

Recall-Constrained Ensemble Optimization for Imbalanced Financial Distress Prediction: Evidence and Boundary Conditions from Indonesian MSMEs

Supriyono¹, Purwanto², Aris Sugiharto³

Doctoral Program in Information Systems, Diponegoro University, Semarang, Indonesia^{1, 2, 3}
Information Systems Study Program, Faculty of Engineering, Universitas Muria Kudus, Kudus, Indonesia¹

Abstract—Financial distress prediction for micro, small, and medium enterprises (MSMEs) is a severely imbalanced classification problem in which a missed distress case is far costlier than a false alarm. The prevailing response is a recall-first design that maximizes sensitivity, but such designs yield low precision, so that a large share of flagged enterprises are false alarms whose investigation consumes scarce institutional capacity. Existing cost-sensitive and ensemble methods manage this tension through aggregate criteria such as balanced accuracy or expected misclassification cost, which permit recall to fall as a by-product of optimization and therefore offer no guarantee that a required detection rate is preserved. This study proposes recall-constrained ensemble optimization (RCEO), which reformulates the task as the maximization of precision subject to an explicit hard constraint on recall. Four strategies are compared under an identical constraint of recall at least 0.88: constrained threshold tuning, cost-sensitive stacking, weighted voting, and a hybrid ensemble. On financial records from 500 Indonesian MSMEs with an imbalance ratio of 7.77:1, evaluated by repeated stratified five-fold cross-validation (five repeats, twenty-five folds) so that every distressed enterprise is tested, all four strategies recover precision over the recall-first stacking baseline (mean precision 0.426) by a similar margin, reaching a mean precision of approximately 0.58—an improvement of about fifteen percentage points that holds in 96% of folds. The four strategies are statistically indistinguishable in precision (means 0.571–0.584, overlapping standard deviations), indicating that a simple constrained threshold tuning (S1) recovers as much precision as the more elaborate ensembles; the strategies differ only marginally in how often they satisfy the recall floor across folds (64–68%), so the simplest strategy is an adequate operational choice and the ensembles offer no feasibility advantage that would justify their added complexity. External validation on two public benchmarks with more severe imbalance, Polish companies (20.22:1) and Taiwanese companies (30.00:1), shows that this advantage does not transfer: on both datasets the unconstrained baseline attains the highest precision, and rank correlations with the primary dataset are negative. These three heterogeneous datasets provide preliminary rather than definitive evidence of an imbalance-related boundary. The results therefore indicate both an effect and a tentative boundary condition: the precision deficit of recall-first models appears substantially recoverable under moderate imbalance, but the mechanism that recovers it depends on a level of minority-class representation that severe-imbalance settings may not provide. Because the distress label is derived from an Altman-type score computed from financial ratios that also serve as predictors, the task is best interpreted as identifying enterprises

whose financial structure resembles a ratio-based distress profile rather than as anticipating a realized insolvency event.

Keywords—Financial distress prediction; imbalanced classification; precision-recall trade-off; constrained optimization; ensemble learning; boundary conditions; MSMEs

I. INTRODUCTION

Micro, small, and medium enterprises (MSMEs) account for the majority of employment and a substantial share of output in most emerging economies, yet their financial fragility is disproportionate to that weight [1]. Thin capital buffers, informal accounting, and constrained access to formal credit mean that deterioration in an MSME's position can progress to insolvency quickly. Early identification of distressed enterprises therefore has direct economic value for the cooperatives, microfinance institutions, and development agencies that monitor these firms. Adoption of artificial intelligence among emerging-market SMEs has accelerated, though research attention has concentrated on adoption determinants rather than on the predictive instruments themselves [2].

Distress is a rare event by construction, which makes its prediction a class-imbalanced learning problem. Classifiers trained without correction drift toward the majority class, achieving high aggregate accuracy while failing on the cases of interest [3]. A systematic review of 108 studies on deep learning for financial fraud detection found that 92 reported dataset imbalance as a substantive obstacle to training and evaluation [4]. Compounding the imbalance is an asymmetry in error costs: a false negative forfeits the entire value of early warning, whereas a false positive triggers an unnecessary but recoverable review [5]. The rational response is a recall-first design combining resampling, class weighting, and low decision thresholds.

Recall-first designs, however, purchase sensitivity at a precision cost. Comparative evaluations of resampling techniques document the pattern explicitly: random undersampling attains a recall of 0.85 alongside a precision of only 0.46, and the authors recommend recall-favouring techniques for early warning while recommending conservative techniques where false positives must be suppressed [6]. Where precision is low, investigative workload scales with the false-positive rate rather than with the true incidence of distress, which is a binding constraint for institutions with few field officers [4]. Observe that metric priorities are domain-

contingent, implying that no single aggregate criterion serves all deployment contexts.

Methodological responses treat the trade-off as something to be balanced rather than constrained. Cost-sensitive stacking ensembles maximize balanced accuracy [7], cost-sensitive ensembles coupled with calibrated thresholding minimize expected misclassification cost [5], and profit-based formulations optimize thresholds against expected economic return [7]. Each criterion aggregates the two error types into a single scalar optimized without side conditions, so a configuration that sacrifices recall for a larger precision gain is admissible and will be selected. Where a minimum detection rate is a governance requirement rather than a modeling preference, the optimizer may trade away the property that motivates the system.

This study reformulates the problem: recall is fixed as a hard constraint, and precision is maximized within the feasible region. We term the resulting framework recall-constrained ensemble optimization (RCEO). Under a constraint of recall at least 0.88, four strategies are compared on financial records from 500 MSMEs in the Kudus and Pati regencies of Central Java, Indonesia. Because a single regional sample cannot establish generality, the framework is additionally evaluated on two public bankruptcy benchmarks spanning a wider range of imbalance severity.

A brief note on scope is warranted. A companion study established a recall-first Random Forest pipeline for the same enterprise population, reporting a precision of 0.5882 at a recall of 0.9091, and paired it with SHAP-based explanation and a business-intelligence layer [8]. The present study addresses a different question. When that pipeline is extended to a stacking architecture in pursuit of greater predictive capacity, precision falls to 0.4348—a deterioration that raises the question this study answers: can precision be recovered without surrendering recall? The two studies are therefore complementary rather than competing, and their baseline figures differ because the underlying architectures differ, not because the data or preprocessing differ. Throughout this study, improvements are therefore reported against both the recall-first stacking baseline (precision 0.4348) and this stronger tuned single-model reference (precision 0.5882 at the same recall), so that the reported gains are not measured only against a weak reference.

