SUST Journal of Science and Technology Vol. 35(2) 2025, 39 - 54

SUST Journal of Science and Technology

journal homepage: https://journals.sust.edu

Predicting At-Risk Primary Students in Sylhet, Bangladesh: Grading Artifacts, Half-Yearly Signals, and the Class 3 Failure Breakpoint

Fazle Rabbi Rakib [1*], Mahfuzur Rahman Emon [1], Md. Mehedi Hasan [2]

[1]*Institute of Information and Communication Technology,* [2]*Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet – 3114, Bangladesh*

ARTICLE INFO

Keywords:

At-risk prediction
Bangladesh
Conjunctive grading
Early warning systems
Machine learning
Primary education
Curriculum transition

ABSTRACT

Primary school failure statistics in low- and middle-income countries are often reported without distinguishing between students who fail broadly and those who fail a single subject under conjunctive grading rules that collapse all failure into one label. This paper reports findings from a multi-institutional study of 4,632 primary school students (Classes 1–5) across nine schools in Sylhet, Bangladesh, spanning government, institutional, private, Islamic, girls-only, and university-affiliated institutions. Three questions guide the analysis. First, can half-yearly examination records reliably identify students who will remain at risk by year-end, and which features carry the strongest predictive signal? Second, what share of students recorded as failing have failed only one subject under a rule that assigns GPA = 0 regardless of how many subjects a student actually fails? Third, does the sharp increase in Mathematics failure at Class 3 replicate independently across different types of schools which points to the national curriculum as the primary driver rather than any individual school factor? Across cohort years, an XGBoost classifier trained on one year's records sustains 74.9% at-risk recall when applied to the following year. Within the same year, Random Forest achieves 88.8% accuracy (AUC = 0.928) under cross-validation, reducing false-alarm interventions by 65% compared with the existing half-yearly flagging approach. An analysis of GPA = 0 records shows that 38.3% of the students had single-subject failures. Mathematics was the failed subject in 55.5% of these cases. Class 3 Mathematics failure rates jumped from 8.5% to 28.8% across the sampled schools. This pattern suggests a shared curriculum-related transition rather than a school-specific factor. The findings suggest two possible responses: a school-deployable early warning system using existing mid-year records and a policy reform replacing the conjunctive grading rule with a compensatory calculation that would reclassify approximately 38% of current GPA = 0 students without changing their actual performance.

1. Introduction

In Bangladesh, the total enrollment rate in primary education had reached 98.0% by 2022 (UNESCO, 2023). However, primary completion rates and grade-level learning benchmarks have not kept pace with enrollment, particularly in urban secondary cities and rural areas outside Dhaka. The gap between enrollment and learning varies across the education system. It is concentrated in particular grades, subjects, and types of institutions, differences that aggregate statistics can obscure.

This study examines three related challenges. The first concerns the information available to schools. Although schools receive mid-year examination results for every student in July, they typically lack a systematic way to identify which students will continue to struggle in December and which will recover without further intervention. Without this distinction, schools may end up intervening with every flagged student, resulting in a high number of false alarms. On the other hand, limiting interventions may leave stu-

* Corresponding author.
E-mail address: rakib-iict@sust.edu (F. R. Rakib)

https://doi.org/10.63512/sustjst.2025.2005

Copyright© 2025 SUST. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

dents with persistent difficulties. The second challenge concerns how academic failure is labeled. Under Bangladesh's national grading framework, the conjunctive rule assigns GPA = 0 to a student who fails even one subject, regardless of their performance in the remaining subjects. In this dataset, 22.1% of students received GPA = 0 at the end of the year. However, this figure does not distinguish between students who failed only one subject and those who failed several subjects. The third concern is the increase in Mathematics failure across schools. The failure rate approximately triples between Classes 2 and 3 and continues to rise through Class 5 in the sampled schools. This pattern is observed across institutions with different funding sources, teacher attributes, class sizes and institutional types. This suggests that the increase may be related to the national curriculum rather than the independent factors of each school.

All three identified problems are connected. The grading rule leads the Class 3 Mathematics failures into a GPA = 0 outcome. The GPA = 0 classification then enters mid-year records as a binary flag, which an early warning model must learn to interpret. However, it does not distinguish between students who fail only one subject and those who experience broader academic difficulties. Acting on mid-year flags without this distinction, or attributing the Class 3 spike to school-level factors, can lead to poorly targeted interventions and misguided policy decisions.

Three research questions (RQs) motivate the analysis:

RQ1: Can half-yearly examination records reliably identify students who will remain at risk by year-end and which features carry the strongest predictive signal?

RQ2: What fraction of GPA = 0 outcomes are artifacts of the national conjunctive grading rule rather than actual broad academic failure?

RQ3: Does the Class 3 Mathematics failure-rate increase replicate consistently across the sampled schools and is the pattern more consistent with a shared curriculum transition than with school-level resource differences?

This paper makes three contributions. First, it presents the first multi-institutional study of at-risk prediction using half-yearly examination data from different types of primary schools in Sylhet, Bangladesh. Second, it quantifies the grading-rule artifact in the GPA = 0 population and identifies Mathematics at the Class 3 transition as the dominant driver. Finally, it shows that both issues can be addressed using existing school data. The early warning model relies on records that schools already collect and the proposed grading reform requires no additional assessment.

2. Related Literature

2.1. Early Warning Systems in Education

Early warning systems (EWS) designed to flag students at risk of academic failure have been studied extensively at the secondary and tertiary levels in high-income countries (Bowers et al., 2013; Heppen & Therriault, 2008). These systems typically combine attendance, behavior, and course performance indicators into composite indices or machine learning classifiers. In the primary school context, EWS evidence is thinner, particularly in low- and middle-income countries where digital record systems are newer and data quality is more variable.

Since the early 2010s, machine learning has been increasingly used to predict student performance. Studies reported better performance from Random Forest and gradient-boosted tree models than logistic regression (Adnan et al., 2021; Kumar et al., 2024; Zhang et al., 2021). Similar findings have been reported in student dropout prediction. A survey within developing-country schools found that ensemble methods generally performed the best at predicting the dropouts (Mduma et al., 2019). A study in Tanzania also reported better dropout prediction performance from automated ensemble methods than from baseline classifiers (Mnyawami et al., 2022).

In a systematic review of 68 student performance prediction studies, Albreiki et al. (2021) found that fewer than one in five reported performance across different years or cohorts, which raises concern on model evaluation practices. This makes it difficult to assess whether the reported accuracy is maintained when models are applied to new student cohorts. The present study addresses this limitation through temporal validation, training on the 2024 cohorts and testing on the 2025 cohort.

In our data, failure breadth (students who were failing more subjects at mid-year) ranked as a strong predictor of chronic year-end risk than any single-subject mark threshold.

2.2. Grading Rules and Measurement Artifacts in Reported Failure Rates

The choice between conjunctive and compensatory grading rules has long been debated in educational settings (Haladyna & Hess, 1999). Conjunctive grading, where passing all component assessments is required for an overall passing grade, is conceptually justified as ensuring minimum competence in every domain. When one component has a substantially higher failure rate than others, however, a conjunctive rule systematically conflates students who fail broadly with students who fail narrowly: the overall failure rate is bounded below by the highest component failure rate and grows as high-failure subjects co-occur. This population-level implication of the conjunctive rule has not previously been quantified at the individual student record level for Bangladesh primary school data.

In Bangladesh, the conjunctive structure of the Primary Education Completion Examination and its class-level precursors means that a single-subject failure propagates directly to a GPA = 0 outcome. National Student Assessment data for Grades 3 and 5 have documented persistent Mathematics learning gaps without disaggregating the grading-rule contribution to reported failure rates (Directorate of Primary Education & ACER, 2022). Studies on learning disparities across Bangladesh primary schools document that measured outcomes are sensitive to household wealth, over-age status, and school context (Ahmmed & Uddin, 2022; Asadullah & Chaudhury, 2015), but the proportion of reported GPA = 0 outcomes attributable to single-subject rule artifacts has not been quantified at the individual student level. This paper provides such a quantification. Grade retention, itself triggered by conjunctive pass-fail rules, has been shown in a recent meta-analysis to produce no aver-

age long-term academic benefit (Goos et al., 2021), reinforcing the case for examining whether the failures that trigger retention decisions are genuinely broad.

2.3. Grade-Level Curriculum Transitions and Mathematics

The transitional period from lower to upper primary school is well-documented as a time of elevated academic risk. Alspaugh (1998) showed that achievement loss associated with school-level transitions is consistent and measurable. Whereas, Eccles et al. (1993) linked declines in motivation during these transitions to a mismatch between students' developmental needs and curriculum demands. In the mathematics domain specifically, a systematic review of primary-to-secondary transition research found that Mathematics is the subject most frequently associated with attainment discontinuities at grade boundaries, driven by the qualitative shift from procedural to conceptual demands (Kaur et al., 2022).

In the Bangladesh primary curriculum, Class 3 introduces multiplication of multi-digit numbers, basic fractions, and abstract word problems that represent a qualitative shift from the number-recognition and additive reasoning tasks of Classes 1–2. National Student Assessment data for Grades 3 and 5 have documented achievement gaps at this transition at the national level (Directorate of Primary Education & ACER, 2022); our contribution is the first multi-institutional replication of this breakpoint at the classroom section level, across nine heterogeneous institution types.

2.4. Teacher Resources and Student Outcomes in South Asia

The relationship between teacher experience and student outcomes in low- and middle-income countries is complex and often non-linear. A comprehensive review of school resource effects across developing countries found that most school inputs have heterogeneous and context-dependent effects: teacher experience benefits are concentrated in early career years, while evidence for later-career experience and for resource inputs such as textbook provision is mixed and sensitive to whether materials match students' actual learning levels (Glewwe et al., 2013; Glewwe & Muralidharan, 2016). In Bangladesh specifically, Asadullah and Chaudhury (2015) found that school-level teacher quality explained only a modest share of within-district learning variation once student socioeconomic background was accounted for, and Ahmmed et al. (2022) found teacher quality to be a significant structural determinant of student achievement across 128 Bangladesh schools ($\beta = 0.588$) but noted that quality measures based on credentials alone are weak proxies for instructional effectiveness. Metzler and Woessmann (2012), exploiting within-teacher variation in Peru, showed that subject-specific teacher knowledge had a causal effect on student achievement in that subject, a finding relevant to the Mathematics-specific failure pattern we report here.

