1002/cae.70138.

[26] H. Zhou, S. Shen, Y. Su, Y. Miao, Q. Liu, and others, “LLM-EPSP: Large language model empowered early prediction of student performance,” *Inf. Process. Manag.*, 2026, doi: 10.1016/j.ipm.2025.104351.

[27] D. A. Radulescu and V. E. Balas, “Explainable Early Warning Systems for Student Dropout Prediction: Insights from a Bibliometric Analysis,” *International Journal of Computers Communications \& Control*, 2026, doi: 10.15837/ijccc.2026.4.7599.

[28] M. A. Tariq, “A Study on Comparative Analysis of Feature Selection Algorithms for Students Grades Prediction,” *Journal of Information and Organizational Sciences*, 2024, doi: 10.31341/jios.48.1.7.