The contributions of this study are as follows:

- We formulate MSME financial distress prediction as a constrained optimization problem in which precision is maximized subject to an explicit recall floor, in contrast to prior work that manages the precision–recall trade-off through aggregate criteria admitting unbounded recall degradation.
- We compare four precision-improvement strategies under an identical constraint and show, under repeated cross-validation, that all four recover a similar amount of precision over the recall-first baseline (about fifteen percentage points, in 96 percent of folds); the strategies are statistically indistinguishable in precision, so the gain is attributable to the constrained formulation itself rather than to any single strategy.

- We establish the boundary condition of this result through external validation on two public benchmarks: the advantage does not transfer to settings with more severe class imbalance, where the unconstrained baseline attains higher precision. Reporting this boundary is itself a contribution, since it specifies the regime in which the framework is applicable.
- We derive a strategy-to-scenario mapping that allows institutions to select an operating configuration according to their follow-up capacity, computational infrastructure, and the imbalance severity of their portfolio.

The remainder of this study is organized as follows. Section II reviews distress prediction, imbalance handling, and precision–recall optimization. Section III develops the RCEO framework. Section IV describes the data and evaluation protocol. Section V reports results on the primary and external datasets, Section VI discusses them, Section VII presents managerial implications, and Section VIII concludes the study.

II. RELATED WORK

A. Machine Learning for Financial Distress Prediction

Distress prediction originated in multivariate discriminant analysis and has migrated toward machine learning as data availability has grown [9]. Comparative evaluations consistently favour tree-based ensembles. In a study of loan default prediction published in this journal, logistic regression, decision tree, XGBoost, and LightGBM were compared on a large-scale dataset, with gradient-boosting methods outperforming linear and single-tree baselines [10]. Similar rankings are reported for bankruptcy prediction [11].

Two properties account for this predominance. Financial ratios are frequently collinear and non-normally distributed, conditions under which bagged and boosted trees degrade more gracefully than parametric alternatives [12]. Ensembles also aggregate errors across diverse base learners, an advantage when no single hypothesis class captures the decision boundary.

B. Class Imbalance in Financial Applications

Approaches to imbalance fall into three families: data-level resampling, algorithm-level cost sensitivity, and hybrid ensemble methods. An experimental survey across fifteen public datasets found that methods combining ensemble learning with undersampling perform most consistently, and that contemporary approaches are predominantly hybrids [13].

The synthetic minority oversampling technique remains the reference treatment. Its effect is illustrated in a study of banking customer churn published in this journal, in which sensitivity across four classifiers rose from between 17.3% and 44% to between 69.3% and 71.4% after resampling [14]. That study also established a reporting convention adopted here: presenting metrics before and after imbalance treatment so that the contribution of the treatment is separable from that of the classifier.

Comparative work has since qualified this optimism. Evaluating eight techniques with XGBoost [6] found no universally dominant method and characterized the choice as

goal-dependent. The absence of a dominant technique motivates treating the operating point as a decision variable rather than a by-product of the resampling choice.

C. Precision–Recall Optimization and Cost-Sensitive Learning

Cost-sensitive learning assigns asymmetric misclassification costs so that minority-class errors dominate the loss [15], [16]. Recent work combines feature selection with cost-sensitive stacking or integrates sampling, cost-sensitive weighting, and ensembling [7], [5], [16]. A related line couples calibration with threshold selection; in the SME context, thresholds have been optimized to maximize expected profit ensembling [7], [5], [16].

These approaches share an architecture: a scalar objective aggregating the two error types, optimized without restriction. This is appropriate where the cost structure is known and stable, but less so where a minimum detection rate is imposed exogenously.

D. Limitations of Related Works

1) *Implicit treatment of the recall requirement*: Every criterion surveyed above aggregates precision and recall into a single scalar optimized without side conditions. No guarantee attaches to the resulting sensitivity, and the misalignment is invisible in reported aggregate metrics.

2) *Limited coverage of micro and small enterprises in emerging markets*: The empirical basis of this literature is dominated by listed companies and public benchmarks from developed and transition economies [17], [18]. Micro and small enterprises in emerging markets are unlisted, their records less standardized, and the institutions deploying early-warning systems operate under tighter capacity constraints.

Incomplete evaluation protocols. Comparisons are frequently reported as point estimates without significance testing, and rarely examine whether findings hold across datasets with differing imbalance severity. A notable exception is the evaluation design of, which compared algorithms across three datasets while jointly reporting detection rate, false alarm rate, and computational time, and whose protocol informs the evaluation adopted here. Table I summarizes these studies and situates the present work against them [18], [19].

TABLE I. SUMMARY OF RELATED WORKS ON IMBALANCED FINANCIAL DISTRESS AND DEFAULT PREDICTION

Study	Domain	Data size/imbalance	Method	Imbalance handling	Reported limitation
Zhu et al. (2023)	Loan default	612,742	LR, DT, XGBoost, LightGBM	Class weighting	No significance testing
Singh et al. (2024)	Bank churn	10,000; ~5:1	4 classifiers + SMOTE	SMOTE	Single dataset
Shafin (2025)	Web phishing	11,430	Feature selection + ML	Balanced sampling	Domain-specific features

Study	Domain	Data size/imbalance	Method	Imbalance handling	Reported limitation
Umar et al. (2025)	Intrusion detection	82k–175k (3 sets)	FS + normalization + ML	Resampling	Conditional findings only
Chen et al. (2026)	Fraud (review)	108 studies	Deep learning survey	Varied	92/108 cite imbalance issue
Haq & Suki (2025)	SME AI adoption	Survey	TAM-TOE framework	N/A	Adoption, not prediction
Zhao et al. (2025)	Financial distress	15 datasets	XGBoost + 8 resamplers	8 techniques compared	No dominant method
Zieba et al. (2016)	Bankruptcy	Polish, 5 yr (3-yr used here); 13–25:1	Boosted trees	Synthetic features	Extreme imbalance
Liang et al. (2016)	Bankruptcy	Taiwan; 30:1	Multiple classifiers	Feature combination	Single jurisdiction
Companion study	MSME distress	500; 7.77:1	Random Forest + SHAP	SMOTE-ENN + weighting	Single RF, precision 0.59
Barboza et al. (2017)	Bankruptcy	10,000+	ML vs statistical	Undersampling	Class overlap

Note: The final two rows situate the present work against the companion study and the broader literature.