School-level teacher experience, STEM teacher count, class size and teacher-student ratio enter the analysis as contextual variables in the early warning model rather than as primary explanatory variables. Their cross-school associations with at-risk rates are reported in Section 4.1 and interpreted with appropriate caution in Section 5.4. The ecological associations observed across these nine schools, including a 4.0-fold range in average teacher experience and a ∼3-fold range in mean class size, are offered as descriptive context rather than as evidence that reallocating these inputs would produce corresponding changes in outcomes.

3. Data and Methods

3.1. Setting and Sample

The study draws on academic records collected from nine primary schools in Sylhet, Bangladesh, covering Classes 1–5 across academic years 2024 and 2025, though temporal coverage is not uniform across all schools (see Section 4.4 for the school-level breakdown). Sylhet is the fourth-largest city in Bangladesh and the administrative center of Sylhet Division, a region with a sizeable expatriate diaspora and a mix of urban government, private, and community-based schools.

Schools were selected through criterion-based purposive sampling from an enumerable frame of 160 registered institutions in the Sylhet district school directory, of which 114 use the Pathshala digital school management platform. The nine selected schools span six institutional types: government (n = 1), girls-only government (n = 1), private (n = 2), Islamic combined school (n = 1), institutional (n = 3; one attached to a government medical college, one to the central bank, and one to a public research institute), and university-affiliated (n = 1). This coverage was chosen deliberately to test whether findings replicate across governance and management structures. The sample is described as a criterion-purposive pilot study drawing from an enumerable frame; it is not a probability sample and findings should not be generalized to Bangladesh as a whole without replication.

The nine schools are: Osmani High School (OHS), Bangladesh Bank School (BBS), Sylhet Udayan School (UDS), Bangladesh Tea Research Institute High School (BTRI), Victoria High School (VHS), Sylhet Government Pilot High School (SGPHS), Govt. Agragami Girls' High School (AGHS), Shahjalal Jamia Islamia School and College (SJIS), and Shahjalal University School and College (SUSC). These abbreviations are used throughout.

The final dataset comprises 4,632 annual student records across the nine schools: OHS (n = 122), BBS (n = 737), UDS (n = 603), BTRI (n = 232), VHS (n = 476), SGPHS (n = 827), AGHS (n = 528), SJIS (n = 939), and SUSC (n = 168). For the early warning model (Section 3.3), the matched half-yearly plus annual subset was used (n = 4,470; students with no half-yearly record were excluded).

3.2. Data Sources and Variables

Academic result data were obtained through three different channels: direct download from school Pathshala portals (seven schools), extraction from PDF result sheets (SUSC) and structured Excel files provided directly by schools (OHS). Data were extracted, cleaned and consolidated using a standardized pipeline. The complete pipeline is illustrated in Figure 1. All student-identifying fields were replaced with SHA-256-based pseudonymous identifiers before any analysis, keeping the original records unmodified. The de-identified dataset is publicly available on Zenodo

Table 1. School-level descriptive statistics across nine Sylhet primary schools

School	Type	Classes	N	At-Risk (%)	Mean GPA (Ann.)	Mean GPA (HY)	Teachers	STEM Tchrs	Avg Tchr Exp (yr)	Mean Class Size
OHS	Institutional	1–5	122	12.3	3.59	2.32	33	7	17.0	24.4
BBS	Institutional	1–5	737	23.2	3.32	2.39	28	3	16.5	37.4
UDS	Private	1–5	603	41.0	2.41	2.73	27	2	4.2	49.7
BTRI	Institutional	1–5	232	37.5	2.58	2.03	16	4	14.9	24.5
VHS	Private	1–5	476	20.6	3.64	3.17	39	9	11.7	48.3
SGPHS	Government	3–5	827	22.7	3.20	2.55	46	6	10.3	70.7
AGHS	Government (girls)	3–5	528	11.6	3.80	2.79	48	11	13.6	66.7
SJIS	Islamic	1–5	939	12.0	4.05	3.38	97	22	11.4	44.6
SUSC	University-affiliated	1–5	168	26.2	3.27	2.70	36	9	N/A	34.1

Note. At-Risk (%) = proportion of students with annual GPA < 2.0. Mean GPA (Annual) and Mean GPA (HY) are computed over all annual and half-yearly records, respectively. Avg Teacher Exp = mean years of teaching experience reported on the school roster; N/A = not available for SUSC. STEM Teachers = teachers assigned to Mathematics, Science, or Technology subjects. Mean Class Size = mean number of students per classroom section. HY = half-yearly assessment; GPA = grade point average; STEM = Science, Technology, Engineering, and Mathematics.

(Rakib et al., 2026).

Outcome variable: *at_risk_label*: binary indicator, 1 if year-end GPA < 2.0 (the national passing threshold). GPA is computed by the school from subject-level grade points under the conjunctive national rule.

Half-yearly predictors (Tier P features): GPA at mid-year exam (*GPA_HY*), number of failed subjects at mid-year (*HY_Fail_Count*), minimum subject percentage across all subjects (*HY_Min_Subject_Pct*), standard deviation of subject percentages (*HY_Subject_Disparity*), within-section rank percentile at mid-year (*HY_Rank_Percentile*), section mean GPA at mid-year (*Section_HY_Avg_GPA*), proportion of classmates at risk at mid-year (*Peer_HY_AtRisk_Rate*), and section-level merit rank at mid-year (*Merit_Section_HY*). Subject-level percentage scores for Bangla, English, Mathematics, Science, Bangladesh and Global Studies/Environment (BGS/Environment), and Religion at mid-year were additionally included; General Knowledge and Skills/Other were excluded from the predictor set on account of high missingness (83.2% and 58.4% respectively). School-level contextual covariates comprised total teacher count, STEM teacher count, average teacher experience, teacher-student ratio, and class size.

The consolidated dataset from which these features were derived contains 363 columns in total: student records were expanded across both examination types (half-yearly and annual) and all assessed subjects (up to 37 subject-name variants across the nine schools), resulting in per-subject mark, grade point, and percentage columns for each combination, together with cross-examination velocity scores and section-level and school-level aggregates. The Tier-P file retains only the half-yearly-derived predictors and school-level contextual features as a 30-column subset; after removing administrative identifiers, the outcome label, and four columns excluded from the predictor set owing to near-zero variance or high missingness ($\geq 58\%$), the 19 named predictors above enter the model alongside class level, examination year, a one-hot-encoded school identifier, and three binary missing-data indicators, yielding 32 features in total.

Missing data: 162 students (3.5% of the annual cohort) had no half-yearly record, predominantly SGPHS Class 3 2024 students (n = 137) for whom half-yearly results were not published, with isolated gaps at BBS, OHS, SJIS, and VHS. These students were excluded from the prediction analysis.

Exclusion on the basis of missing half-yearly records is unlikely to be Missing Completely at Random (MCAR). The SGPHS Class 3 2024 exclusions reflect a gap in the school's published records rather than selective omission. The most consequential bias is likely conservative in direction: SGPHS Class 3 students have among the highest at-risk rates in the dataset, so their exclusion probably understates the model's real false-negative rate in deployment at that school and class level.

No imputation was applied to the binary outcome. General Knowledge (83.2% null) and Skills/Other (58.4% null) were excluded from the predictor set entirely and not imputed; these are the two features noted as excluded in the Tier-P paragraph above. Three subject-level features with partial missingness (Science, BGS/Environment, and Religion) were imputed with the school–class-level median where available and otherwise set to 0; a binary missing-indicator column was added for each of these three, yielding the three missing-data indicators included among the 32 model features.

A comparison of excluded and included students confirms the direction of this bias. Excluded students (n = 162) have a year-end at-risk rate of 44.4% [95% CI: 37.0%–52.1%] and a mean annual GPA of 2.32, compared with 21.3% [95% CI: 20.1%–22.5%] and 3.42 for the 4,470 included students. SGPHS Class 3 2024 accounts for 137 of the 162 excluded records (84.6%) at an at-risk rate of 44.5%, confirming that the model's real false-negative burden at that school and class level is higher than cross-validation estimates reflect. Importantly, the 162 excluded students are absent from the prediction model but remain in the full annual dataset: the reported at-risk rate of 22.1% counts all 4,632 students (952 at-risk in the model cohort plus 72 at-risk among those ex-

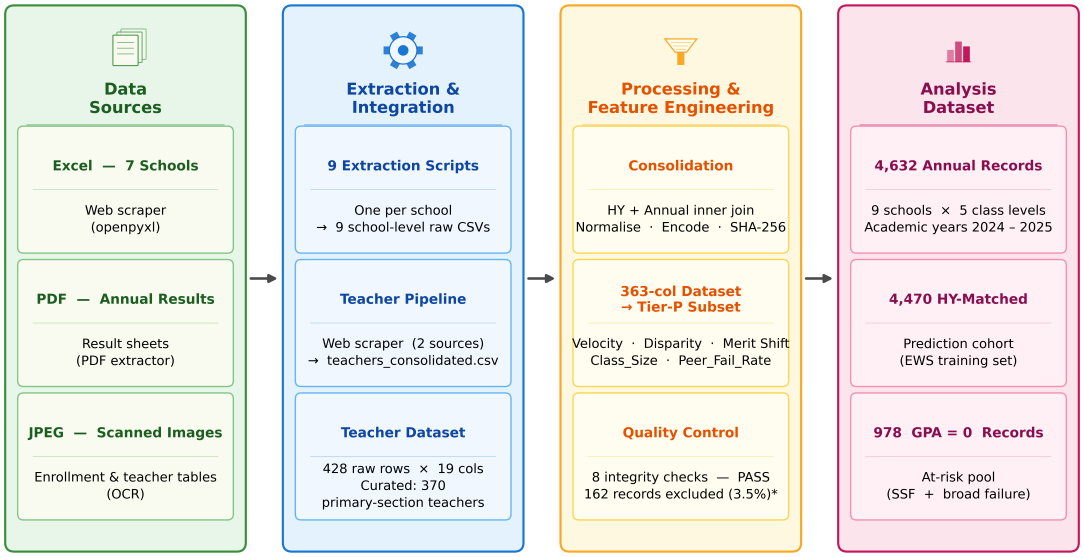


Figure 1. Data collection and curation pipeline for nine Sylhet primary schools

Note: Four grouped columns show the end-to-end pipeline from raw data collection to the analysis-ready dataset. Arrows indicate left-to-right data flow. 162 records excluded = 3.5% of annual dataset; excluded records had no half-yearly entry. The 428 teacher rows represent raw roster entries before deduplication and removal of secondary-level staff; 370 primary-section teachers remain after quality control, as reported in Table 1. HY = half-yearly; OCR = optical character recognition; SHA-256 = Secure Hash Algorithm 256-bit; SUSC = Shahjalal University School and College; AGHS = Govt. Agragami Girls' High School.

cluded = 1,024 total). The model therefore operates on a slightly lower observed prevalence (21.3%) than the full cohort. As an extreme sensitivity bound, treating all 162 excluded students as at-risk raises the full-cohort rate to (952 + 162)/4,632 = 24.1%, a 2.0 percentage-point change that does not alter any substantive claim in this paper.

