Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Explainable Machine Learning for Predicting Indonesian Vocational School Accreditation: Geographic Validation, Probability Calibration, and Subgroup Auditing

Muhamad Riyan Maulana [1,*], Putu Sudira [1], Priyanto [1], Yanuar Agung Fadlullah [1], Dian Nurdiana [2], and Amir Hamzah bin Sofhi @ Subhi [3]

[1] Department of Technology and Vocational Education, Graduate School, Universitas Negeri Yogyakarta, 55281, Yogyakarta, Indonesia; e-mail : muhamadriyan.2025@student.uny.ac.id; putupanji@uny.ac.id; priyanto@uny.ac.id; yanuaragung.2025@student.uny.ac.id.
[2] Data Science Study Program, Faculty of Science and Technology, Universitas Terbuka, 15437, Tangerang Selatan, Indonesia; e-mail : dian.nurdiana@ecampus.ut.ac.id.
[3] Practicum Unit, Institut Pendidikan Guru Kampus Pendidikan Teknik, 71760, Bandar Baru Enstek, Malaysia; e-mail : jurnaledd@gmail.com
* Corresponding Author : Muhamad Riyan Maulana

Abstract: Vocational school accreditation is a high-stakes institutional classification, yet the geographic transferability and probability reliability of models based on routinely collected administrative data remain underexplored. This study developed a reproducible, leakage-controlled machine-learning framework for classifying the recorded accreditation classes C, B, and A among Indonesian vocational schools. Its main contribution is an integrated evaluation design that combines province-disjoint validation, probability calibration, cluster-aware uncertainty, and subgroup auditing to test reliability beyond conventional random-validation performance. A cross-sectional dataset containing 149,376 program-level records was deterministically aggregated into 14,391 schools, including 14,134 with valid labels. Eight provinces comprising 1,536 schools were locked before feature selection, model comparison, tuning, and calibration; 12,598 schools from 31 provinces supported development. The selected CatBoost model used 35 structural and vocational features with grouped out-of-fold temperature scaling. On the locked holdout, the calibrated model achieved balanced accuracy of 0.586, Macro F1 of 0.533, quadratic weighted kappa of 0.499, Macro ROC-AUC of 0.757, and expected calibration error of 0.049. School size, teacher resources, vocational staffing, program diversity, and status were most influential. Class B had the lowest recall, and performance varied across school groups and provinces. The framework supports calibrated, auditable preliminary screening, but not automated accreditation or replacement of professional assessors.

Keywords: CatBoost; Educational Data Mining; Explainable Artificial Intelligence; Geographic Validation; Machine Learning; Probability Calibration; Subgroup Auditing; Vocational School Accreditation.

Received: August, 2nd 2026
Revised: August, 28th 2026
Accepted: September, 1st 2026
Published: September, 12th 2026

1. Introduction

Vocational school accreditation is a high-stakes institutional classification because it condenses evidence on governance, teacher resources, learning processes, and educational outcomes into judgments that shape accountability and school-improvement priorities [1]. In Indonesia, these dimensions are interdependent [2] and operate across a geographically diverse vocational system in which labour-market demand, institutional capacity, leadership, and teacher preparation vary substantially between regions [3]. Routinely collected administrative data and artificial intelligence may support earlier institutional screening [4], but predictive systems can reproduce existing disparities [5] and remain constrained by generalizability, transparency, and ethical integration into educational decisions [6]. The relevant question is therefore not simply whether accreditation labels can be predicted, but whether

predictions remain transferable, probabilistically reliable, interpretable, and appropriately bounded.

Existing evidence indicates that accreditation-related outcomes contain a learnable predictive signal. Lee and Cho showed that disclosure data could predict higher-education accreditation outcomes [7]. More broadly, reviews of predictive learning analytics identify persistent limitations in external validity, transparent interpretation, and translation into defensible educational action [8]. These limitations are especially consequential for institutional accreditation, where prediction errors may influence which schools receive additional scrutiny. Accreditation labels also reflect assessment processes and institutional context that are only partially captured by administrative counts; consequently, an apparently accurate model can still fail when transferred to provinces with different school compositions.

Geographic transfer is a central methodological challenge because a model developed using schools from some provinces may be applied to provinces absent from model development. Distribution-shift benchmarks show that conventional development performance can deteriorate in genuinely different environments [9], while internal-external validation can expose between-cluster heterogeneity hidden by ordinary resampling [10]. Predictive confidence may also degrade under dataset shift [11]; calibration must therefore be assessed separately from discrimination [12], particularly in multiclass settings where aggregate accuracy cannot represent class-specific probability behaviour [13].

Trustworthy screening requires evidence beyond predictive performance. Human-centred educational-AI research reports uneven treatment of human control, safety, reliability, and trustworthiness [14]. Leakage can produce severely optimistic claims and undermine reproducibility [15], while algorithmic auditing should cover problem formulation, data boundaries, model development, evaluation, and downstream risks [16]. Lifecycle assurance likewise treats validation, documentation, explanations, and monitoring as interdependent safeguards [17]. Accordingly, complete and transparent reporting of model development and evaluation remains essential [18].

Against this background, the principal research gap is the absence of a unified, leakage-controlled evaluation design for multiclass vocational-school accreditation that jointly addresses geographic transfer, probability calibration, clustered uncertainty, explainability, subgroup heterogeneity, robustness, and reproducibility. This study addresses that gap by deterministically aggregating program-level records to schools; comparing random and province-grouped validation; locking eight provinces before feature selection, model comparison, tuning, and calibration; evaluating the selected model once on the untouched holdout; and auditing explanations, subgroups, school-status dependence, sensitivity, and reproducibility. Unlike performance-only comparisons, this design treats geographic discrimination, probability reliability, clustered uncertainty, and subgroup behaviour as distinct but connected evidence for responsible use.

Four research questions guide the analysis: RQ1 examines how accurately structural and vocational administrative features distinguish accreditation classes C, B, and A under random, province-grouped, and locked geographic evaluation; RQ2 assesses probability calibration, statistical stability, and influential school characteristics; RQ3 evaluates performance across school status, school size, and held-out provinces, including the effect of removing school status; and RQ4 examines robustness under alternative samples and outcome definitions, together with model responses to consistently varied vocational-teacher inputs. The study aims to develop a reproducible, calibrated, explainable, and subgroup-audited framework for preliminary review prioritization, not an automated accreditation mechanism or a substitute for professional assessors.

2. Related Work

2.1. Artificial Intelligence and Predictive Analytics in Education

AI in education increasingly supports assessment, prediction, recommendation, and institutional decisions [19]. Yet educator roles, ethical safeguards, transparency, and implementation remain uneven, and predictive dashboards often omit guidance for interpreting probabilities or acting on them [20]. These gaps justify evaluations that link performance with uncertainty, explanation, and bounded use.

2.2. Accreditation and School-Quality Prediction

Administrative data have supported accreditation and school-quality modelling across schools and universities using linear, tree, ensemble, and neural methods [21], [22]. Yet accreditation also depends on teacher development, assessment management, assessor roles, testing, and institutional conditions not fully represented by routine counts [23]. Indonesian evidence confirms that ensemble models can distinguish accreditation categories [24], but most studies rely on pooled or conventional splits and rarely combine unseen-region transfer, calibration, clustered uncertainty, subgroup analysis, and reproducibility.