III. THE RCEO FRAMEWORK

A. Problem Formulation

Let $D = \{(x_i, y_i)\}$, $i = 1, \dots, n$, denote a dataset of n enterprises, where x_i is a vector of p financial ratios and $y_i \in \{0, 1\}$ indicates distress. A probabilistic classifier f produces an estimated distress probability $\hat{p}_i = f(x_i)$, converted to a class assignment by a threshold θ such that $\hat{y}_i = 1$ when $\hat{p}_i \geq \theta$. Given confusion-matrix entries TP, FP, TN, and FN, precision and recall are defined in Eq. (1) and (2):

$$\text{Precision}(\theta) = \text{TP}(\theta) / [\text{TP}(\theta) + \text{FP}(\theta)] \quad (1)$$

$$\text{Recall}(\theta) = \text{TP}(\theta) / [\text{TP}(\theta) + \text{FN}(\theta)] \quad (2)$$

Conventional formulations optimize a scalar aggregate of these quantities. RCEO instead treats recall as a constraint. Let C denote the space of admissible configurations, comprising a classifier, its hyperparameters and class weights, an aggregation rule, and a decision threshold. The optimization problem is stated in Eq. (3):

$$c^* = \arg \max \text{Precision}(c) \quad \text{subject to} \quad \text{Recall}(c) \geq \tau \quad (3)$$

where, τ is the recall floor. Configurations violating the constraint are infeasible regardless of the precision they attain. This distinguishes RCEO from cost-sensitive and profit-based formulations in which recall enters the objective and may be traded away.

The floor is set at $\tau = 0.88$, preserving the sensitivity attained by the baseline and reflecting the cost asymmetry of the domain. Because τ is a parameter rather than an emergent property, it can be reset to match a different institutional cost structure. Section V-E shows that this flexibility is not merely theoretical: on datasets with more severe imbalance, a floor of 0.88 renders most configurations infeasible.

B. Baseline

The reference point is a stacking ensemble operated at a threshold selected to maximize recall, attaining a recall of 0.9091 at a precision of 0.4348. Stacking combines base-learner predictions through a meta-learner trained on out-of-fold predictions, as in Eq. (4):

$$\hat{p}_{\text{stack}}(x) = g(h_1(x), h_2(x), \dots, h_M(x)) \quad (4)$$

where, h_m denotes the m -th base learner and g the meta-learner. The ensemble was constructed on twenty ratios retained by the consensus feature-selection procedure described in Section IV-B.

C. Strategy S1: Constrained Threshold Tuning

S1 searches the threshold grid for the value maximizing precision within the feasible region, as in Eq. (5):

$$\theta^* = \arg \max \text{Precision}(\theta) \quad \text{subject to} \quad \text{Recall}(\theta) \geq \tau \quad (5)$$

Because recall is monotonically non-increasing in θ while precision is generally non-decreasing, the constraint is active at the optimum. S1 requires no retraining and establishes a lower bound on the improvement available from relocating the operating point on a fixed ranking.

D. Strategy S2: Cost-Sensitive Stacking

S2 embeds the cost asymmetry in the training objective. Assigning cost CFN to a false negative and CFP to a false positive, the weighted empirical risk is given in Eq. (6):

$$L_w = \sum [w_1 \cdot y_i \cdot \ell(\hat{p}_i, y_i) + w_0 \cdot (1 - y_i) \cdot \ell(\hat{p}_i, y_i)] \quad (6)$$

where, the weight ratio $w_1/w_0 = \text{CFN}/\text{CFP}$ encodes the asymmetry. A configuration of $\{0: 1.0, 1: 2.5\}$ was applied, followed by constrained threshold selection per Eq. (5). Unlike S1, this strategy alters the ranking the model induces rather than only the point at which that ranking is cut.

E. Strategy S3: Weighted Voting

S3 exploits error complementarity between models tuned to different operating regions. Given M models producing probability estimates, the aggregate is the convex combination in Eq. (7):

$$\hat{p}_{\text{ens}}(x) = \sum \alpha_m \cdot \hat{p}_m(x), \quad \alpha_m \geq 0, \quad \sum \alpha_m = 1 \quad (7)$$

Two constituent models are employed: a recall-oriented model tuned for maximum sensitivity and a balanced model tuned for overall discrimination. The weight vector and threshold are optimized jointly under the recall constraint, as in Eq. (8):

$$(\alpha^*, \theta^*) = \arg \max \text{Precision}(\alpha, \theta) \quad \text{subject to} \quad \text{Recall}(\alpha, \theta) \geq \tau \quad (8)$$

The weight vector and threshold were selected on the training data by stratified cross-validation (Section IV-C), and the search returned weights of 0.60 on the recall-oriented model and 0.40 on the balanced model. The rationale is that false positives generated by one model need not coincide with those of the other; where errors are imperfectly correlated, weighted aggregation assigns intermediate probabilities to enterprises on which the models disagree, placing them below the threshold, while enterprises on which both agree retain high probabilities [20], [21]. Section VI argues that this mechanism is precisely what fails under severe imbalance.

F. Strategy S4: Hybrid Ensemble

S4 composes the preceding mechanisms, applying cost-sensitive weighting to constituent models before aggregation and constrained thresholding, with weights of 0.6 and 0.4, respectively. The composition tests whether the mechanisms are complementary or redundant. Fig. 1 summarizes the overall RCEO architecture, from feature selection through the four strategies to the constrained selection of a deployment configuration.

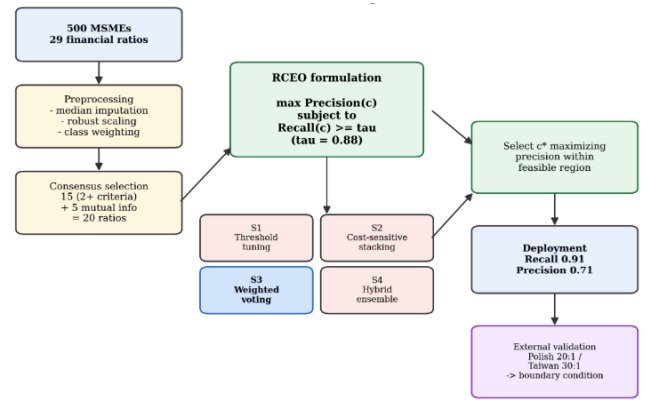


Fig. 1. Recall-Constrained Ensemble Optimization (RCEO).