Figure 2 provides a schematic overview of the three parallel analytical tracks described in Sections 3.3–3.5, each corresponding to one research question.

3.3. Early Warning Classification Model

Three classifiers were trained and evaluated. Logistic Regression (LR; L2 regularization, C = 1.0), Random Forest (Breiman, 2001; RF; 300 trees, max depth = 10, class weight balanced) and XGBoost (Chen & Guestrin, 2016; max depth $\in \{3, 5\}$, learning rate $\in \{0.05, 0.1\}$, n_estimators $\in \{100, 200\}$, optimal hyperparameters selected per fold via inner 3-fold grid search, scale_pos_weight set to majority/minority ratio (negatives ÷ positives), which upweights the minority at-risk class during training). All features were standardized (z-scored) before LR training and tree-based models received raw values.

Validation strategy. Four validation approaches were applied:

1. *Stratified 5-fold cross-validation*: It was used as the primary evaluation method, maintaining class balance across folds. Accuracy and AUC are reported as fold means with standard deviations.
2. *Group K-fold cross-validation*: The entire classroom sections were kept together during validation (n = 103 groups, each group represents one classroom section, not one school). This prevents classmate-level leakage involving peer-context features (*Section_HY_Avg_GPA*, *Peer_HY_AtRisk_Rate*) and tests whether these features provide useful predictive information beyond correlations within individual sections.
3. *Temporal validation*: Training on 2024 students only and testing on 2025 students only. This is the most deployment-relevant validation: it mirrors the scenario in which a school uses last year's cohort to train a model and applies it to the current year.
4. *Leave-one-school-out cross-validation (LOSO)*: Each school's students were held out in turn as the test set while the remaining eight schools' data were used for training. This tests cross-institution generalization by testing whether a model trained on students from eight school types predicts adequately for a ninth unseen institution. XGBoost was used for this validation given its superior temporal performance. Per-school results are reported in Section 4.2.4.

Decision thresholds were selected per validation split rather than fixed at the default 0.5. For each fold, a grid search over candidate thresholds from 0.1 to 0.9 (in 0.01 steps) was applied to the training fold; the F1-maximizing threshold was then applied to the held-out fold without modification. The F1-maximizing criterion was chosen because the dataset is class-imbalanced (21.3% positive rate): a fixed 0.5 threshold over-penalizes recall in this setting, generating many missed at-risk cases. For temporal and LOSO validation, the threshold was likewise selected on the training split and applied without modification to the test split, preventing any look-ahead into unseen data.

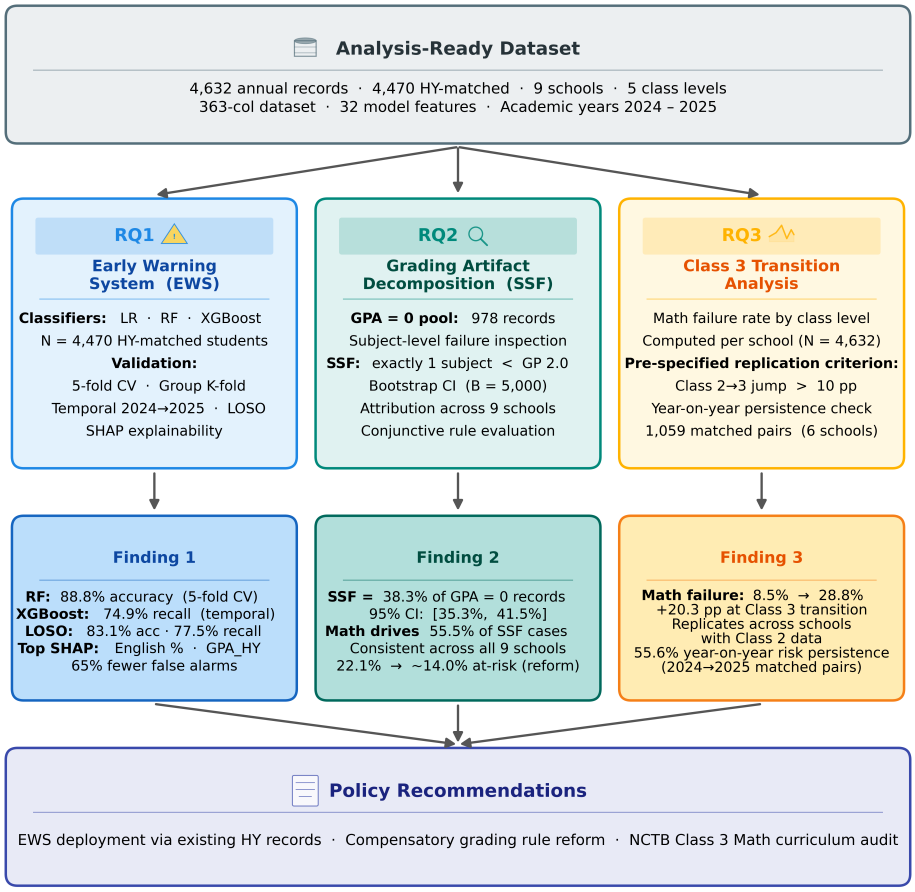


Figure 2. Methodological overview showing three parallel analysis tracks

Note: All three tracks draw from the same analysis-ready dataset (top node). RQ1, RQ2, and RQ3 labels correspond to the three research questions in Section 1. Findings nodes summarize results in Sections 4.2, 4.3, and 4.4 respectively. EWS = early warning system; SSF = single-subject failure; HY = half-yearly; CV = cross-validation; LOSO = leave-one-school-out cross-validation; LR = logistic regression; RF = random forest; CI = confidence interval (nonparametric bootstrap); GPA = grade point average; pp = percentage points.

3.4. Single-Subject Failure Decomposition

To separate single-subject failures from genuine broad failure, every GPA = 0 student record was inspected at the subject level. Throughout this paper, "grading-rule artifact" refers specifically to a measurement and reporting consequence of the conjunctive rule, not to an error in the underlying competency standard itself. A student who fails one required subject is correctly classified as failing under the current rule. The artifact lies in the aggregate GPA = 0 label, assigning the same label to this student and to a student who fails all subjects. Thus, making the two populations analytically indistinguishable at the reporting level. A record was classified as a **single-subject failure (SSF)** if: (a) the student received GPA = 0 overall and (b) exactly one canonical subject had a grade point of 0 (the fail grade under the national scheme). The dominant failing subject in each SSF record was recorded for subject-attribution analysis.

This decomposition is conservative: it uses the strictest possible definition of SSF (exactly one canonical subject with a grade point of 0).

3.5. Class 3 Transition Analysis

Mathematics failure rates (grade point < 2.0) were computed by class level for each school separately and for the pooled dataset. Replication across schools was assessed by examining whether the Class 2-to-Class 3 failure rate jump exceeded 10 percentage points in each individual school, independent of whether any other schools showed the same pattern. The replication criterion (present in all nine schools) was pre-specified before computing school-level figures. To formally test the Class 3 failure rate increase while controlling for between-school and between-year differences, a logistic regression model was fitted with Mathematics failure (subject grade point < 2.0 at year-end) as the binary outcome and class level (Class 1 as reference category), school identity (eight indicator variables), and examination year as predictors.

4. Results

4.1. At-Risk Prevalence and Variation Across Schools

1,024 students (22.1%) among the 4,632 annual student records finished the academic year at risk (GPA < 2.0). This overall figure conceals a 3.5-fold range across institutions,

Table 2. At-risk prevalence by school, sorted by descending at-risk rate

Rank	School	N	At-Risk (n)	At-Risk (%)	Mean Annual GPA	Median Annual GPA	Mean HY GPA	GPA Trajectory (HY→FIN)
1	UDS	603	247	41.0	2.41	3.17	2.73	−0.325
2	BTRI	232	87	37.5	2.58	3.00	2.03	+0.557
3	SUSC	168	44	26.2	3.27	4.29	2.70	+0.570
4	BBS	737	171	23.2	3.32	4.08	2.39	+0.946
5	SGPHS	827	188	22.7	3.20	3.83	2.55	+0.828
6	VHS	476	98	20.6	3.64	4.70	3.17	+0.481
7	OHS	122	15	12.3	3.59	4.14	2.32	+1.271
8	SJIS	939	113	12.0	4.05	4.75	3.38	+0.674
9	AGHS	528	61	11.6	3.80	4.38	2.79	+1.010
—	**Pooled**	4632	1024	22.1	3.38	—	2.79	—

Note: At-risk status is defined as annual GPA < 2.0. Schools are ranked 1 (highest at-risk rate) to 9 (lowest). GPA Trajectory (HY→FIN) is the mean within-student difference between annual and half-yearly GPA; a positive value indicates students improved from mid-year to year-end on average; UDS is the only school with a negative trajectory (−0.325). Pooled row weights mean GPA by school enrollment; median and trajectory are not computed for the pooled row. HY = half-yearly assessment; GPA = grade point average.