2.3. Transferability, Calibration, and Responsible Evaluation

Context-aware feature analysis is essential because explanation and fairness claims cannot rest on black-box outputs [25]; representative multi-metric comparisons within one dataset are also preferable to unsupported state-of-the-art claims [26]. Responsible evaluation should trace harms across the machine-learning lifecycle [27] and treat fairness as multidimensional [28]. Educational early-warning evidence further shows that satisfactory internal performance may fail at unseen institutions despite explainability and fairness calibration [29]. Accreditation screening should therefore combine class-specific calibration [12], leakage-controlled reporting [15], [18], evaluation across separated contexts [9], [10], and human oversight with subgroup monitoring [14], [29], [30].

2.4. Research Gap and Study Positioning

Prior studies do not jointly test geographic transfer, class-probability reliability, province-cluster uncertainty, subgroup failures, and robustness to analytical definitions in multiclass vocational accreditation. This study addresses that gap through a sequenced framework: deterministic school-level aggregation; province-disjoint development and evaluation; grouped out-of-fold calibration; cluster-aware uncertainty; interpretation; subgroup and status audits; sensitivity analysis; and reproducibility checks. The contribution is their integration, not any single classifier or explanation method.

3. Proposed Method

3.1. Study design and analytical workflow

This retrospective cross-sectional study followed a fixed leakage-control sequence: rule-based processing and school-level aggregation; eligibility assessment; locking eight holdout provinces; development-only feature and model comparison, grouped tuning, and out-of-fold calibration; specification freezing; and one-time holdout evaluation with uncertainty, explanation, subgroup, sensitivity, and reproducibility audits. Figure 1 presents this workflow. Predictors and labels came from one administrative snapshot without dated decisions, so the model classifies contemporaneous status for retrospective review prioritization rather than forecasting future accreditation; prospective use requires temporal alignment and external revalidation.

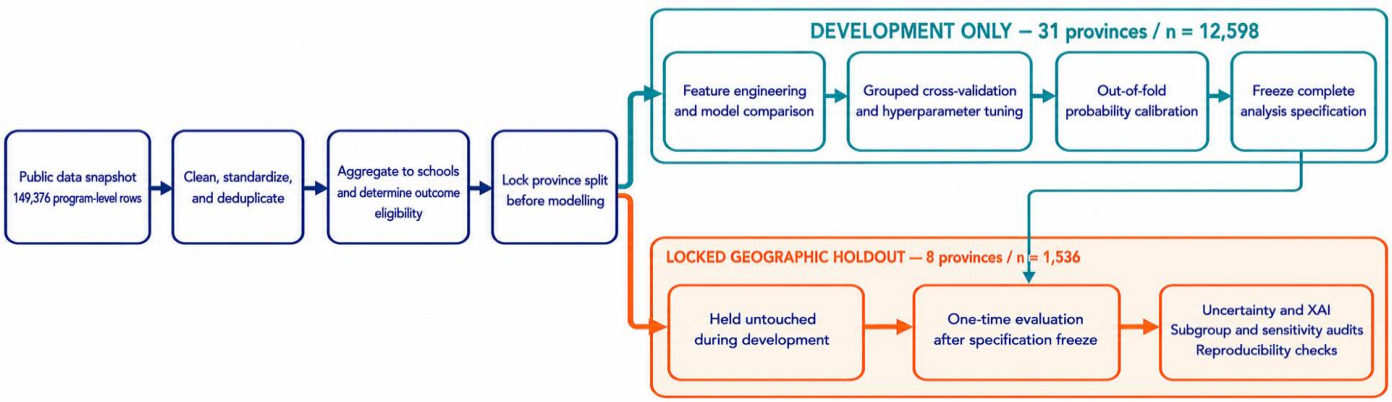

Figure 1. Research Workflow

3.2. Data source, eligibility, and school-level aggregation

The study used the official Kemendikdasmen vocational-school snapshot retrieved on 10 July 2026, covering identifiers, ownership, location, vocational fields, programs, competencies, accreditation, enrolment, and teachers. After NPSN-header detection, semicolon parsing, standardization, and exact-duplicate removal, repeated NPSN rows were retained only when they represented distinct programs or competencies. NPSN defined schools: distinct-program student counts were summed; recurring school-wide teacher totals used the maximum; and identity, location, ownership, and accreditation used the mode. Positive enrolment defined active programs and competencies, and vocational fields were harmonized before breadth, share, and concentration calculations. Counts, ratios, and shares remained observed; missing values were not imputed and observed zeros were retained. Primary modelling included C, B, and A; other or missing labels remained only for descriptive and sensitivity analyses. The 149,376 program rows became 141,429 after deduplication, representing 14,391 schools, 14,134 with valid labels. The final 35 predictors covered 12,598 development schools in 31 provinces and 1,536 locked-holdout schools in eight provinces.

3.3. Outcome, predictors, and feature engineering

The ordered outcome remained C = 0, B = 1, A = 2. Three prespecified candidates were compared: 16 core structural variables, 35 structural-plus-vocational variables, and the 35 variables plus province. The augmented set covered size, teacher resources, ratios, vocational staffing, grade shares, structural-zero indicators, program and competency breadth, activity, diversity and concentration, field enrolment shares, and status. Province was optional because the final specification had to transfer beyond observed regions. Non-negative counts were log-transformed using Eq. (1).

Feature blocks were conceptually prespecified. Size captured operating scale; teacher counts and ratios represented staffing capacity and pressure; program and competency breadth represented complexity and curricular reach; vocational shares and HHI represented specialization; grade shares and activity represented enrolment structure; zero indicators distinguished no reported activity from small values; and status represented governance. These are institutional-capacity proxies, not measures of instructional quality or causal determinants.

Feature-set selection used development data only. On identical province-grouped folds, Macro F1 was primary, with quadratic weighted kappa and Macro ROC-AUC supporting. The 35-feature augmented set without province was frozen because it had the highest grouped Macro F1 and remained applicable to unseen provinces. Holdout features and labels never entered construction, selection, or tuning.

$$x_{\log} = \ln(1 + x) \tag{1}$$

Here, x is the original count, $x_{\log}$ the transformed value, and ln the natural logarithm. Vocational-field concentration used the Herfindahl-Hirschman Index in Eq. (2).

$$\mathrm{HHI} = \sum_{j=1}^{J} s_j^2 \tag{2}$$

In Eq. (2), s_j is the enrolment share in field j, and J is the number of represented fields; higher HHI denotes greater concentration and lower HHI denotes greater diversity. Ratios and shares used consistent numerator-denominator pairs, and structural zeros remained distinct from missing observations.