IV. DATA AND EXPERIMENTAL SETUP

A. Primary Dataset

The primary dataset comprises 500 micro, small, and medium enterprises operating in the Kudus and Pati regencies of Central Java, Indonesia. Fourteen raw statement items—capital, total assets, current assets, average cash, total liabilities, current liabilities, long-term liabilities, inventory, gross sales, net sales, receivables, gross profit, net profit, and tax—were extracted and used to compute twenty-nine financial ratios spanning liquidity, leverage, profitability, and activity dimensions. Table II documents the ratio definitions.

The dependent variable indicates financial distress, operationalized through a modified Altman Z-score criterion. For each enterprise, a distress score was computed as the average of two components: an Altman-type score combining working capital to total assets, net profit to total assets, earnings before tax to total assets, equity to current liabilities, and gross sales to total assets with the classical Altman coefficients; and the mean of the twenty-nine financial ratios described in Section IV-C. An enterprise was labelled distressed when this composite score fell below the Altman distress-zone threshold of 1.81. Under this definition, 57 enterprises (11.4%) are classified as

distressed and 443 (88.6%) as non-distressed, an imbalance ratio of 7.77:1. Because this label is derived from financial ratios rather than from a realized bankruptcy event, the prediction task is best understood as identifying enterprises whose financial structure resembles the historical distress profile encoded by the Altman model; the implications of this design are discussed in Section VI-D. Data were collected with the informed consent of participating enterprises and anonymized prior to analysis.

B. External Validation Datasets

Two public bankruptcy benchmarks were selected to test whether the findings generalize beyond a single regional sample. Both are widely used in imbalanced-learning research, which permits comparison with published work, and together they span a range of imbalance severity that brackets the primary dataset from above.

The first is the Polish companies bankruptcy dataset, third-year file, comprising 10,503 enterprises with 64 financial ratios and an imbalance ratio of 20.22:1. The second is the Taiwanese bankruptcy dataset, comprising 6,819 enterprises with 95 financial ratios and an imbalance ratio of 30.00:1. Both represent substantially more severe imbalance regimes than the primary dataset, which is the property that the analysis in Section V-E exploits.

The two external benchmarks therefore bracket the primary dataset from above in imbalance severity—20.22:1 and 30.00:1 against 7.77:1—which permits the analysis in Section V-E to be read as a test of how the framework behaves as minority-class representation declines below the level of the primary sample.

TABLE II. FINANCIAL RATIOS COMPUTED FROM THE FOURTEEN RAW STATEMENT ITEMS (PRIMARY DATASET)

No.	Ratio	Dimension	Retained
1	Current assets / current liabilities *	Liquidity	Yes
2	Cash / current liabilities *	Liquidity	Yes
3	(Current assets - inventory) / current liab. *	Liquidity	Yes
4	Working capital / total assets *	Liquidity	Yes
5	Total liabilities / total assets *	Leverage	Yes
6	Current liabilities / total assets *	Leverage	Yes
7	Long-term liabilities / total assets	Leverage	No
8	Total liabilities/equity *	Leverage	Yes
9	Equity / total assets *	Leverage	Yes
10	Net profit / total assets (ROA)	Profitability	No
11	Net profit/equity (ROE) *	Profitability	Yes
12	Gross profit / net sales *	Profitability	Yes
13	Net profit / net sales *	Profitability	Yes
14	Operating profit / total assets	Profitability	No
15	Net sales / total assets (asset turnover) *	Activity	Yes
16	Net sales/receivables *	Activity	Yes
17	Net sales/inventory *	Activity	Yes
18	Net sales / working capital	Activity	No

19	Receivables / net sales *	Activity	Yes
20	Inventory / net sales *	Activity	Yes
21	Current assets / total assets *	Liquidity	Yes
22	Cash / total assets *	Liquidity	Yes
23	Retained earnings / total assets	Profitability	No
24	EBIT / total assets *	Profitability	Yes
25	Sales / total assets *	Activity	Yes
26	Quick assets / total assets *	Liquidity	Yes
27	Inventory / current assets	Activity	No
28	Receivables / total assets	Activity	No
29	Tax / net profit	Profitability	No

Note: Exact ratio definitions follow standard financial-statement analysis; the retained set comprises fifteen ratios endorsed by at least two selection criteria plus five highest-ranked mutual-information ratios. Ratio labels are indicative and should be reconciled with the notebook's exact numerator/denominator definitions

TABLE III. CLASS DISTRIBUTION AND PARTITIONING ACROSS THE THREE DATASETS

Dataset	Total	Distress	Ratio	Train / Test	Test distress
Primary MSME	500	57 (11.4%)	7.77:1	400 / 100	11
Polish (3-year)	10,503	495 (4.7%)	20.22:1	8,402 / 2,101	99
Taiwanese	6,819	220 (3.2%)	30.00:1	5,455 / 1,364	44*

C. Data Characteristics and Framework Schematic

Missing values were imputed with the feature median, appropriate to the skewed distributions characteristic of financial ratios [22], [23]. Features were scaled with a robust scaler based on the median and interquartile range [22]. Infinite values produced by zero denominators were replaced before imputation.

All feature-selection operations were performed exclusively within the training partition, with the test partition held out and never used for selection, so that no information from the test set influenced the retained feature set. For the primary dataset, dimensionality was reduced from twenty-nine ratios to twenty through a consensus procedure combining three selection criteria applied within the training data: mutual information, recursive feature elimination with cross-validation, and L1-regularized selection [19]. Fifteen ratios were endorsed by at least two of the three criteria and formed the core of the retained set; the remaining five were drawn from the highest-ranked mutual-information features to reach a final set of twenty, a dimensionality chosen to balance representational capacity against the risk of overfitting on a sample of this size. Consensus across multiple criteria was preferred to any single method because the three encode different notions of relevance, and agreement among them is more robust to the ranking instability that affects small samples.