Table 3. Model performance — stratified 5-fold cross-validation and group K-fold cross-validation

Model	CV Acc. (%)	CV Recall (%)	CV Prec. (%)	CV F1	CV ROC-AUC	CV PR-AUC	GKF Acc. (%)	GKF Recall (%)
Naive HY Rule	79.6	83.4	51.3	63.5	81.0	46.3	—	—
Logistic Regression	84.2±2.6	80.9±9.4	60.6±7.3	68.6±2.4	92.2±0.9	79.0±1.4	83.4±3.3	84.2±4.4
Random Forest	88.8±0.8	71.3±5.2	75.3±3.6	73.1±2.2	92.8±1.3	80.9±1.7	88.2±1.0	63.5±6.8
XGBoost	86.7±1.4	81.2±1.7	65.2±3.5	72.3±2.1	92.6±1.0	80.3±0.7	85.8±2.1	79.1±5.5

Note: CV metrics are means ± standard deviations across five folds (n = 4,470 matched half-yearly / annual records; 952 at-risk, 21.3%). GKF = Group K-fold cross-validation, in which entire classroom sections are held out together to guard against within-section correlation inflating CV estimates; GKF accuracy and recall are reported for the leakage check. The Naive HY Rule flags any student with half-yearly GPA < 2.0 as at-risk; it has no variance across folds and was not re-evaluated under GKF. Recall, Precision, F1, and ROC-AUC are reported for the at-risk class (positive class = GPA < 2.0 at year-end). PR-AUC = area under the precision-recall curve. CV = stratified 5-fold cross-validation; HY = half-yearly assessment; GPA = grade point average.

from 11.6% at AGHS to 41.0% at UDS. The variation appear to follow a pattern. Schools with less experienced teaching staff and larger class sizes tend toward higher at-risk rates, though this association is ecological and subject to confounding (Asadullah & Chaudhury, 2015; Ahmmed et al., 2022).

At-risk rates also vary sharply by grade level. Classes 1 and 2 show relatively low at-risk proportions (15.6% and 14.5%, respectively). While Class 3 marks a sharp system-wide increase to 29.2% more than double the Class 2 rate. This jump is the most distinctive feature of the grade-level distribution and is investigated in Section 4.4.

4.2. Early Warning Model Performance

4.2.1. Cross-validation results

Of the 4,470 students with matched half-yearly and annual records, 952 (21.3%) were at risk by year-end. All three classifiers substantially outperform the naive baseline (flagging as at-risk any student whose half-yearly GPA falls below 2.0), which achieves 79.6% accuracy but generates approximately 755 false alarms (safe students incorrectly flagged) per prediction run.

Random Forest achieves the highest accuracy at 88.8% ± 0.8% (AUC = 0.928 ± 0.013), meeting the pre-specified grant target of $\geq 85\%$. The gain over the naive rule is not primarily in recall (the naive rule's recall of 83.4% exceeds RF's 71.3%) but in precision: per 100 at-risk students correctly identified, the naive rule generates approximately 95 unnecessary interventions while Random Forest cuts this to roughly 33, a 65% reduction that frees substantial capacity for cases that actually need attention.

4.2.2. Leakage guard — Group K-fold

To verify that peer-context features (*Section_HY_Avg_GPA*, *Peer_HY_AtRisk_Rate*) do not inflate CV scores through within-section correlation, Group K-fold cross-validation grouped entire classroom sections. The accuracy drop is 0.6 percentage points for Random Forest (88.8% → 88.2%) and at most 0.9 points for any model. Peer features add genuine predictive signal; they do not introduce artifactual leakage.

4.2.3. Temporal validation (deployment-relevant)

Temporal validation reveals an important asymmetry between models. Random Forest's recall falls from 71.3% (CV) to 55.8% when tested on a new cohort-year, resulting in a 15.5 percentage-point drop, reflecting RF's tendency to memorize section-specific within-year patterns. XGBoost sustains 74.9% recall across the cohort-year boundary, only 6.3 points below its CV figure. For schools using records from previous years to predict risk in the current cohort, XGBoost is the recommended model when SHAP-based explanations are needed to direct targeted support (Section 4.2.5). Schools prioritizing maximum case detection over interpretability should consider Logistic Regression. It achieves the highest temporal recall (81.0%) with the smallest cohort-year performance drop (+0.1 pp). Logistic Regression's consistent performance suggests that the core linear relationship between mid-year performance and final risk is stable across cohort years.

4.2.4. Cross-school generalization (LOSO)

Leave-one-school-out cross-validation tests whether an XGBoost model trained on eight schools can predict the at-

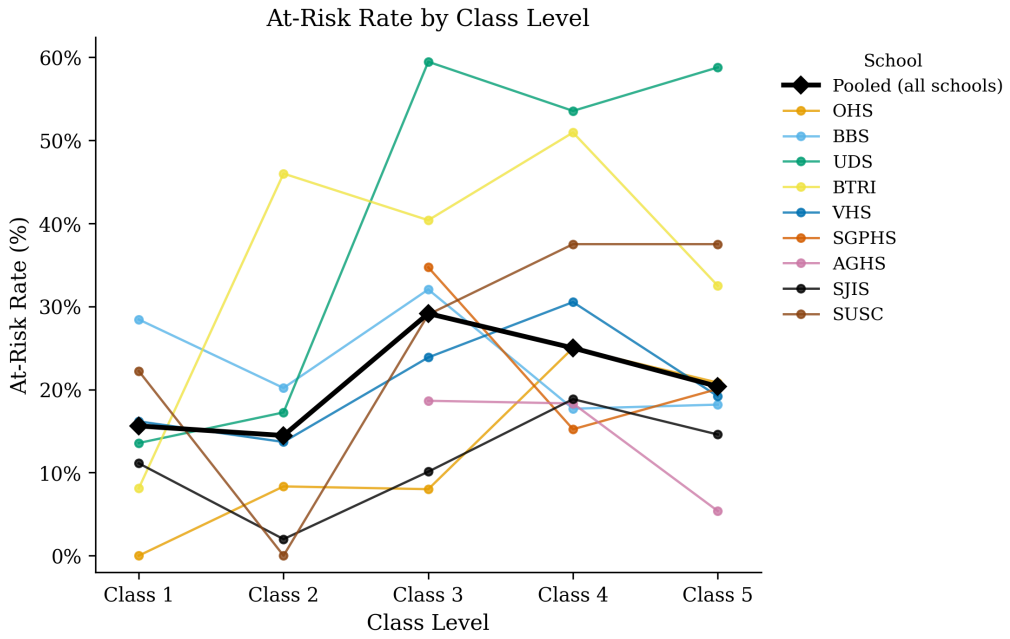


Figure 3. At-risk rate by class level across nine Sylhet primary schools

Note: Each thin colored line represents one of the nine schools in the study; the bold black line shows the pooled rate across all schools (n = 4,632 annual records). At-risk status is defined as annual GPA $<$ 2.0. Class 1 and Class 2 show relatively stable and low pooled at-risk rates (15.6% and 14.5%, respectively). Class 3 marks a sharp system-wide inflection to 29.2% — more than double the Class 2 pooled rate — followed by modest variation at Classes 4 and 5. School-level trajectories show wide dispersion: UDS climbs to approximately 60% at-risk by Class 5 while AGHS remains below 20% throughout. Despite this dispersion, the Class 2→3 upward inflection is present in nearly all individual school lines, indicating the pattern is not driven by a single institution. Schools AGHS and SGPHS contribute data from Class 3 onward only; their lines begin at Class 3. HY = half-yearly assessment; GPA = grade point average.

Table 4. Temporal validation results — model trained on 2024 cohort, tested on 2025 cohort

Model	CV Acc. (%)	Temp. Acc. (%)	Δ Acc. (pp)	CV Recall (%)	Temp. Recall (%)	Δ Recall (pp)	CV F1	Temp. F1	Temp. ROC-AUC
Logistic Regression	84.2	84.1	0.0	80.9	81.0	+0.1	68.6	67.6	91.1
Random Forest	88.8	87.5	−1.3	71.3	55.8	−15.5	73.1	64.6	89.9
XGBoost	86.7	86.9	+0.2	81.2	74.9	−6.3	72.3	70.0	90.2

Note: CV metrics are stratified 5-fold cross-validation means (within-year, 2024 training data). Temporal metrics are single-split out-of-sample estimates (train: 2024 annual records; test: 2025 half-yearly + annual matched records). Δ = temporal minus CV value; a negative value indicates generalization loss when moving to a new cohort year. Recall is reported for the at-risk class (positive class = annual GPA $<$ 2.0). XGBoost is recommended as the primary deployment model because its SHAP feature importance (Section 4.2.5) shows which factors contribute to at-risk prediction. Under temporal validation, Logistic Regression achieves the highest at-risk recall (81.0%) with the smallest cohort-boundary drop (+0.1 pp), making it the preferred choice when maximizing case detection is the only priority and explanatory output is not required. pp = percentage points; CV = cross-validation; GPA = grade point average.

risk outcomes of the ninth unseen school's . Table 5 reports per-school results.

Across nine held-out schools, mean LOSO accuracy is 83.1% (range: 71.8%–91.6%) and mean at-risk recall is 77.5% (range: 53.1%–94.6%). Generalization quality varies substantially. BBS returns the lowest accuracy (71.8%), reflecting that its student population, characterized by high teacher experience and a moderate at-risk rate, is underrepresented in the remaining eight schools. SJIS returns the lowest recall (53.1%), attributable to its distinctively low at-risk rate (12.0%) and Islamic school context, which the model trained on a more heterogeneous corpus does not fully capture. At the other end, SGPHS and OHS achieve recall above 89%, indicating that their at-risk patterns are well-represented in the training corpus.

UDS has the highest at-risk rate in the dataset (41.0%), achieves LOSO accuracy of 83.6% and recall of 71.7%. The results place it in the mid-range on both metrics. Although UDS had a particularly high proportion of at-risk students, this did not prevent the model from generalization. A model trained on the other eight schools was still able to identify at-risk patterns among UDS students. In contrast, generalization was more difficult for SJIS and BBS because their distinctive institutional characteristics and demographic composition were not well represented in the other schools, respectively.