3.4. Descriptive analysis and validation architecture

Continuous features were compared by accreditation class using Kruskal-Wallis tests, Benjamini-Hochberg correction, and ε^2 effect sizes [30], [31]; status used chi-square and Cramer's V, and selected continuous associations used Spearman correlations. Before any feature or model decision, eight provinces (1,536 schools) were locked; 12,598 schools in 31 provinces formed development data. Random stratified five-fold validation provided a benchmark, while five-fold StratifiedGroupKFold by province estimated geographic transfer [32]. Three repeated grouped evaluations assessed stability, and the frozen model was evaluated once on the holdout.

Holdout provinces were chosen purposively before modelling to span western, central, and eastern Indonesia while retaining all three classes collectively; no model metric influenced selection. The set was DKI Jakarta, Sumatera Barat, Riau, Bengkulu, Kalimantan Selatan, Sulawesi Tengah, Papua Barat Daya, and Papua Pegunungan, with all other eligible provinces used for development. This holdout design tests named unseen regions but is not statistically representative of every province.

3.5. Model development and performance evaluation

The comparison covered complementary assumptions: the majority dummy set a prevalence floor; multinomial logistic regression supplied a transparent unordered linear baseline; cumulative ordinal logistic regression tested the C-B-A order under proportional odds; random forest represented bagged nonlinear trees; and CatBoost represented regularized boosting for heterogeneous tabular features. CatBoost was well suited to this mix of log-counts, continuous ratios and shares, binary structural-zero indicators, and categorical status, where accreditation may reflect nonlinear interactions among scale, staffing, and program breadth. Because the curated dataset lacks a standard benchmark or fixed feature space, no state-of-the-art claim is made; identical folds, features, and metrics provide an internally valid comparison.

All five estimators were compared, and CatBoost's ordered boosting was used to reduce target leakage and prediction shift while retaining nonlinear interactions [33]. Twelve configurations were tested only by province-grouped cross-validation. The selected model used 800 iterations, depth 7, learning rate 0.03, L_2 regularization 8, balanced class weights, random strength 2, and bagging temperature 0.5.

F_1 was primary because it weights classes equally under imbalance. Accuracy, balanced accuracy, class-wise precision and recall, quadratic weighted kappa, Matthews correlation, one-versus-rest Macro ROC-AUC, log loss, Brier score, and ECE were also reported [34]. Kappa preserves ordering; ROC-AUC evaluates probability ranking; lower log loss and Brier values indicate better probabilities. Eq. (3) defines multiclass log loss.

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{3} y_{ic}\ln(p_{ic}) \tag{3}$$

In Eq. (3), n is the number of schools, i indexes schools, c indexes classes, y_{ic} is the one-hot outcome, and p_{ic} the predicted probability.

3.6. Probability calibration, uncertainty, and explainability

Calibration used training-only province-grouped out-of-fold probabilities. A temperature T minimizing multiclass log loss over $0.25 \leq \mathrm{T} \leq 5.00$ adjusted confidence without changing class decisions [35]; Eq. (4) gives the calibrated probability for class c.

$$p_{\text{cal}}(c \mid x) = \frac{\exp(z_c(x)/T)}{\sum_{j=1}^{3}\exp\big(z_j(x)/T\big)} \tag{4}$$

In Eq. (4), $p_c^{\text{cal}}(x)$ is the calibrated probability for class c and feature vector x, $z_c(x)$ the logit, T the fitted temperature, and j the class index. Calibration used log loss, Brier score, class-specific reliability curves, and ECE with 10 bins, as defined in Eq. (5).

$$\text{ECE} = \sum_{m=1}^{M}\frac{\mathrm{N}(B_m)}{n}\,\text{abs}\big(\text{accuracy}(B_m) - \text{confidence}(B_m)\big) \tag{5}$$

In Eq. (5), ECE is expected calibration error; B_m is bin m; $|B_m|$ its size; M the bin count; and $\text{accuracy}(B_m)$ and $\text{confidence}(B_m)$ its observed accuracy and mean confidence. Lower ECE indicates better agreement.

Uncertainty used 1,000 school-level and 1,000 province-cluster bootstrap samples, emphasizing cluster resampling to preserve provincial dependence [36]. Learning curves compared training and province-grouped validation F_1 across fixed fractions. Global explanation combined CatBoost importance, holdout permutation importance, mean absolute SHAP, and

one local explanation; SHAP decomposed predictions against a reference expectation [37]. All explanations describe model behaviour, not causality.

3.7. Subgroup auditing, sensitivity analysis, and reproducibility

Holdout performance was audited by status, school-size quartile, and province; a status ablation repeated training without status. Sensitivity tests required positive enrolment and/or vocational-teacher counts and used alternative three-class and binary outcomes. Representative-school scenarios varied vocational-teacher counts with all dependent ratios and shares, measuring model response rather than intervention. Reproducibility records covered seeds, versions, identifiers, hashes, artefacts, calibration metadata, predictions, tables, figures, and a model card. Integrity checks verified province and school disjointness, probability sums, reload equivalence, and holdout hashes.

4. Results and Discussion

4.1. Data audit and descriptive analysis

Deterministic aggregation reduced 149,376 program records to 14,391 schools after removing 7,947 exact duplicates (5.3%). Table 1 reports the full sample flow and confirms that schools, not programs, were the modelling unit.

Table 1. Data audit and analytic sample flow

Sample-flow item	Count
Raw program-level rows	149,376
Exact duplicate rows removed	7,947
Rows after duplicate removal	141,429
Unique school records	14,391
Schools with primary C/B/A outcome	14,134
Training schools	12,598
Locked geographic holdout schools	1,536
Training provinces	31
Held-out provinces	8

Only 639 program rows (0.4%) lacked accreditation; 65,176 (43.6%) had zero students, while zero total-teacher and vocational-teacher values occurred in 0.3% and 3.7%. These zeros were treated as structural, not missing. Among 14,134 valid-label schools, B included 7,248 (51.3%), A 3,931 (27.8%), and C 2,955 (20.9%), motivating the use of class-balanced metrics. Status was associated with accreditation ($\chi^2 = 365.13$, $p < 0.001$, $V = 0.161$), and Table 2 gives the corrected continuous-feature comparisons.

Table 2. Kruskal-Wallis differences across accreditation classes

Variable	H statistic	ε^2	FDR-adjusted p
total_students	3821.51	0.270	<0.001
total_teachers	3542.11	0.251	<0.001
vocational_teachers_total	3346.01	0.237	<0.001
students_per_teacher	1690.28	0.120	<0.001
number_of_programs	2702.48	0.191	<0.001
number_of_competencies	2944.06	0.208	<0.001
active_programs	2244.22	0.159	<0.001
students_per_vocational_teacher	695.11	0.053	<0.001
field_concentration_hhi	327.62	0.023	<0.001

Table 2 shows the strongest separation for students ($\varepsilon^2 = 0.270$), total teachers ($\varepsilon^2 = 0.251$), and vocational teachers ($\varepsilon^2 = 0.237$). Median enrolment rose from 67 in C to 143 in B and 523 in A. Correlations linked students with total teachers ($\rho = 0.818$) and vocational

teachers (ρ= 0.755), and program with competency counts (ρ= 0.922), indicating that scale and breadth separated classes more than ratios or field concentration.