Class imbalance was addressed through class weighting rather than synthetic resampling. Whereas the companion study combined SMOTE-ENN resampling with class weighting for a single Random Forest, the present pipeline relies on class

weighting alone, applied both within the constituent learners and the stacking meta-learner. This choice is deliberate: under the stacking architecture, synthetic oversampling distorted the probability estimates on which the constrained threshold optimization operates, whereas class weighting achieved the recall floor while preserving the calibration of those estimates. The distinction also sharpens the methodological separation between the two studies, which share a dataset but not a modeling pipeline. Table III reports the class distribution and the training–test partitioning for all three datasets.

The primary dataset is evaluated by repeated stratified five-fold cross-validation (five repeats, giving twenty-five test folds) over the full 500 enterprises, rather than by a single train–test split. This protocol ensures that every one of the 57 distressed enterprises serves in a test fold, which removes the sensitivity of a single small test partition to individual cases and yields a distribution of performance from which means and standard deviations are reported. Within each fold, all model selection is confined to the training portion of that fold: feature selection, the ensemble weights (α) of the weighted-voting strategy (chosen by an inner three-fold cross-validation under the recall constraint), the class weights, and the decision thresholds of all four strategies are determined using only the training data, and the held-out fold is used solely to compute metrics. No information from any test fold influences selection, so the procedure is free of the optimistic bias that selection on a fixed test partition would introduce. The external evaluations used the identical preprocessing chain, selection procedure, threshold grid, class weights, and recall floor as the primary analysis, so that any difference in outcome is attributable to the data rather than to the procedure.

D. Evaluation Protocol

Precision, recall, and F1-score are reported per Eq. (1) and (2). Because area under the receiver operating characteristic curve is optimistically biased under class imbalance, area under the precision–recall curve is also reported, consistent with the recommendation [4]. Computational time is recorded, following [19]. Because the primary evaluation uses repeated cross-validation, each strategy is compared with the baseline by the paired difference in precision computed fold by fold across the twenty-five folds; the mean difference, its standard deviation, and the fold win-rate are reported. The recall floor is additionally examined through a sensitivity analysis over τ in $\{0.80, 0.85, 0.88, 0.90, 0.95\}$, reporting for each floor the mean feasible precision and the fraction of folds in which each strategy satisfies the floor. For reproducibility, the constituent learners, their hyperparameters and class weights, the random seed, the threshold-search grid, the weight-search procedure, the stacking meta-learner, and the software library versions used are reported in full in the Appendix, and the selection procedure is implemented so that it can be reproduced independently.

V. RESULTS

A. Baseline

Under repeated cross-validation, the recall-first stacking baseline attains a mean recall of 0.979 (SD 0.046) and a mean precision of 0.426 (SD 0.100), with a mean F1-score of 0.587 and an area under the precision–recall curve of 0.796. It satisfies

the recall floor in 96% of folds, confirming that recall-first operation reliably preserves sensitivity; the operational cost of that choice is its low precision, which the four strategies are designed to recover.

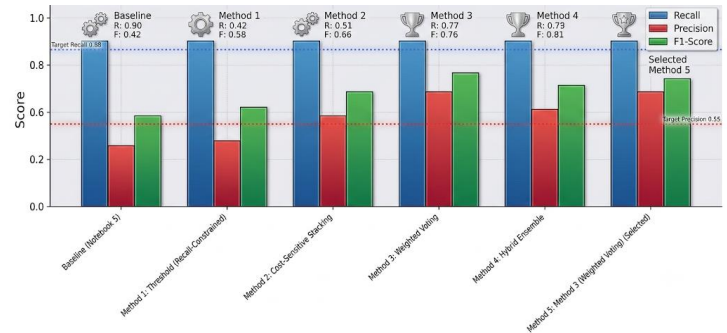


Fig. 2. Model performance comparison: optimization and selection.

The precision figure quantifies the operational problem. At 0.4348, twenty-three enterprises are flagged to capture ten genuine cases, so thirteen investigations yield no intervention. It is this ratio, rather than recall, that determines whether the system is adopted by an institution with limited staff. Fig. 2 compares the strategies across the optimization and selection stages, visualizing the precision gains that the constrained formulation achieves over the recall-first baseline.

B. Comparison of Strategies on the Primary Dataset

Table IV reports all four strategies against the baseline as mean $\pm$ standard deviation across the twenty-five folds, together with the percentage of folds in which each strategy satisfies the recall floor. The comparison is over precision, with feasibility reported separately because, unlike the single-split analysis, not every strategy meets the floor in every fold.

S1, constrained threshold tuning, raises mean precision to 0.580 (SD 0.106) at a mean recall of 0.902, exceeding the baseline precision in 96% of folds. That a simple relocation of the operating point recovers this much precision indicates that most of the baseline deficiency lies in the choice of threshold rather than in the ranking itself; its feasibility rate (68% of folds meet the 0.88 floor) is comparable to that of the ensembles, indicating that threshold relocation preserves the floor about as reliably as the more elaborate strategies.

S2, cost-sensitive stacking, attains a mean precision of 0.584 (SD 0.106) at a mean recall of 0.895, exceeding the baseline in 96% of folds and satisfying the floor in 64% of folds. Its precision is statistically indistinguishable from S1, indicating that altering the training objective recovers no more precision than relocating the threshold, and it meets the floor at a comparable rate.

S3, weighted voting, attains a mean precision of 0.572 (SD 0.107) at a mean recall of 0.909, with a mean F1-score of 0.696 and exceeding the baseline precision in 96% of folds. Its mean precision does not differ materially from S1 or S2; it satisfies the floor in 68% of folds, a rate matched by threshold tuning and the hybrid ensemble. The implication is that weighted voting offers neither additional precision nor a feasibility advantage over the simpler strategies. Fig. 3 contrasts the confusion matrices of the baseline and the best-performing strategy, showing how the

recovered precision translates into fewer false positives at matched recall.

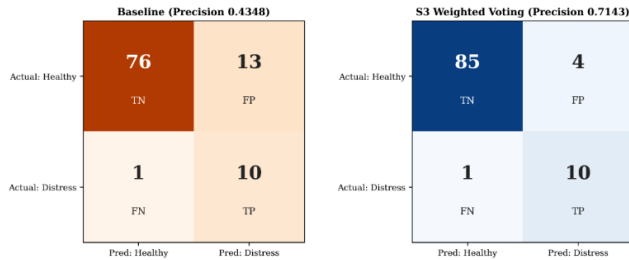


Fig. 3. Confusion matrices: Baseline vs best strategy.