4.2.5. What drives the prediction

Students who recover between mid-year and year-end fail an average of 1.8 subjects at mid-year; those who remain at risk fail an average of 3.6 subjects, a two-subject gap that is more consistent across schools and cohort years than any single-subject mark threshold. At the same mid-year GPA, a student failing three subjects is at substantially higher risk than a student failing one with very low marks: failure breadth is the clearest differentiating signal beyond

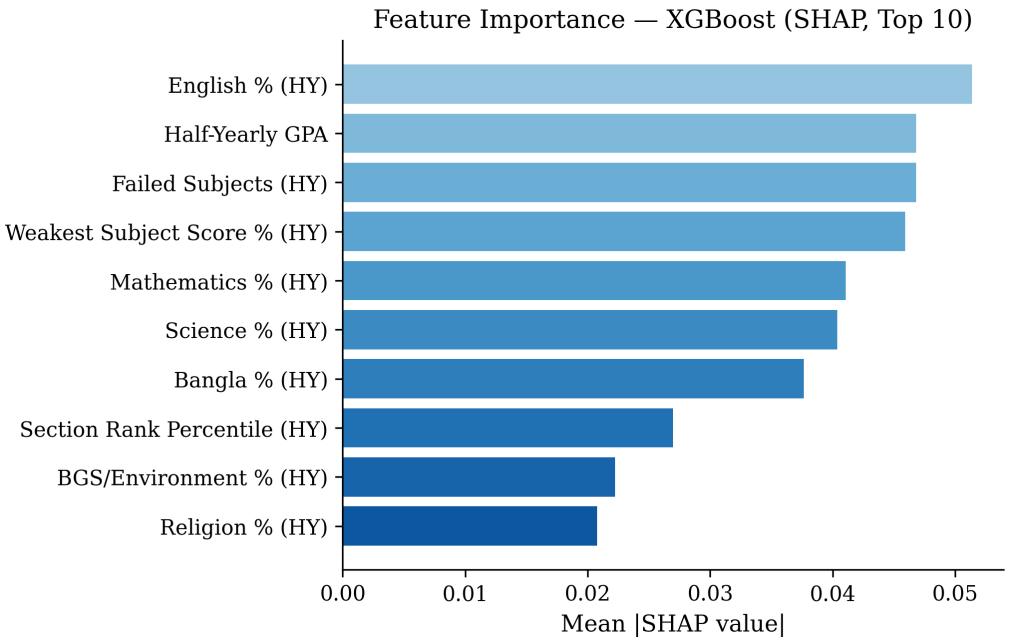


Figure 4. XGBoost SHAP feature importance — top 10 predictors of year-end at-risk status

Note: Bars represent mean absolute SHAP values computed over the full training set using the tuned XGBoost classifier. Higher values indicate greater average contribution to the model's output irrespective of direction. Features are ordered from highest to lowest importance (top to bottom). All features are derived from half-yearly (HY) examination records collected in July. English % (HY) leads marginally (mean |SHAP| = 0.051), followed by Half-Yearly GPA (0.047), Failed Subjects (HY, 0.047), and Weakest Subject Score % (HY, 0.046); these four have similar contributions and together account for the largest share of predictive signal. Subject-specific percentage scores (Mathematics, Science, Bangla, BGS/Environment, Religion) and Section Rank Percentile contribute smaller incremental signal. No teacher-level or school-level feature appears in the top 10. SHAP = SHapley Additive exPlanations; HY = half-yearly assessment; BGS = Bangladesh and Global Studies.

Table 5. Leave-one-school-out (LOSO) cross-validation results — XGBoost, per held-out school

Held-Out School	N Test	Acc. (%)	At-Risk Recall (%)	F1	ROC-AUC
BBS	726	71.8	94.6	60.5	91.8
SGPHS	681	79.6	89.5	61.5	91.2
BTRI	232	81.0	82.8	76.6	89.7
OHS	121	82.6	93.3	57.1	94.5
UDS	603	83.6	71.7	78.2	87.0
AGHS	528	83.9	88.5	56.0	92.5
SUSC	168	84.5	56.8	65.8	86.1
VHS	475	89.0	67.0	71.4	92.6
SJIS	936	91.6	53.1	59.9	91.3
Mean	—	83.1	77.5	65.2	90.8

Note. In each LOSO fold, the held-out school's students form the test set and the remaining eight schools' students form the training set. Schools are ordered from lowest to highest accuracy (hardest to easiest to generalize to). XGBoost was used for all LOSO folds given its superior temporal validation performance. At-Risk Recall = proportion of year-end at-risk students (annual GPA < 2.0) correctly identified; F1 is the harmonic mean of precision and recall for the at-risk class. The Mean row is the unweighted mean across nine held-out schools. LOSO = leave-one-school-out; GPA = grade point average; ROC-AUC = receiver operating characteristic area under the curve.

overall GPA.

English % at mid-year marginally leads the XGBoost feature importance rankings (mean |SHAP| = 0.051), followed by Half-Yearly GPA (0.047), Failed Subjects at mid-year (0.047), and Weakest Subject Score % (0.046) — four predictors so tightly clustered as to be effectively co-equal (Figure 4).

The remaining top-10 features are subject-specific percentage scores (Mathematics, Science, Bangla, BGS/Environment, Religion) and Section Rank Percentile, each contributing smaller incremental signal beyond the top four. No teacher-level or school-level feature appears in the top 10, consistent with the ablation finding that teacher covariates add negligible signal beyond student-level half-yearly performance (Section 5.5).

4.3. Single-Subject Failures Under the Grading Rule

Of the 978 students with GPA = 0, decomposition at the subject level shows that 375 (38.3%, 95% CI: 35.3%–41.5%) are single-subject failures: students who received a fail grade in exactly one canonical subject while receiving passing grades across all others. Under the current conjunctive grading rule, these students receive the same GPA = 0 label as students who fail every subject.

The subject-attribution of SSF cases is highly concentrated. Mathematics accounts for 55.5% of all SSF cases (208 of 375), followed by English (14.9%), BGS & Environment (8.8%), Religion (7.5%), Bangla (4.8%), Science (3.7%), and Skills & Other (3.7%), with General Knowledge accounting for the remaining 1.1%. Three findings make the Mathematics concentration noteworthy. First, it is not proportional to Mathematics' share of the curriculum; it is disproportionately high. Second, it is consistent across school types: Mathematics is the dominant SSF driver in all nine schools, spanning government, institutional, private, Islamic, and girls-only institutions. Third, it is grade-level specific: as shown in Section 4.4, the Mathematics SSF concentration

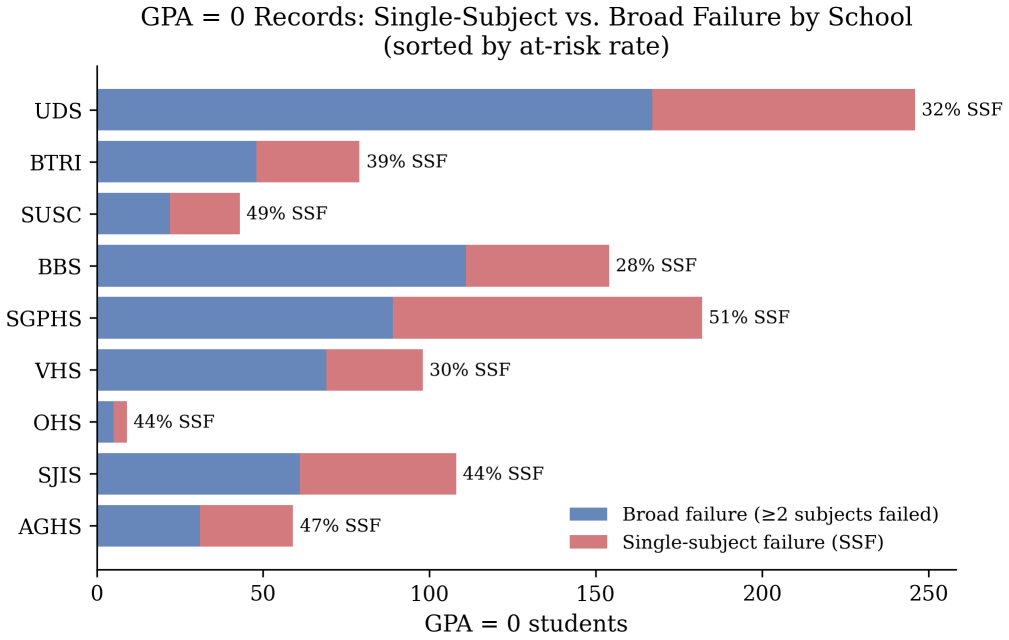


Figure 5. Decomposition of GPA = 0 records into single-subject failure and broad failure, by school

Note: Schools are ordered from top to bottom by descending at-risk rate (UDS = 41.0%, AGHS = 11.6%). Each bar represents the total count of GPA = 0 students at that school; bar segments distinguish broad failure (blue; ≥2 canonical subjects failed) from single-subject failure (SSF; pink; exactly 1 canonical subject failed). The SSF percentage annotated at the right of each bar is the school-level SSF share of GPA = 0 students. Across all nine schools, 375 of 978 GPA = 0 students (38.3%, 95% CI: 35.3%–41.5%) are SSF cases. School-level SSF rates range from 28% (BBS) to 51% (SGPHS), indicating that the grading-rule artifact is not concentrated in any single institution. SGPHS and AGHS contributed data for Classes 3–5 only; their GPA = 0 counts and SSF proportions therefore reflect a three-class range rather than the full five-class span of the remaining schools. SSF = single-subject failure; GPA = grade point average; CI = confidence interval (nonparametric bootstrap, B = 5,000).