4.2. Model development and geographic validation

In Table 3, CatBoost plus province led random validation at F_1 = 0.599 but fell to 0.541 under grouped validation. Without province, the augmented model moved only from 0.562 to 0.543 and achieved the best grouped Macro F_1, showing that province identity improved within-fold fit more than unseen-region transfer.

Table 3. Model comparison under random and province-grouped cross-validation

Model	Feature set	Random F1	Grouped F1	Δ F1	Grouped QWK	Grouped AUC
CatBoost	Vocational augmented	0.562	0.543	-0.019	0.505	0.750
CatBoost	Vocational augmented plus province	0.599	0.541	-0.058	0.507	0.761
Random forest	Vocational augmented	0.565	0.538	-0.027	0.461	0.746
Multinomial logistic regression	Vocational augmented	0.560	0.535	-0.025	0.502	0.752
CatBoost	Core structural	0.540	0.531	-0.009	0.492	0.735
Ordinal cumulative logistic regression	Vocational augmented	0.391	0.378	-0.013	0.450	0.748
Dummy majority	Vocational augmented	0.226	0.230	0.004	0.000	0.500

Feature-set ablation favoured the augmented specification: it outperformed the core set in grouped Macro F1 and kappa, whereas adding province slightly reduced Macro F1. Repeated grouped evaluations remained above the majority baseline but varied with withheld provinces. Differences among CatBoost (0.543), random forest (0.538), and multinomial logistic regression (0.535) were small, so algorithmic dominance is not claimed. The ordinal model ranked classes reasonably (AUC 0.748) but had lower Macro F1 (0.378); the dummy scored 0.230. The learning plateau and persistent training-validation gap suggest that richer institutional indicators, rather than more records with the same features, are needed for geographic transfer.

4.3. Probability calibration and locked-holdout performance

Grouped out-of-fold scaling fitted T = 1.208 and improved probability quality without changing predicted classes. Table 4 reports raw versus calibrated metrics, and Figure 2 shows class-specific reliability on the locked holdout.

Table 4. Probability-calibration audit

Evaluation probabilities	Log loss	ECE
Training grouped OOF, raw	0.881	0.083
Training grouped OOF, calibrated	0.875	0.064
Locked holdout, raw	0.882	0.078
Locked holdout, calibrated	0.877	0.049

Table 4 shows that calibration improved probability quality on the locked holdout, reducing log loss from 0.882 to 0.877 and ECE from 0.078 to 0.049. However, Figure 2 shows that the improvement was not uniform across classes: B remained underconfident at middle probabilities, C was overconfident at moderate-to-high probabilities, while A was better calibrated mainly at low probabilities. Thus, although aggregate ECE improved, the calibrated probabilities were not equally reliable across all classes, as a single global temperature cannot fully correct these class-specific deviations. Class-B probabilities should therefore be interpreted cautiously, particularly when the model is applied to new geographic or temporal settings. Table 5 further shows moderate performance on the one-time holdout, with balanced accuracy of 0.586, Macro F1 of 0.533, quadratic weighted kappa of 0.499, and Macro ROC-AUC of 0.757. The difference between the probability-ranking performance and hard-label

classification results further suggests that the model is more suitable for probabilistic triage than for definitive class assignment.

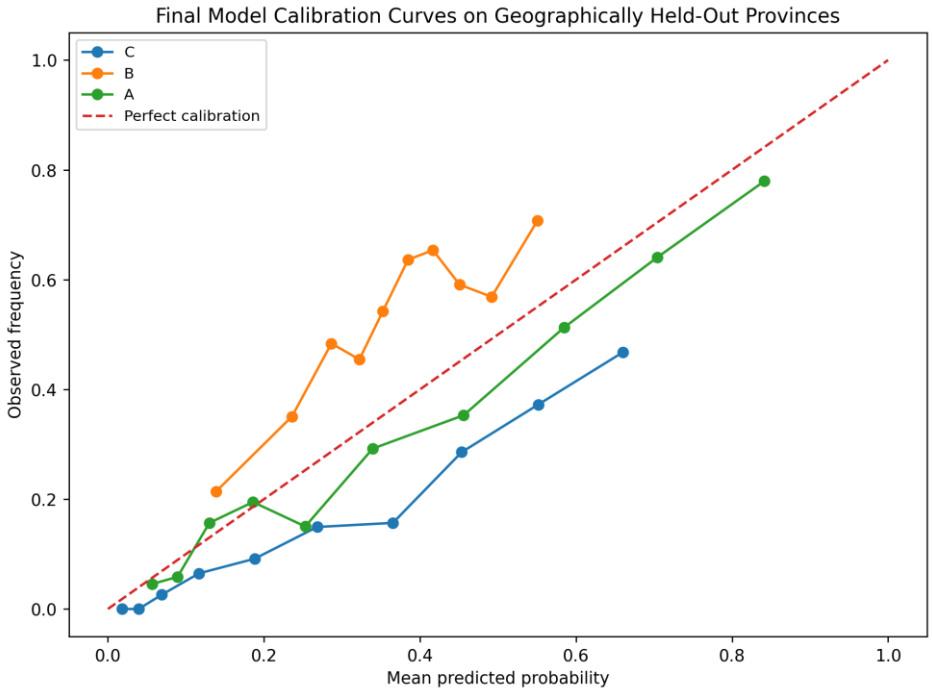

Figure 2. Class-specific calibration curves on the locked geographic holdout

Table 5. Final calibrated performance on the locked geographic holdout

Metric	Locked holdout value
Accuracy	0.538
Balanced accuracy	0.586
Macro precision	0.530
Macro recall	0.586
Macro F1	0.533
Quadratic weighted kappa	0.499
Matthews correlation	0.316
Macro ROC-AUC	0.757
Log loss	0.877
Multiclass Brier score	0.548
Expected calibration error	0.049

Table 6 identifies the main error: 65.7% of C and 69.3% of A schools were correct, but only 40.7% of B schools. B errors were nearly symmetric toward C and A, while extreme C-to-A and A-to-C errors were uncommon, largely preserving ordinal structure.

Table 6. Final holdout confusion matrix

Actual class	Predicted C	Predicted B	Predicted A
C	163	77	8
B	246	325	228
A	44	106	339

Table 7 confirms distinct failure profiles: A had the strongest $F_1 = 0.637$; C combined recall = 0.657 with precision = 0.360; and B paired the highest precision (0.640) with the lowest recall (0.407). One-versus-rest AUC was 0.817 for C, 0.647 for B, and 0.807 for A, again showing that the middle class was least separable.

Table 7. Class-specific performance on the locked holdout

Class	Support	Precision	Recall	F1
C	248	0.360	0.657	0.465
B	799	0.640	0.407	0.497
A	489	0.590	0.693	0.637

Class B overlaps both endpoints: some schools resemble C in scale and resource constraints, yet resemble A in program breadth and staffing. This explains its AUC of 0.647 and nearly symmetric displacement toward C and A; temporal mismatch between predictors and labels may add ambiguity. Because extreme errors were rare, the task remains ordinal, but hard class-B assignments are unsuitable for final decisions. These cases should be referred to assessors with the full probability vector and calibration evidence.