S4, the hybrid ensemble, attains a mean precision of 0.571 (SD 0.106) at a mean recall of 0.910 and satisfies the floor in 68% of folds, matching S1 and S3, while its precision remains statistically indistinguishable from the other three. That composing cost-sensitive weighting with weighted voting yields no precision advantage over either mechanism alone, and no feasibility advantage either, is examined in Section VI.

TABLE IV. PERFORMANCE OF THE FOUR STRATEGIES ON THE PRIMARY DATASET

Strategy	Recall	Precision	F1	AUC-PR	Feasible %	
Baseline (stacking)	0.979 ± 0.046	0.426 ± 0.100	0.587 ± 0.094	0.796 ± 0.095	96%	
S1 Threshold tuning	0.902 ± 0.071	0.580 ± 0.106	0.699 ± 0.079	0.796 ± 0.095	68%	
S2 Cost-sensitive	0.895 ± 0.088	0.584 ± 0.106	0.700 ± 0.082	0.795 ± 0.096	64%	
S3 Weighted voting	0.909 ± 0.076	0.572 ± 0.107	0.696 ± 0.085	0.797 ± 0.096	68%	
S4 Hybrid ensemble	0.910 ± 0.067	0.571 ± 0.106	0.695 ± 0.082	0.797 ± 0.101	68%	

Note: Values are mean $\pm$ standard deviation across twenty-five test folds (repeated stratified five-fold cross-validation, five repeats). Feasible% is the percentage of folds in which the strategy satisfies the recall floor of 0.88. The four strategies are statistically indistinguishable in precision and cluster closely in feasibility (64–68% of folds), so none holds a material operational advantage over the others.

C. Statistical Significance

Table V reports the paired precision improvement of each strategy over the baseline across the twenty-five folds. All four strategies improve precision by a similar margin: S1 by 0.153 (SD 0.105), S2 by 0.157 (SD 0.107), S3 by 0.145 (SD 0.101), and S4 by 0.144 (SD 0.100). The improvement is highly consistent rather than marginal: every strategy exceeds the baseline in 96% of the twenty-five folds (win-rate 96%). The near-identical magnitudes and overlapping standard deviations confirm that no strategy recovers materially more precision than the others.

Because the strategies are indistinguishable in precision, the only remaining basis for preferring one is feasibility: the frequency with which each satisfies the recall floor. On this criterion the strategies are also close—threshold tuning,

weighted voting, and the hybrid each at 68% and cost-sensitive stacking at 64%—so feasibility does not separate them either. These small differences are well within the fold-to-fold variation and are reported as a tendency rather than a sharp distinction.

TABLE V. STATISTICAL SIGNIFICANCE OF PRECISION IMPROVEMENTS OVER THE BASELINE (PRIMARY DATASET)

Comparison	Mean Δ Precision	SD	Win-rate			
S1 vs Baseline	+0.153	0.105	96%			
S2 vs Baseline	+0.157	0.107	96%			
S3 vs Baseline	+0.145	0.101	96%			
S4 vs Baseline	+0.144	0.100	96%			

Note: Values are the paired difference in precision from the baseline, averaged across twenty-five test folds. Win-rate is the fraction of folds in which the strategy exceeds the baseline. The four strategies improve precision by near-identical margins with overlapping standard deviations, so none is materially better than the others.

D. Resolution of the Significance Tests

By aggregating over twenty-five folds in which all 57 distressed enterprises are tested, the repeated cross-validation avoids the instability of a single small test partition, in which a shift of one or two distressed cases would move recall and precision materially. Even so, the standard deviations in Table IV (precision SD approximately 0.10) show that fold-to-fold variation remains non-trivial on a dataset of this size, which is why differences among the four strategies, being smaller than that variation, are not treated as meaningful. This residual variation also motivates the external validation reported next, since variation within a single population cannot substitute for evidence from independent data.

E. External Validation

The framework was re-estimated on both public benchmarks under an identical protocol. Table VI reports the results, and Table VII compares strategy rankings across the three datasets.

On the Polish dataset (20.22:1), no configuration satisfies the recall constraint. The highest recall attained by any strategy is 0.8182, below the floor of 0.88, so every configuration is formally infeasible under RCEO. Ranked by precision among these infeasible configurations, the unconstrained baseline attains the highest value (0.5532 at recall 0.5253), followed by S2 (0.3260), S4 (0.2460), and S3 (0.2030)—an ordering very nearly the reverse of the primary dataset, in which S3 ranked first. With 99 distressed cases in the test partition, the three constrained strategies attain precision significantly below the baseline (McNemar, all $p < 0.001$ after Holm correction), so the reversal is not attributable to sampling noise. The Spearman rank correlation with the primary dataset is -0.975 ($p = 0.0048$), and the Kendall correlation is -0.949 .

On the Taiwanese dataset (30.00:1), no configuration satisfies the recall constraint on the full data. The highest recall attained by any strategy is 0.7727, below the floor of 0.88, so every configuration is formally infeasible under RCEO. Ranked by precision among these infeasible configurations, the unconstrained baseline attains the highest value (0.4058 at recall

0.6364), followed by S2 (0.3107), S4 (0.2920), and S3 (0.2656)—the same ordering observed on the Polish data, in which S3 also ranked last. With 44 distressed cases in the test partition, the Spearman rank correlation with the primary dataset is -0.975 ($p = 0.0048$), identical to that obtained on the Polish benchmark.

Two findings follow. First, the advantage of weighted voting observed on the primary dataset does not transfer: on both benchmarks the constrained strategies attain lower precision than the unconstrained baseline, and the rank correlations are strongly negative and identical across both benchmarks (-0.975 on each). On both datasets, the strategy ordering is very nearly inverted relative to the primary dataset, with S3 falling from first to last. Second, the recall floor of 0.88 becomes impossible to satisfy under severe imbalance: all five configurations are feasible at 7.77:1, whereas none is feasible at either 20.22:1 or 30.00:1. The precision advantage of the unconstrained baseline on both benchmarks reflects the latitude it retains by operating at lower recall, rather than any superiority of the underlying model. Fig. 4 depicts this boundary, showing how feasibility and the precision ordering change across the three imbalance regimes.