Table 6. Single-subject failure (SSF) attribution by canonical subject group across nine schools

Subject	SSF (n)	% Total SSF	% All GPA=0	Schools (n)
Mathematics	208	55.5	21.3	9
English	56	14.9	5.7	9
BGS & Environment	33	8.8	3.4	8
Religion	28	7.5	2.9	6
Bangla	18	4.8	1.8	4
Science	14	3.7	1.4	5
Skills & Other	14	3.7	1.4	3
General Knowledge	4	1.1	0.4	1
Total	375	100.0	38.3	—

Note. SSF = single-subject failure; defined as a GPA = 0 record where exactly one canonical subject group received a grade point of 0 (fail grade) while all remaining subjects received passing grades. "% Total SSF" is computed over all SSF cases (n = 375); "% All GPA = 0" is computed over all GPA = 0 records (n = 978). "Schools (n)" = number of schools in which this subject appears as the sole failing subject in at least one SSF case. BGS = Bangladesh and Global Studies; GPA = grade point average.

is driven almost entirely by Classes 3–5, not by Classes 1–2.

A conservative estimate of the policy implication: if a compensatory grading rule were applied that computed GPA as the weighted average of subject performance rather than assigning 0 for any single failure, approximately up to 38% of current GPA = 0 records could be affected (reclassified to passing status). No student's actual academic performance would change; only the labeling method would change. This would reduce the reported whole-dataset at-risk rate from 22.1% to approximately 14.0%.

4.4. The Class 3 Curriculum Breakpoint

The most consistent structural finding in this dataset is the Mathematics failure rate discontinuity at Class 3. In the pooled dataset, the Mathematics subject failure rate (grade point < 2.0 at year-end) is 8.5% at Class 2 [95% CI: 6.3%–10.8%] and 28.8% at Class 3 [95% CI: 26.1%–31.6%], a 20 percentage-point increase in a single grade transition. The rate remains elevated at Class 4 (30.8%) and Class 5 (30.6%), with no meaningful recovery after the Class 3 transition. Controlling for school and year fixed effects in a logistic regression model, Class 3 students have 3.3 times the odds of Mathematics failure compared with Class 1 students (OR = 3.33, 95% CI: 2.49–4.47, $p < .001$), and approximately 5 times the odds compared with Class 2 students (implied OR ≈ 5.1). The Class 3 effect is present and statistically significant after accounting for compositional differences across the nine schools and the two cohort years.

This pattern is not driven by a single school. It replicates across the sampled schools with available Class 2 data, including schools with very different at-risk rates, class sizes, teacher experience levels, and institutional types.

The jump is abrupt: Classes 1 and 2 show relatively low failure rates (12.3% and 8.5% respectively), and the increase to 28.8% occurs at a single grade boundary, not as a gradual trend. It is also symmetric across institutional types: the pattern appears in the institutional OHS (which has the most experienced teachers and the lowest overall at-risk rate), in private UDS (the highest at-risk school), in the Islamic school SJIS, and in the girls-only government school

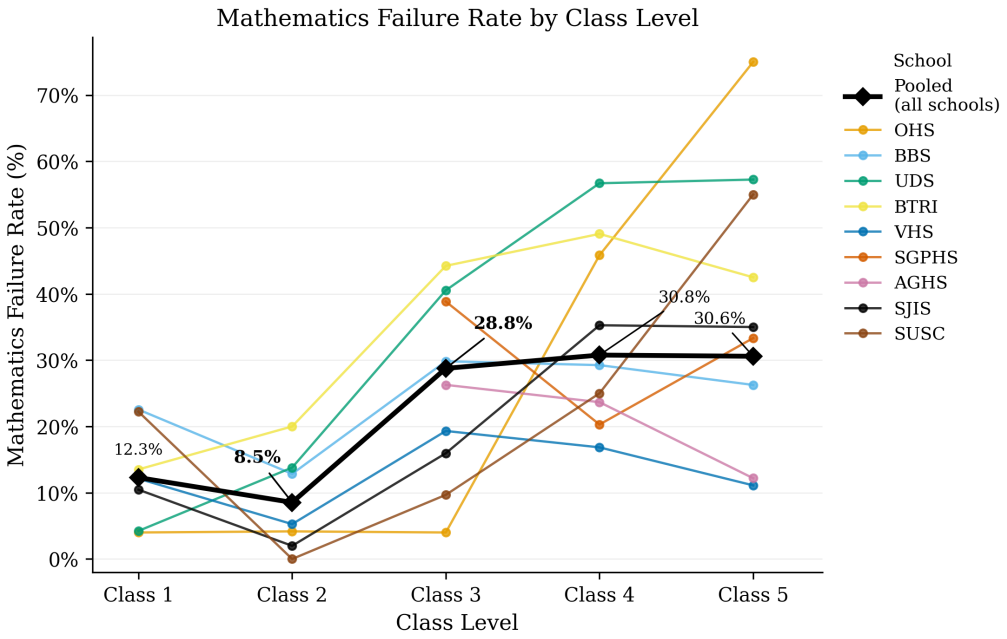


Figure 6. Mathematics subject failure rate by class level across nine Sylhet primary schools

Note: Mathematics failure is defined as a subject-level grade point < 2.0 at the year-end assessment. Each thin colored line represents one school; the bold black line shows the pooled failure rate across all nine schools. Pooled rates are annotated for all five class levels: Class 1 (12.3%), Class 2 (8.5% [95% CI: 6.3%–10.8%]), Class 3 (28.8% [95% CI: 26.1%–31.6%]), Class 4 (30.8%), and Class 5 (30.6%); Class 2 and Class 3 annotations are bold to emphasize the key transition. The 20 percentage-point jump from Class 2 (8.5%) to Class 3 (28.8%) in the pooled line replicates across the individual school lines with Class 2 data, including schools with highly different overall at-risk rates, teacher experience levels, and institutional types (government, institutional, private, Islamic, girls-only). The elevated failure rate at Class 3 does not recover at Class 4 or Class 5. Schools AGHS and SGPHS contribute data from Class 3 onward only; their lines begin at Class 3.

AGHS, among others. And it persists: the elevated failure rate does not recover at Class 4 or 5; students who fall behind at Class 3 largely remain behind through the end of primary school.

The cross-institutional replication effectively rules out school-level explanations (teacher turnover, class size, institutional management) as primary causes. What nine heterogeneous schools share in common is the national Mathematics curriculum; the specific content demands of the Bangladesh National Curriculum and Textbook Board (NCTB) Class 3 syllabus and their policy implications are discussed in Section 5.3.

Year-over-year risk persistence. Among cohort-matched students in schools with two years of data (2024–2025), 55.6% of students at risk in 2024 remained at risk in 2025 (1,059 matched pairs). Pairs were identified by school and roll number from source records (roll numbers are not included in the public dataset release), with class level advanced by one year across consecutive annual cohorts; six schools contributed (BBS, BTRI, SGPHS, SJIS, UDS, and VHS), as OHS and SUSC had only one year of records, and the AGHS Class 4 2024 cohort had only half-yearly results published. The within-year persistence rate (51.3% of mid-year at-risk students remaining at-risk by year-end) is not directly comparable — it covers a different population over a shorter time window — but the cross-year figure [95% CI: 48.9%–62.0%] independently indicates that more than half of students recorded as at-risk in one academic year remain at risk the following year without intervention.

5. Discussion

5.1. RQ1: The Case for a Within-School Early Warning System

Half-yearly examination records support reliable identification of students who will remain at risk by year-end. This conclusion holds across four independent validation designs, where each examines a different aspect of the early warning system. Stratified cross-validation evaluates the model's ability to identify at-risk students within the same year. Random Forest achieves 88.8% accuracy (AUC = 0.928) and reduces false-alarm interventions by 65% compared with naive mid-year flagging. Group K-fold validation assures that the peer-context features (*Section_HY_Avg_GPA*, *Peer_HY_AtRisk_Rate*) contribute genuine signal rather than within-section correlation artifacts. The RF accuracy has dropped by only 0.6 percentage points under this stricter grouping. Temporal validation examines whether the model can maintain its performance across cohort years. When trained on 2024 records and tested on 2025 students, XGBoost achieves 74.9% at-risk recall. Only a fraction (one in five) of published early warning studies report this type of cross-year evaluation (Albreiki et al., 2021). Leave-one-school-out cross-validation extends the test to a cross-institutional setting. A model trained on eight schools generalizes to a ninth unseen institution at a mean accuracy of 83.1% and mean at-risk recall of 77.5%. The reliability finding is not an artifact of one evaluation design. Rather, it holds across time, sections and institutions.

The four leading SHAP magnitudes indicate failure breadth rather than failure depth in any single subject (Sec-

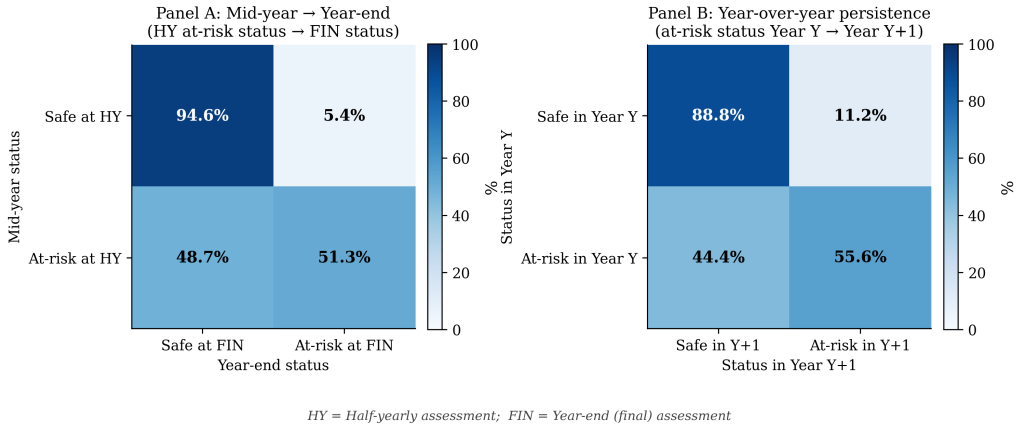


Figure 7. Risk status transition matrices: mid-year to year-end (Panel A) and year-over-year persistence (Panel B)

Note: Cell values are row percentages (i.e., conditional on the row status, the percentage transitioning to each column status). Color intensity reflects cell magnitude on a 0–100 scale (Blues palette); white text is used for cells exceeding 60%. Panel A (Mid-year → Year-end): of students classified as safe at the half-yearly assessment, 94.6% remained safe at year-end and 5.4% became at-risk — a low false-negative rate indicating mid-year safe status is highly stable. Of students at-risk at mid-year, 51.3% remained at-risk by year-end (within-year persistence) while 48.7% recovered. Panel B (Year-over-year persistence, 2024→2025): of students at-risk in Year Y (2024), 55.6% were still at-risk in Year Y+1 (2025), based on 1,059 matched student pairs across schools with two years of data. The cross-year persistence rate (55.6% [95% CI: 48.9%–62.0%]) indicates that more than half of at-risk students remain at risk the following year; this figure and the within-year rate (51.3%) involve different populations and time spans and should not be compared inferentially. HY = half-yearly assessment; FIN = year-end (final) assessment.

tion 4.2.5). This provides a more consistent distinction between students who recover and those who remain at risk. English has a slightly higher SHAP value (0.051, compared with 0.047 for the next three predictors). This small difference is consistent with its role as a gating subject in the national curriculum. For students with equivalent overall GPA, a lower mid-year performance in English may indicate somewhat greater chronic risk than a comparable performance in most other subjects. However, the difference is too small to support a strong claim that English is the dominant predictor.