4.4. Uncertainty and explainability

Figure 3 and Table 8 show that province clustering widened transfer uncertainty. For Macro F_1, the school-level 95% interval was 0.510-0.559, versus 0.476-0.553 by province; balanced-accuracy intervals widened from 0.561-0.612 to 0.521-0.624, and the cluster kappa lower bound fell to 0.412. School-level resampling would therefore overstate precision for unseen provinces.

Table 8. Point estimates and bootstrap 95% confidence intervals

Metric	Point estimate	School-level 95% CI	Province-cluster 95% CI
Macro F1	0.533	0.510-0.559	0.476-0.553
Balanced accuracy	0.586	0.561-0.612	0.521-0.624
Quadratic weighted kappa	0.499	0.462-0.536	0.412-0.535

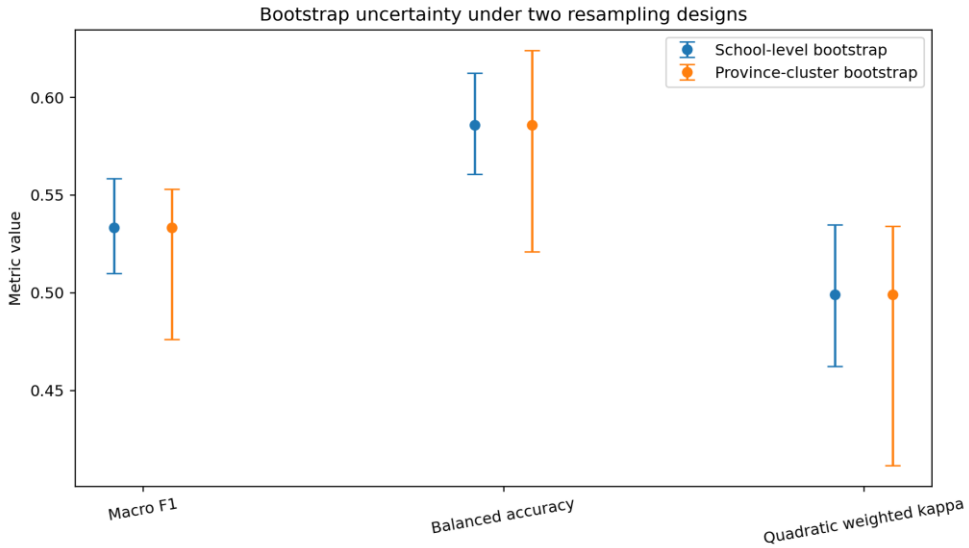

Figure 3. School-level and province-cluster bootstrap confidence intervals

Figure 4 shows global SHAP attribution dominated by institutional scale and breadth: log-enrolment ranked first with mean absolute SHAP = 0.287, followed by teachers, programs, vocational teachers, status, and competencies (0.092-0.113). Table 9 complements this with holdout permutation importance. Table 9 shows the largest Macro F1 loss (0.0298) after shuffling enrolment, followed by status (0.0200) and vocational-field count (0.0076); most other decreases were below 0.004 and near their permutation standard deviations. SHAP distributes attribution within the model, whereas permutation importance identifies features whose disruption consistently harms holdout performance.

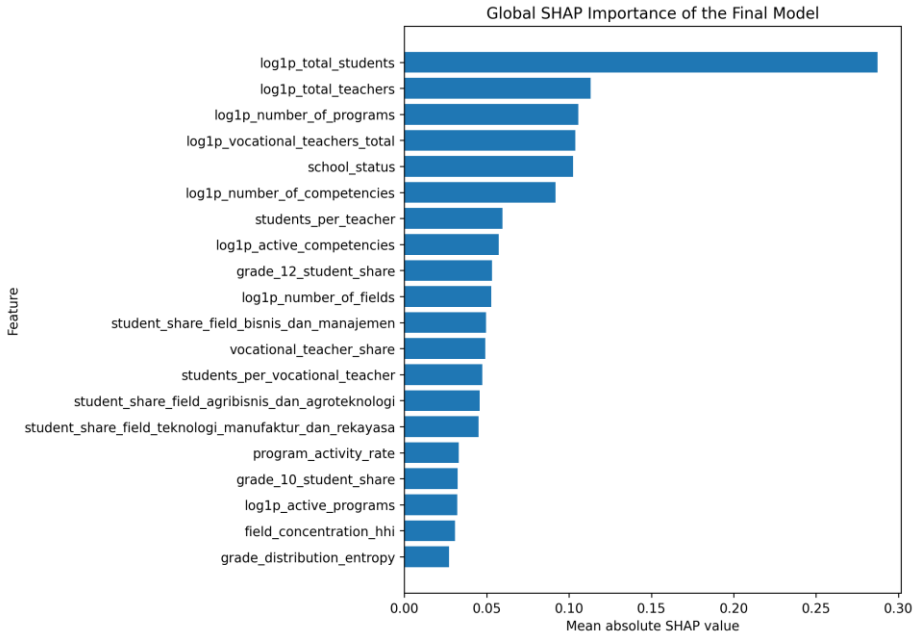

Figure 4. Global SHAP importance on the locked geographic holdout

Table 9. Holdout permutation importance for the ten highest-ranked features

Feature	Mean Δ Macro F1	SD
log1p_total_students	0.0298	0.0111
school_status	0.0200	0.0034
log1p_number_of_fields	0.0076	0.0032
student_share_field_pariwisata	0.0033	0.0019
field_concentration_hhi	0.0029	0.0031
student_share_field_energi_dan_pertambangan	0.0014	0.0019
student_share_field_teknologi_manufaktur_dan_rekayasa	0.0013	0.0018
student_share_field_agribisnis_dan_agroteknologi	0.0009	0.0033
student_share_field_teknologi_konstruksi_dan_properti	0.0005	0.0010
grade_11_student_share	0.0004	0.0041

4.5. Subgroup auditing and sensitivity analysis

Table 10 descriptively shows higher public-school Macro F_1, kappa, and AUC. The public-private Macro F_1 gap was 0.065, but its province-cluster 95% interval (-0.001 to 0.213) crossed zero. Calibration ranked groups differently, so stronger discrimination did not imply lower calibration error.

These differences indicate subgroup heterogeneity, not algorithmic unfairness: the gap interval crossed zero, and status is entangled with province, scale, staffing, and label processes. The audit identifies monitoring priorities, not discrimination or causality.

Table 10. Performance by school status

Status	n	Macro F1	Balanced accuracy	QWK	Macro AUC	ECE
Public	592	0.572	0.600	0.556	0.785	0.066
Private	944	0.508	0.572	0.464	0.741	0.057

Removing status slightly reduced Macro F_1 from 0.533 to 0.528, kappa from 0.499 to 0.483, and AUC from 0.757 to 0.751, while ECE rose from 0.049 to 0.055. The public-private Macro F_1 gap remained 0.061, and class-C recall disparity increased, implying that correlated structural features retained status-related information.

Figure 5A shows non-monotonic quartile Macro F_1 values of 0.383, 0.384, 0.435, and 0.320; the largest quartile had only 6 C schools and none were correctly identified. Figure 5B reports province Macro F_1 from 0.426 in Sulawesi Tengah to 0.561 in Riau. Papua Barat Daya's 37 schools produced a wide interval, while Papua Pegunungan's 21 schools were too few for a stable three-class estimate. Aggregate performance therefore masks compositional and regional heterogeneity.