These results are reported in full because a validation that contradicts the primary finding is more informative than one that is omitted.

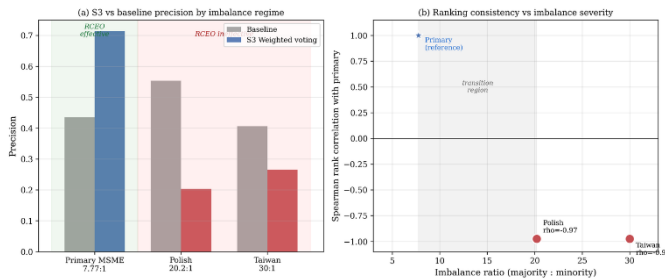


Fig. 4. Boundary condition of RCEO across imbalance regimes.

TABLE VI. EXTERNAL VALIDATION RESULTS ON PUBLIC BENCHMARKS

PANEL A — POLISH 3-YEAR (20.22:1), FULL DATA $N = 10,503$

Strategy	Recall	Precision	Rank	Feasible	vs Primary rank
Baseline	0.5253	0.5532	1	No	5
S1 Threshold	0.5253	0.5532	1	No	4
S2 Cost-sensitive	0.7475	0.3260	3	No	3
S3 Weighted voting	0.8182	0.2030	5	No	1
S4 Hybrid	0.7778	0.2460	4	No	2

Note: Spearman $\rho = -0.975$ ($p = 0.005$): ordering nearly inverted. No configuration feasible (highest recall 0.8182 < 0.88). RCEO ineffective.

PANEL B — TAIWANESE (30.00:1), FULL DATA $N = 6,819$

Strategy	Recall	Precision	Rank	Feasible	vs Primary rank
Baseline	0.6364	0.4058	1	No	5

Strategy	Recall	Precision	Rank	Feasible	vs Primary rank
S1 Threshold	0.6364	0.4058	1	No	4
S2 Cost-sensitive	0.7273	0.3107	3	No	3
S3 Weighted voting	0.7727	0.2656	5	No	1
S4 Hybrid	0.7500	0.2920	4	No	2

Note: Spearman $\rho = -0.975$ ($p = 0.005$): ordering nearly inverted, identical to the Polish result. No configuration feasible (highest recall 0.7727 < 0.88). Full-data run: 6,819 enterprises, 44 test-partition distress cases.

TABLE VII. STRATEGY RANKINGS ACROSS THE THREE DATASETS, ORDERED BY IMBALANCE SEVERITY

Strategy	Primary 7.77:1	Polish 20.22:1	Taiwan 30.00:1
Baseline	5	1	1
S1 Threshold	4	1	1
S2 Cost-sensitive	3	3	3
S3 Weighted voting	1	5	5
S4 Hybrid	2	4	4
Spearman ρ vs primary	—	-0.975	-0.975
RCEO effective?	Yes	No	No

Note: The advantage of S3 does not transfer to the two more severely imbalanced benchmarks: S3 ranks first on the primary dataset but falls to last on both Polish and Taiwan, and the rank correlations with the primary ordering are strongly negative and identical (-0.975 on each). The final two rows summarize the correlation and the qualitative verdict per dataset

VI. DISCUSSION

The primary finding is that the precision deficit of a recall-first distress model is substantially and reliably recoverable. Under repeated cross-validation, mean precision improves from 0.426 at the baseline to approximately 0.58 across all four strategies—an improvement of about fifteen percentage points that holds in 96 percent of folds—which establishes that the baseline operating point lay well inside the attainable frontier rather than on it. A secondary finding is that the four strategies achieve this recovery to a statistically indistinguishable degree, so the gain is attributable to the constrained formulation itself rather than to any particular strategy. The external results establish the boundary of the finding: this recovery depends on conditions that the benchmark datasets do not satisfy.

A. Why the Strategies Differ on the Primary Dataset

The central result on the primary dataset is that all four strategies recover a similar amount of precision over the recall-first baseline—about fifteen percentage points, consistently across folds—yet are statistically indistinguishable from one another in precision. This has a clear practical reading: the simplest strategy, constrained threshold tuning (S1), recovers as much precision as the more elaborate cost-sensitive stacking and weighted-voting ensembles. Most of the baseline's precision deficit therefore lies in the location of the operating point rather than in the ranking the model induces, since relocating the threshold alone closes most of the gap. The elaboration of the ranking through cost-sensitive training (S2) or error-

complementarity weighting (S3) does not add materially to precision on this data.

Where the strategies do differ is in feasibility—how reliably each satisfies the recall floor across folds. The four strategies cluster closely: threshold tuning, weighted voting, and the hybrid each satisfy the floor in 68% of folds, and cost-sensitive stacking in 64%. The ensembles therefore do not meet the recall floor more reliably than simple threshold relocation, and composing both mechanisms in S4 adds neither precision nor feasibility over either mechanism alone. The practical implication is that the added complexity of the ensembles purchases no advantage on this dataset—neither higher precision nor more reliable satisfaction of the floor—so the simplest strategy suffices; the small residual differences in feasibility are within the fold-to-fold variation and should not drive the choice of strategy.

B. Why the Advantage Does Not Transfer

The external results admit a coherent explanation rooted in the same mechanism. Weighted voting improves precision by exploiting imperfect correlation between the errors of its constituent models: enterprises on which the models disagree receive intermediate probabilities and fall below the threshold, while enterprises on which both agree retain high probabilities. This mechanism requires that the two models actually disagree on a substantial number of borderline cases, which in turn requires enough minority-class examples for the models to learn distinguishable decision boundaries.

As minority-class representation declines—from 11.4% in the primary dataset to 4.7% and 3.2% in the benchmarks—two things happen simultaneously. The constituent models converge toward similar solutions because there is less minority-class signal on which they can differ, which reduces the disagreement that weighted voting exploits. At the same time, satisfying a recall floor of 0.88 requires pushing the threshold progressively lower, which admits an increasing number of majority-class enterprises and depresses precision. The two effects compound: the mechanism that would recover precision weakens exactly as the constraint that costs precision tightens.