Two constraints bound the deployment picture. Since the model was trained and tested within a single city, its ability to generalize to schools in Dhaka, Chittagong, or rural areas remains untested. The LOSO pattern sharpens this limit: generalization quality is determined by institutional distinctiveness, not by at-risk prevalence. BBS (71.8% accuracy) and SJIS (53.1% at-risk recall) are harder for a cross-trained model to predict than UDS (83.6% accuracy, 71.7% recall) despite UDS having the dataset's highest at-risk rate, suggesting that schools with character underrepresented in the training corpus drive generalization difficulty. Cross-district validation is the necessary next step before any national deployment recommendation. On temporal scope, XGBoost's 6.3-percentage-point recall drop from within-year cross-validation to the cross-cohort test is the quantified cost of training on prior-year records; it is small enough to support annual reuse of a trained model without rebuilding, but it should be monitored as the school's student population changes.

Prediction is not intervention. Identifying a student as at-risk at the half-yearly examination does not specify what support they need; the feature profile can guide that decision. Students failing one or two subjects at mid-year are candidates for targeted subject-specific support. Those failing four or more subjects need broader academic support before the increased demands of Class 3, which will further compound their difficulties.

5.2. RQ2: The Grading Rule as a Policy-Correctable Measurement Problem

The early warning model identifies students who may need intervention but interpreting these flags requires understanding what a year-end GPA = 0 represents. Among the 978 students recorded as GPA = 0, 375 (38.3%; 95% CI: 35.3%–41.5%) failed exactly one canonical subject while passing all others. This classification results from the conjunctive grading rule, which assigns GPA = 0 regardless of how many subjects a student actually fails. More than one in three recorded failures in this dataset is not a reflection of broad academic underperformance. This reporting consequence is not limited to particular institutions. School-level SSF rates range from 28% at BBS to 51% at SGPHS and also present in all nine schools in the sample dataset irrespective of institutional type. A reporting consequence of this magnitude and consistency warrants systematic analysis rather than treatment as a marginal correction.

The subject attribution of SSF cases is highly concentrated and the concentration is analytically informative beyond the aggregate figure. Mathematics is the only failing subject in 55.5% of all SSF cases (208 of 375). It also appears as the dominant SSF driver in all nine schools disproportionate to its share of the curriculum. This pattern also varies across grade levels. Mathematics-related SSF cases occur predominantly in Classes 3–5, whereas Classes 1–2 have lower Mathematics failure rates and fewer SSF cases. These findings suggest that the Mathematics-related SSF and the Class 3 Mathematics breakpoint are closely connected. A Mathematics-specific failure at the Class 3 transition propagates directly into a GPA = 0 label under the conjunctive rule, inflating the apparent at-risk rate and obscuring the curricular origin of the failure. Treating them as separate

problems would understate the paper's central argument.

During the Class 3 transition, when a student fails in Mathematics the conjunctive grading rule assigns a GPA = 0 regardless of performance in other subjects. This adds to the reported at-risk rate and makes it harder to identify the origin of the failure. Treating them as separate problems would understate the paper's central argument.

The rationale behind the conjunctive grading rule is defensible. Requiring students to achieve minimum competence in every core subject prevents an overall GPA from hiding a critical weakness in any individual subject through cross-subject averaging. However, the way the rule classifies students creates a reporting problem. A student who fails only Mathematics receives the same GPA = 0 as a student who fails every subject. As a result, their records become indifferent in combined measures of school failure rates. This creates three practical problems. First, it misdirects intervention resources towards comprehensive remediation for students who actually need subject-specific support. Second, it distorts comparisons of at-risk rates among schools with greater exposure to Class 3 Mathematics failure. Third, the same classification can trigger grade-retention decisions, though meta-analytic evidence showing no average long-term academic benefit from grade retention (Goos et al., 2021). In all three cases, the problem arises from using the same label for two qualitatively different types of academic failure.

The reporting problem can be addressed without changing the minimum competency requirement. Applying a compensatory grading calculation to aggregate reporting with the existing conjunctive rule would reclassify approximately 38% of current GPA = 0 records. This would reduce the reported at-risk rate from 22.1% to 14.0% without changing any student's subject-level assessment or removing the minimum-competency requirement. The conjunctive rule can still be used to track individual students' competency and identify subject-specific difficulties. However, the compensatory calculation would provide an alternative measure for aggregate reporting. Separating these two functions does not lower academic standards. It makes the reported statistics more representative of students' actual performance. The information required for proposed reporting is already available in every school's half-yearly and annual records.

5.3. RQ3: The Class 3 Breakpoint: What Schools Cannot Fix Alone

Mathematics accounts for 55.5% of single-subject failures. This raises an important question: "Why mathematics, and why does this pattern emerge at Class 3?" Previous curriculum research identifies Mathematics as the most frequently associated subject with changes in academic attainment during the transitions from primary to upper-primary grades. These changes have been linked to a shift from procedural to conceptual learning demands (Kaur et al., 2022). However, whether this pattern occurs across different types of institutions in a Bangladeshi urban district or it is associated with the national curriculum rather than individual school characteristics has not previously been examined at the multi-institutional level. The Mathematics failure rate in the sampled schools increases from 8.5% to 28.8% at Class 3 (note that SGPHS and AGHS provided data exclusively for Classes 3–5). Across the sampled schools, there are clear differences in teacher experience, class size, funding, and management structure. Despite these differences, the national curriculum remains the common factor. This suggests that school-level resource allocation alone may not resolve a problem that appears to be linked to the structure of the curriculum.

The NCTB Class 3 Mathematics curriculum introduces abstraction compared with the content covered in Classes 1–2. Three system-level responses could address this problem. They differ in implementation cost and in how directly they address the source of the difficulty. A *bridging curriculum* would introduce a structured sequence of readiness activities at the end of Class 2 to help students prepare for the conceptual demands of Class 3. This approach addresses the problem before students make the transition and is the structural response most consistently recommended in the curriculum-transition literature (Kaur et al., 2022). However, it would require curriculum revision and involve the highest implementation cost. A *diagnostic assessment* at the beginning of Class 3 could identify the specific concepts that students struggle with. This would allow additional support to focus on those concepts rather than covering the entire Class 3 syllabus. The half-yearly records examined in this study already provide some of this information by identifying students who are still failing Mathematics when they enter the next half. A *subject-specific in-service training* programme for Class 3 teachers would be the lowest-cost option and would not require curriculum revision. However, its effectiveness would depend on whether gaps in teacher knowledge or differences in teaching approaches are the main constraint at each school.

The present study provides evidence from within a single city, but the findings still need to be examined cross-districts. The failure breadth finding from RQ1 offers further insight into RQ3. Across all nine schools, the number of subjects a student is failing at mid-year consistently predicts chronic risk, though there are differences in teacher experience or class size. This indicates that failure breadth provides an individual-level signal that is not limited to individual schools. Both signals, Mathematics failure at the Class 3 transition and the number of subjects a student is failing at mid-year, are available in the half-yearly records that schools already collect. Using these records for system-level monitoring requires treating them as a shared source of information rather than keeping them separate. Section 5.4 examines whether differences in resources across the nine schools can explain any part of the Class 3 breakpoint.

5.4. Resource Disparities and Their Limits as an Explanation

The curriculum attribution rests on ruling out school-level explanations. The nine schools span a 4.0-fold range in average teacher experience (4.2 years at UDS to 17.0 years at OHS) and a ∼3-fold range in mean class size (24 students per section at OHS and BTRI to 71 at SGPHS). These gaps are large enough to be policy-relevant but do not straightforwardly explain at-risk rates. AGHS, a government girls' school with large class sizes (mean 67 stu-

dents per section) and experienced teachers (13.6 years), has one of the lowest at-risk rates (11.6%). BBS has experienced teachers (16.5 years) and modest class sizes (37 students) yet an at-risk rate of 23.2%. The ecological association between teacher experience and at-risk rate is visible at the extremes (UDS: 4.2 years, 41.0% at-risk; OHS: 17.0 years, 12.3%) but does not hold monotonically across all nine schools.

This heterogeneity is a caution against simplistic resource-based explanations. Teacher experience and class size are the most observable and oft-cited school-level inputs; yet as constants shared by every student within a school, they cannot explain within-school variation across class levels, and their cross-school associations are confounded by selection (more experienced teachers may sort to schools with easier-to-teach populations). The within-school finding, that Class 3 failure rates spike in all nine schools despite their very different resource profiles, is more informative than any cross-school teacher experience correlation.

5.5. Limitations

Four limitations should be considered when interpreting these findings. First, the sample includes nine schools from a single city. Sylhet's socioeconomic profile, diaspora remittance flows and urban school density may differ in other parts of Bangladesh. The findings should be treated as a pilot study. Replication across other districts is required before drawing conclusions at the national level.