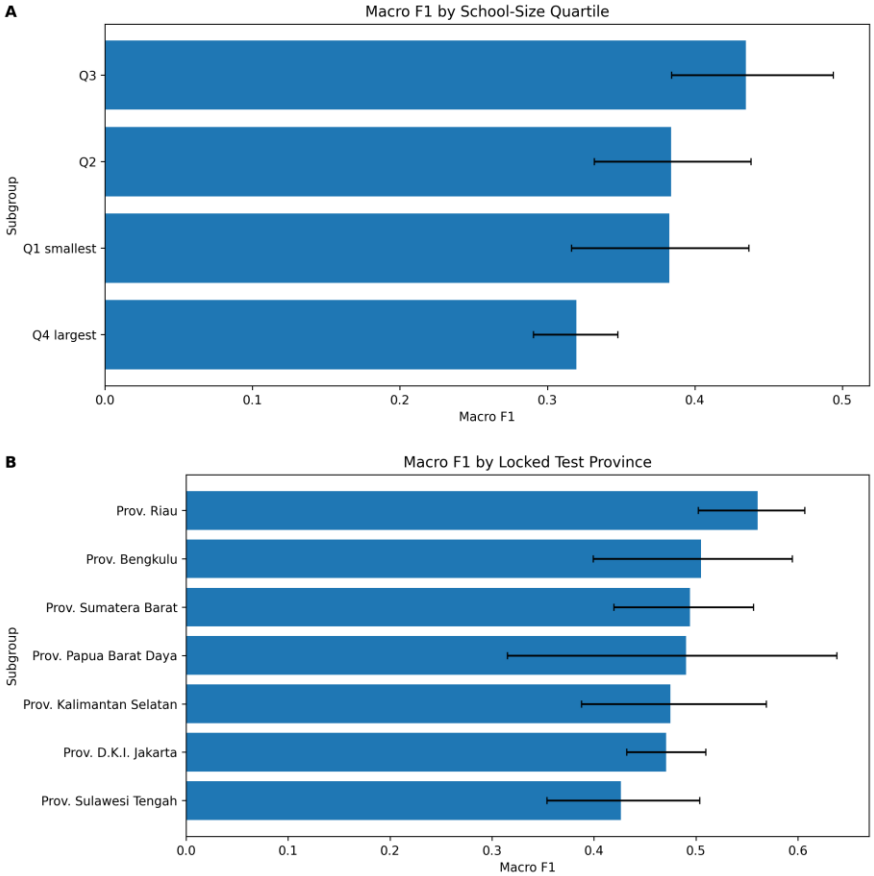

Figure 5. Subgroup performance on the locked geographic holdout: (**a**) Macro F1 across school-size quartiles; (**b**) Macro F1 across held-out provinces with reportable three-class estimates

Table 11 shows stable structural-zero sensitivity: Macro F_1 stayed between 0.533 and 0.536, AUC between 0.757 and 0.759, and ECE reached about 0.044 under the strictest restriction. Unusual zero-count records did not drive the holdout result.

Table 11. Sample-restriction sensitivity analysis

Sample restriction	n	Macro F1	Balanced accuracy	QWK	Macro AUC	ECE
All holdout schools	1536	0.533	0.586	0.499	0.757	0.049
Schools with students	1447	0.533	0.581	0.497	0.757	0.046
Schools with vocational teachers	1457	0.534	0.581	0.498	0.757	0.044
Schools with students and vocational teachers	1398	0.536	0.579	0.498	0.759	0.044

Table 12 shows that adding not-accredited schools to C changed Macro F_1 only from 0.533 to 0.538 and AUC from 0.757 to 0.761. The binary C/not-accredited versus A/B task achieved ROC-AUC = 0.814 and accuracy = 0.757, but positive-class F_1 remained 0.470. Broad risk ranking was easier than separating adjacent grades, yet still missed at-risk schools. In a complementary analysis, Figure 6 shows a gradual directional response as standardized vocational teachers rose from 0 to 20 with dependent features updated. Mean P(A) increased from 0.349 to 0.445, mean P(C) fell from 0.295 to 0.216, and mean P(B)

changed from 0.356 to 0.340; P(A) crossed P(B) after about one added teacher. This is model responsiveness, not a causal staffing effect.

Table 12. Outcome-definition sensitivity analysis

Outcome definition	Task	Test n	Macro F1	Balanced accuracy	AUC	Positive-class F1
Primary: C, B, A	three-class	1536	0.533	0.586	0.757	
Secondary three-class: C/Not Accredited, B, A	three-class	1541	0.538	0.586	0.761	
Binary exploratory risk: C/Not Accredited versus A/B	binary	1541	0.656	0.717	0.814	0.470

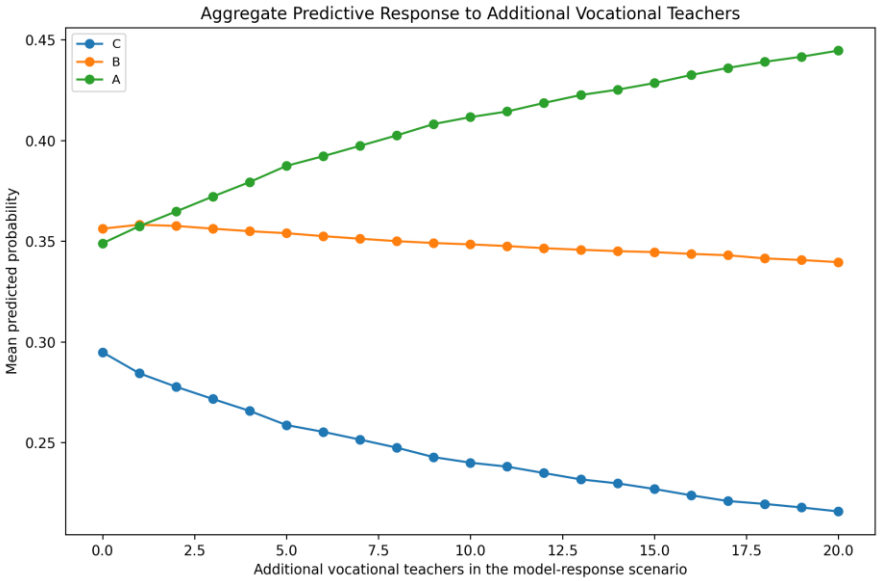

Figure 6. Aggregate predictive response to standardized increases in vocational-teacher counts

4.6. Discussion

The three validation regimes answer progressively stricter questions. Random stratified cross-validation estimates performance when training and validation folds share the national and provincial mixture, so it can benefit from relationships and geographic composition that recur across folds. Province-grouped cross-validation instead removes province overlap and approximates transfer to a region absent from the corresponding training fold, while the locked geographic holdout is stricter because eight named provinces remained untouched until the final evaluation. The decline was largest when province was included directly (Macro F1 0.599 under random validation versus 0.541 under grouped validation), whereas the specification without province declined less (0.562 to 0.543); the final locked-holdout Macro F1 was 0.533. This pattern indicates that part of the learned relationship is context-dependent: province identity and province-linked school composition improved within-distribution fit more than transferable discrimination. Random validation is therefore informative for the narrower within-distribution question, but good performance under that design does not establish geographic transferability [10]. Because the holdout was purposively constructed rather than sampled to represent every province, its result should also be interpreted as evidence for transfer to the observed held-out contexts, not as a universal national guarantee.

The proximity of the province-grouped scores for CatBoost (0.543), random forest (0.538), and multinomial logistic regression (0.535) reinforces this interpretation. These classifiers differ substantially in their capacity to represent nonlinearities and interactions, yet they converged on a similar geographic-validation ceiling when supplied with the same administrative feature space. CatBoost remained the selected model because it provided the highest grouped Macro F1 and handled the mixed tabular inputs efficiently, but the small differences do not support algorithmic dominance. In conjunction with the learning curve, which plateaued while a training-validation gap persisted, the comparison suggests that classifier sophistication is not the primary constraint. Adding more records with the same structure is unlikely by itself to recover information that the features do not encode. More transferable gains are

likely to require temporally aligned indicators of governance, instructional quality, facilities, student outcomes, financing, industry partnerships, and assessor evidence, rather than a more complex classifier applied to the existing administrative variables.

The class-specific errors, calibration deviations, and cluster-aware intervals describe related forms of uncertainty. Class B is an intermediate accreditation category and therefore borders both C and A; schools in B can resemble C in scale or resource constraints while resembling A in program breadth or staffing. Consistent with this overlap, B had recall of 0.407 and one-versus-rest AUC of 0.647, and its errors were distributed almost symmetrically toward C and A, whereas direct C-to-A and A-to-C errors were uncommon. The class-specific reliability curves add a probability perspective: B was underconfident at middle probabilities, C was overconfident at moderate-to-high probabilities, and A was better calibrated mainly at low probabilities. These opposing deviations explain why one global temperature improved aggregate ECE without making every class uniformly reliable [12]. They are compatible with overlapping decision boundaries and potentially ambiguous or temporally misaligned cases, although they do not by themselves prove label error. Finally, the wider province-cluster interval for Macro F1 (0.476-0.553) than the school-level interval (0.510-0.559) shows that uncertainty increases when entire geographic contexts, rather than nominally independent schools, are allowed to vary [36]. Taken together, these findings identify class-B, low-margin, and geographically atypical cases as priorities for human review rather than as candidates for automatic classification.

The subgroup, ablation, and sensitivity results caution against interpreting a single aggregate score as a property of the model alone. Removing school status had only a modest effect on overall performance, while the public–private Macro F1 gap remained 0.061 and the class-C recall disparity increased. This indicates that correlated structural features, such as enrolment, staffing, program breadth, and province, may still encode status-related information. However, this does not establish discrimination or causality because the cluster interval crossed zero [16]. Excluding structural-zero records barely changed the three-class results, whereas redefining the outcome as C/not-accredited versus A/B increased ROC-AUC to 0.814 but yielded a positive-class F1 of only 0.470. Thus, model performance must be interpreted according to the specific outcome definition and operational error costs.

Operationally, the model can support preliminary review prioritization through a bounded triage workflow. First, review teams may use the calibrated probability vector, rather than only the highest-probability class, to rank schools for initial document checking; for a risk-oriented workflow, the relevant class probability may be used to order a queue but must not be treated as an accreditation decision. Second, class-B assignments, low maximum probabilities, or small differences between the two largest class probabilities can flag ambiguous cases for a more comprehensive assessor review. Third, province-level and subgroup performance, calibration, sample size, and province-cluster uncertainty should accompany the score so that users can escalate cases from sparsely represented or unstable contexts and trigger local recalibration or revalidation. Any operational thresholds must be prespecified according to review capacity and the relative cost of missed at-risk schools, then validated before use. The model thus supports screening and triage, not autonomous accreditation or replacement of professional judgment [25], [38]. Moreover, because predictors and labels came from the same cross-sectional snapshot, the reported results are contemporaneous classification results rather than evidence of prospective prediction. Prospective use would require temporally ordered data, external temporal and geographic validation, drift monitoring, and recalibration before deployment.

5. Conclusions

This study establishes a reproducible, leakage-controlled, and geographically validated framework for using administrative data in Indonesian vocational-school accreditation screening, integrating school-level aggregation, province-aware validation, calibration, explainability, subgroup auditing, sensitivity analysis, and reproducibility controls. Administrative data provide a meaningful but incomplete signal for prioritizing institutional review, while uneven performance across classes, school groups, and provinces precludes autonomous decisions. Holdout performance (Macro F1 = 0.533, balanced accuracy = 0.586, class-B recall = 0.407, and ECE = 0.049) supports probability-based triage rather than final class assignment and requires continued monitoring of class reliability, subgroup performance, and

geographic drift. Interpretation is limited by the cross-sectional design, potentially imperfect or temporally misaligned labels, restricted geographic coverage, and missing indicators of leadership, instructional quality, facilities, student outcomes, financing, industry partnerships, and assessor observations. Use should therefore remain human-supervised, transparent, and periodically revalidated. Future studies should incorporate temporally aligned indicators, validate across more provinces and years, and assess whether calibrated predictions improve professional review without reinforcing institutional inequities. Accordingly, the present results should be understood as contemporaneous classification, not evidence of prospective prediction; prospective deployment requires temporally ordered evaluation and external revalidation.

Author Contributions: Conceptualization, M.R.M., P.S., and P.; methodology, M.R.M., P.S., D.N., and A.H.S.; software, M.R.M., Y.A.F., and D.N.; validation, P.S., P., D.N., and A.H.S.; formal analysis, M.R.M. and D.N.; investigation, M.R.M. and Y.A.F.; resources, M.R.M., Y.A.F., and A.H.S.; data curation, M.R.M. and Y.A.F.; writing original draft, M.R.M.; writing—review and editing, P.S., P., Y.A.F., D.N., and A.H.S.; visualization, M.R.M. and Y.A.F.; supervision, P.S. and P.; project administration, M.R.M. and P.S.; funding acquisition, M.R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Directorate General of Higher Education, Research, and Technology (Ditjen Diktiristek) through the *Pendidikan Magister menuju Doktor untuk Sarjana Unggul* (PMDSU) program under contract number 249.2/DST/UN34.9/T/PT.01.03/2026.

Data Availability Statement: The source data are publicly available through the Kemendikdasmen vocational-school portal at https://smk.kemendikdasmen.go.id/data-sekolah. The processed dataset, analysis code, predictions, and supporting reproducibility materials are publicly available at https://doi.org/10.5281/zenodo.22123602. No personally identifiable student data were used.