Second, the study uses each school's internal examination records rather than standardized assessments. Marks obtained in the half-yearly examination at SJIS are not directly comparable to the same grade level at UDS because the examinations differ in content, marking standards and grade distributions. Therefore, the early warning models predict within-school risk rather than absolute learning levels. Similarly, the subject-level failure rates reported in the grading-rule analysis reflect each school's internal standards rather than a uniform competency benchmark. Although the national letter-grade thresholds used to convert percentage scores into GPA are standardized, the school-specific examinations that produce those scores are not. For example, a mark of 65% on a more demanding school's half-yearly examination may indicate greater subject mastery than the same mark on a less demanding one. However, both marks enter the model in the same way. As a result, the model implicitly adapts to each school's mark distribution rather than an external competency scale. This calibration remains consistent across cohort years and is appropriate for the intended deployment setting within an individual school. However, differences in mark distributions across school can affect generalization. The model may perform less reliably in schools whose mark distributions differ substantially from the training data. This is consistent with the variation in LOSO recall, which ranged from 53.1% to 94.6%, as reported in Section 4.2.4.

Third, the observational design does not allow causal conclusions to be drawn. Although the Class 3 Mathematics breakpoint is observed consistently across schools, attributing this pattern to the curriculum is based on the evidence against school-level explanations rather than experimental evidence. A quasi-experimental design examining differences in the timing of curriculum implementation or comparing schools using different editions of the Class 3 textbook could provide stronger evidence of a causal relationship.

Fourth, the teacher-related data have some limitations. The raw rosters were cleaned to identify 370 primary-section teachers (see Figure 1 note). However, some uncertainty remains because the role classifications in the source records are not always clear. Therefore, the teacher-student ratios reported in Table 1 should be considered approximate. These limitations affect the contextual covariates in the prediction model but do not affect the primary findings on SSF decomposition or the Class 3 transition, which are based on student-level data.

To assess the contribution of teacher covariates to prediction performance, a feature ablation was conducted removing all four school-level teacher variables (Avg_Teacher_Experience, Teacher_Student_Ratio, Total_Teachers, STEM_Teacher_Count) from the XGBoost model. Under stratified 5-fold cross-validation, the ablated model yields 86.67% accuracy and 80.78% at-risk recall, a change of -0.4 percentage points in recall and less than 0.1 percentage points in accuracy relative to the full-feature model (86.71% accuracy, 81.20% recall). Under temporal validation (train 2024, test 2025), the no-teacher model achieves 76.0% at-risk recall compared with 73.5% for the full-feature model, a marginal improvement of $+2.5$ percentage points consistent with the teacher variables introducing slight school-composition overfitting that reduces cross-year portability.

Teacher-related variables add negligible predictive signal compared with student-level half-yearly records. Since these variables remain constant for all students within a school, they cannot distinguish between individual students or separate sections. The early warning classifier therefore relies primarily on half-yearly examination records rather than the teachersínformation. This is a practical advantage for deployment in schools where complete teacher roster data are unavailable.

6. Conclusion

Academic failure in Bangladesh's primary schools originates in one place and is mislabeled in another. Despite this mislabeling, the affected students can still be identified before the problem becomes chronic. This study provides evidence for all three claims across nine heterogeneous institutions in Sylhet, using records that schools already collect.

The pattern of failure appears to be connected. The Mathematics failure rate shows a consistent increase from 8.5% to 28.8% at Class 3 across the sampled schools with available data. This increase is observed across schools with differences in teacher experience, class size and funding sources. The consistency of this finding suggests that the pattern cannot be explained by school-level variables. Instead, it highlights the national Class 3 curriculum and also the sharp rise in cognitive load it at class 3. Sim-

ilar patterns have been documented in the international grade-transition literature (Kaur et al., 2022). The present study also provides evidence of this pattern for the first time across heterogeneous institution types within a single Bangladeshi city.

The mislabeling problem lies in the grading rule. The conjunctive rule assigns GPA = 0 regardless of how many subjects a student fails, grouping together students who failed only Mathematics at the Class 3 transition with students who failed broadly. In this dataset, at least 38% of GPA = 0 records belong to the first group. This conflation misdirects support resources and distorts reported failure rates. A compensatory calculation applied alongside the existing rule would reduce the reported at-risk rate from 22.1% to 14.0% without requiring a new assessment. It would also restore the distinction between two populations that need fundamentally different supports.

The results indicate that at-risk students can be identified before the end of the school year. An XGBoost classifier trained on half-yearly records achieved 74.9% at-risk recall when applied to a new cohort year. This cross-year evaluation is particularly relevant to deployment, yet fewer than one in five published early warning studies report this type of evaluation (Albreiki et al., 2021). Failure breadth at mid-year (the number of subjects in which a student is failing in July) predicts chronic year-end risk more reliably than any single-subject mark. It also provides a basis for selecting appropriate interventions: subject-specific support for students failing one or two subjects and broader academic support for those failing four or more.

The three problems are closely connected. The curriculum generates the failure, the grading rule obscures its origin and the early-warning model identifies affected students before their year-end results are available. These findings suggest two complementary actions. At the school level, the classifier can be deployed using existing half-yearly records. At the system level, the grading rule can be reformed alongside an audit of the Class 3 Mathematics transition. Both are achievable within current institutional capacity. The data is publicly available (Rakib et al., 2026) to support replication across other districts and educational contexts.

Data Availability Statement

The de-identified dataset supporting this study, including 4,632 anonymized student records, teacher roster features, and model input features, is publicly available on Zenodo (Rakib et al., 2026; https://doi.org/10.5281/zenodo.21761209), licensed under CC BY 4.0. Student names have been replaced with SHA-256-derived pseudonymous identifiers (*STU_* prefix); all academic and demographic variables are retained. The identification mapping is held by the authors and will not be publicly released.

Ethics Declaration

This study analyses retrospective school administrative records; no students were recruited, contacted, or intervened upon. Academic results for seven schools were downloaded from their publicly accessible Pathshala school management portals; results for SUSC were extracted from published PDF result sheets; and data for OHS were provided directly by the school under institutional permission. All student-identifying fields were pseudonymized before any analysis (see Data Availability Statement).

Acknowledgments

The authors thank the principals and administrative staff of the nine participating schools for providing access to academic records. This work was supported by the Promotional Research Grant 2025-26, provided by the SUST Research Center, Shahjalal University of Science and Technology.

References

Adnan, M., Habib, A., Asharf, J., Mussadiq, S., Raza, A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access, 9*, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049877

Ahmmed, S., Saha, J., & Tamal, M. A. (2022). An empirical study for determining the quality indicators for the primary and secondary school of Bangladesh: A structural equation modeling approach. *Heliyon, 8*(10), e10870. https://doi.org/10.1016/j.heliyon.2022.e10870

Ahmmed, S., & Uddin, M. (2022). Exploring the disparities in learning outcomes among the primary school students of Bangladesh. *International Journal of Educational Development, 93*, 102644. https://doi.org/10.1016/j.ijedudev.2022.102644

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552. https://doi.org/10.3390/educsci11090552

Alspaugh, J. W. (1998). Achievement loss associated with the transition to middle school and high school. *Journal of Educational Research, 92*(1), 20–25. https://doi.org/10.1080/00220679809597572

Asadullah, M. N., & Chaudhury, N. (2015). The dissonance between schooling and learning: Evidence from rural Bangladesh. *Comparative Education Review, 59*(3), 447–472. https://doi.org/10.1086/681929

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal, 96*(2), 77–100. https://doi.org/10.1353/hsj.2013.0000

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Directorate of Primary Education & Australian Council for Educational Research. (2022). *The National Student Assessment 2022: Grades 3 and 5.* Government of Bangladesh/UNICEF/UKAID/ACER. https://research.acer.edu.au/ar_misc/69/

Eccles, J. S., Wigfield, A., Midgley, C., Reuman, D., Mac Iver, D., & Feldlaufer, H. (1993). Negative effects of traditional middle schools on students' motivation. *The Elementary School Journal, 93*(5), 553–574. https://doi.org/10.1086/461740

Glewwe, P., Hanushek, E. A., Humpage, S., & Ravina, R. (2013). School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. In P. Glewwe (Ed.), *Education policy in developing countries* (pp. 13–64). University of Chicago Press. https://doi.org/10.7208/9780226078854-002

Glewwe, P., & Muralidharan, K. (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 5, pp. 653–743). Elsevier. https://doi.org/10.1016/B978-0-444-63459-7.00010-5

Goos, M., Pipa, J., & Peixoto, F. (2021). Effectiveness of grade retention: A systematic review and meta-analysis. *Educational Research Review, 34*, 100401. https://doi.org/10.1016/j.edurev.2021.100401

Haladyna, T. M., & Hess, R. K. (1999). An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment, 6*(2), 129–153. https://doi.org/10.1207/S15326977EA0602_03

Heppen, J. B., & Therriault, S. B. (2008). *Developing early warning systems to identify potential high school dropouts.* National High School Center, American Institutes for Research.

Kaur, T., McLoughlin, E., & Grimes, P. (2022). Mathematics and science across the transition from primary to secondary school: A systematic literature review. *International Journal of STEM Education, 9*, 13. https://doi.org/10.1186/s40594-022-00328-0

Kumar, M., Singh, N., Wadhwa, J., Singh, P., Kumar, G., & Qtaishat, A. (2024). Utilizing Random Forest and XGBoost data mining algorithms for anticipating students' academic performance. *International Journal of Modern Education and Computer Science, 16*(2), 29–44. https://doi.org/10.5815/ijmecs.2024.02.03

Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal, 18*(1), 14. https://doi.org/10.5334/dsj-2019-014

Metzler, J., & Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics, 99*(2), 486–496. https://doi.org/10.1016/j.jdeveco.2012.06.002

Mnyawami, Y. N., Maziku, H. H., & Mushi, J. C. (2022). Enhanced model for predicting student dropouts in developing countries using automated machine learning approach: A case of Tanzanian's secondary schools. *Applied Artificial Intelligence, 36*(1). https://doi.org/10.1080/08839514.2022.2071406

Rakib, F. R., et al. (2026). *Predicting at-risk primary students in Sylhet, Bangladesh: Grading artifacts, half-yearly signals, and the Class 3 failure breakpoint* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.21761209

UNESCO Institute for Statistics. (2023). *Education and literacy: Bangladesh.* https://uis.unesco.org

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in Psychology, 12*, 698490. https://doi.org/10.3389/fpsyg.2021.698490