Acknowledgments: The authors thank the Directorate General of Higher Education, Research, and Technology for support through the *Pendidikan Magister menuju Doktor untuk Sarjana Unggul* (PMDSU) program and Kemendikdasmen for providing the public vocational-school data. Generative AI tools were used only for language refinement; all analyses, interpretations, and final content were verified by the authors.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] T. Vanpham and L. Dangnguyen, "Management of Self-Assessment and Quality Accreditation Activities at Vocational Colleges in Vietnam: Policy, Practice and Some Solutions," *JETT*, vol. 14, no. 3, pp. 326–334, 2023, doi: 10.47750/jett.2023.14.03.04.

[2] B. Susetyo and A. Lie, "Quality indicators of vocational schools: Partial least squares-SEM for small sample," *MethodsX*, vol. 15, p. 103620, 2025, doi: 10.1016/j.mex.2025.103620.

[3] OECD, "Teachers and leaders in vocational education and training," OECD Publishing, 2021. doi: 10.1787/59d4fbb1-en.

[4] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, and Z. Du, "Artificial intelligence in education: A systematic literature review," *Expert Syst. Appl.*, vol. 252, p. 124167, Oct. 2024, doi: 10.1016/j.eswa.2024.124167.

[5] R. S. Baker and A. Hawn, "Algorithmic bias in education," *Int. J. Artif. Intell. Educ.*, vol. 32, no. 4, pp. 1052–1092, 2022, doi: 10.1007/s40593-021-00285-9.

[6] A. Mathrani, T. Susnjak, G. Ramaswami, and A. Barczak, "Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics," *Comput. Educ. Open*, vol. 2, p. 100060, 2021, doi: 10.1016/j.caeo.2021.100060.

[7] D. D. Lee and S. J. Cho, "Predicting the outcomes of the Korean national accreditation system for higher education institutions: A method using disclosure data for outsiders," *Asia Pacific Educ. Rev.*, vol. 22, no. 4, pp. 715–728, 2021, doi: 10.1007/s12564-021-09710-z.

[8] N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022)," *Educ. Inf. Technol.*, vol. 28, no. 7, pp. 8299–8333, Jul. 2023, doi: 10.1007/s10639-022-11536-0.

[9] P. W. Koh *et al.*, "WILDS: A benchmark of in-the-wild distribution shifts," *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. PMLR, pp. 5637–5664, 2021.

[10] T. Takada *et al.*, "Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets," *J. Clin. Epidemiol.*, vol. 137, pp. 83–91, 2021, doi: 10.1016/j.jclinepi.2021.03.025.
[11] Y. Ovadia *et al.*, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," *Advances in Neural Information Processing Systems 32*, vol. 32. Curran Associates, pp. 13969–13980, 2019.
[12] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: A survey on how to assess and improve predicted class probabilities," *Mach. Learn.*, vol. 112, no. 9, pp. 3211–3260, 2023, doi: 10.1007/s10994-023-06336-7.
[13] U. Johansson, T. Löfström, and H. Boström, "Calibrating multi-class models," *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, vol. 152. PMLR, pp. 111–130, 2021.
[14] R. Alfredo *et al.*, "Human-centred learning analytics and AI in education: A systematic literature review," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100215, 2024, doi: 10.1016/j.caeai.2024.100215.
[15] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
[16] I. D. Raji *et al.*, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, pp. 33–44, 2020. doi: 10.1145/3351095.3372873.
[17] R. Ashmore, R. Calinescu, and C. Paterson, "Assuring the machine learning lifecycle: Desiderata, methods, and challenges," *ACM Comput. Surv.*, vol. 54, no. 5, pp. 111, 1–39, 2021, doi: 10.1145/3453444.
[18] G. S. Collins *et al.*, "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, 2024, doi: 10.1136/bmj-2023-078378.
[19] H. Crompton and D. Burke, "Artificial intelligence in higher education: the state of the field," *Int. J. Educ. Technol. High. Educ.*, vol. 20, no. 1, p. 22, Apr. 2023, doi: 10.1186/s41239-023-00392-8.
[20] G. Ramaswami, T. Susnjak, A. Mathrani, and R. Umer, "Use of predictive analytics within learning analytics dashboards: A review of case studies," *Technol. Knowl. Learn.*, vol. 28, no. 3, pp. 959–980, 2023, doi: 10.1007/s10758-022-09613-x.
[21] Noripansyah, A. Kadir, D. Kusumaningsih, and Haderiansyah, "Accreditation prediction of early childhood education institutions using machine learning techniques," *J. Tek. Inform.*, vol. 4, no. 3, pp. 647–655, 2023, doi: 10.52436/1.jutif.2023.4.3.999.
[22] M. B. Musthafa, N. Ngatmari, C. Rahmad, R. A. Asmara, and F. Rahutomo, "Evaluation of university accreditation prediction system," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 732, no. 1, p. 12041, 2020, doi: 10.1088/1757-899X/732/1/012041.
[23] A. S. K. Rashid, "The extent of the teacher academic development from the accreditation evaluation system perspective using machine learning," *J. Exp. Theor. Artif. Intell.*, vol. 35, no. 4, pp. 535–555, 2023, doi: 10.1080/0952813X.2021.1960635.
[24] M. Jannah and P. P. Izati, "Evaluation of classification methods for predicting junior high school accreditation ranks in Indonesia," *Int. J. Eng. Comput. Sci. Appl.*, vol. 5, no. 1, pp. 43–54, 2026, doi: 10.30812/ijecsa.v5i1.6032.
[25] B. I. Igoche, O. Matthew, P. Bednar, and A. Gegov, "Integrating Structural Causal Model Ontologies with LIME for Fair Machine Learning Explanations in Educational Admissions," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 65–85, Jun. 2024, doi: 10.62411/jcta.10501.
[26] J. P. Ntayagabiri, Y. Bentaleb, J. Ndikumagenge, and H. El Makhtoum, "A Comparative Analysis of Supervised Machine Learning Algorithms for IoT Attack Detection and Classification," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 395–409, Feb. 2025, doi: 10.62411/jcta.11901.
[27] H. Suresh and J. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, Oct. 2021, pp. 1–9. doi: 10.1145/3465416.3483305.
[28] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2022, doi: 10.1145/3457607.
[29] H. Xia and K. Chen, "Bridging algorithmic prediction and teacher judgment: An explainable AI framework with institutional fairness calibration for early warning systems," *Front. Educ.*, vol. 11, p. 1881571, 2026, doi: 10.3389/feduc.2026.1881571.
[30] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J. Am. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952, doi: 10.1080/01621459.1952.10483441.
[31] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.
[32] D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography (Cop.).*, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.
[33] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Jan. 2018. [Online]. Available: http://arxiv.org/abs/1706.09516
[34] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
[35] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning,* Aug. 2017, pp. 1321–1330. [Online]. Available: http://arxiv.org/abs/1706.04599
[36] B. Efron and R. J. Tibshirani, "An introduction to the bootstrap New York," *NY Chapman Hall*, vol. 473, 1993.
[37] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Nov. 2017, pp. 4768–4777. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230
[38] G.-J. Hwang, H. Xie, B. W. Wah, and D. Gašević, "Vision, challenges, roles and research issues of artificial intelligence in education," *Comput. Educ. Artif. Intell.*, vol. 1, p. 100001, 2020, doi: 10.1016/j.caeai.2020.100001.