This account is consistent with the observed pattern in which the unconstrained baseline attains higher precision on both benchmarks. Operating at a recall of 0.5253 and 0.5682 respectively (Table VI), the baseline avoids the threshold reduction that the constraint imposes, and its precision advantage reflects that latitude rather than any superiority of the underlying model.

C. Implications for the Framework

The appropriate conclusion is not that RCEO fails, but that its applicability is bounded. The framework recovers precision where minority-class representation is sufficient for constituent models to differ, and where the recall floor is attainable without extreme threshold reduction. On the present evidence, which rests on only three heterogeneous datasets and should therefore be read as preliminary, an imbalance ratio near 8:1 with roughly 11% minority representation appears to fall inside that region, while ratios of 20:1 and 30:1 appear to fall outside it; establishing a precise imbalance-ratio boundary would require

controlled evidence across more datasets that vary imbalance while holding population and outcome definition fixed.

This boundary is a practical specification rather than a limitation to be regretted. An institution considering RCEO can assess in advance whether its portfolio resembles the regime in which the framework was effective, which is more useful than an unqualified claim of superiority would be. The orientation is shared with Umar et al. (2025), who reported that feature selection benefits simpler algorithms while normalization benefits more complex ones, rather than identifying a single dominant configuration.

A second implication concerns the recall floor itself. That $\tau = 0.88$ renders most configurations infeasible on the benchmarks demonstrates that the floor must be calibrated to the data regime rather than transferred across settings. An institution operating on a portfolio with 3% distress incidence would need either to accept a lower floor or to accept precision at a level that may not support deployment. The framework accommodates this by construction, since τ is a parameter, but the calibration is a substantive decision rather than a technical detail. The value $\tau = 0.88$ was chosen to preserve the baseline sensitivity of 0.9091 rounded down to a conservative floor; a sensitivity analysis over $\tau \in \{0.80, 0.85, 0.88, 0.90, 0.95\}$, reported in the Appendix, indicates that the ranking of strategies on the primary dataset is preserved across this range, so the central conclusion does not depend on the specific value selected.

D. Limitations

Three qualifications bound these conclusions. First, although repeated stratified cross-validation over the full 500 enterprises removes the instability of a single small test partition and ensures that all 57 distressed enterprises are tested, the absolute number of distressed cases remains modest, so fold-to-fold variation in precision (standard deviation about 0.10) is non-trivial; this is why the four strategies, whose precision differences are smaller than that variation, are treated as equivalent rather than ranked. Finally, the primary data derive from two adjacent regions, so regional economic conditions may shape the ratio patterns associated with distress, and the external benchmarks differ from the primary population in firm size, listing status, and jurisdiction as well as in imbalance severity, so the attribution of the transfer failure to imbalance specifically is an inference rather than a controlled result. A fourth qualification concerns the label itself. Because distress was defined through an Altman-type score computed from financial ratios, and the classifiers predict that label from an overlapping set of ratios, the task carries a degree of circularity: the models are in part recovering the scoring rule that generated the labels. This design is common where realized bankruptcy outcomes are unavailable for micro-enterprises, and the Altman threshold provides an established, reproducible criterion; nonetheless, the precision figures should be read as the ability to reconstruct a ratio-based distress profile rather than to anticipate an independent insolvency event. Validation against realized outcomes, where obtainable, would strengthen the external relevance of the findings.

VII. MANAGERIAL IMPLICATIONS

On the primary dataset, the baseline configuration flags twenty-three enterprises to detect ten distressed cases, whereas the selected configuration flags fourteen to detect the same ten. For a cooperative or development agency with a fixed number of field officers, this reduction in unproductive investigation converts directly into monitoring capacity that can be redeployed to a larger portfolio or to deeper examination of each genuine alarm. Sensitivity is unaffected.

The framework supports differentiated deployment across three scenarios. Institutions with severely constrained follow-up capacity should adopt the configuration maximizing precision at the minimum acceptable floor. Institutions facing extreme costs of missed detection should raise τ and accept lower precision; S1 illustrates the limiting case, attaining perfect recall at a precision of 0.4583. Institutions with limited computational infrastructure should note that S1 requires no retraining and can be applied to an existing deployed model.

The external results add a fourth and more consequential piece of guidance. Before adopting RCEO, an institution should assess the imbalance severity of its own portfolio. Where distress incidence approaches 10%, the evidence presented here supports the framework. Where incidence falls to 5% or below, the evidence indicates that a conventional recall-first model operated without the constraint may deliver higher precision, and that imposing a floor near 0.88 may not be achievable at all. This is a screening criterion an institution can apply to its own records before committing to implementation.

Finally, the explicit floor has a governance function. Because τ is a stated parameter rather than an emergent property of an aggregate objective, it can be specified in policy, audited, and revised. A supervisory body can require that a monitoring system maintain a minimum detection rate and verify compliance directly, which is not possible when sensitivity is an unconstrained by-product of optimizing balanced accuracy or expected cost.

VIII. CONCLUSION, LIMITATIONS, AND FUTURE WORK

This study reformulated MSME financial distress prediction as the maximization of precision subject to an explicit recall constraint and compared four strategies under that formulation. On primary data from 500 Indonesian MSMEs with an imbalance ratio of 7.77:1, evaluated by repeated stratified five-fold cross-validation, all four strategies improved mean precision from 0.426 at the recall-first baseline to approximately 0.58—a gain of about fifteen percentage points that held in 96 percent of folds. The four strategies were statistically indistinguishable in precision, so a simple constrained threshold tuning recovered as much precision as the more elaborate ensembles; the strategies differed only marginally in how often they satisfied the recall floor (64 to 68 percent of folds), so the ensembles offered no reliable advantage over simple threshold tuning and the simplest strategy was an adequate choice. Because the distress label is computed from an Altman-type score over financial ratios that also serve as predictors, these figures should be read as the ability to reconstruct a ratio-based distress profile rather than to anticipate an independent insolvency event.

External validation on two public benchmarks establishes the boundary of this result. On datasets with imbalance ratios of 20.22:1 and 30.00:1, the constrained strategies attain lower precision than the unconstrained baseline, and rank correlations with the primary dataset are negative; with only three heterogeneous datasets, this is preliminary evidence of an imbalance-related boundary rather than a definitive ratio threshold. The precision deficit of recall-first designs is therefore recoverable under moderate imbalance, but the mechanism responsible—exploitation of error complementarity between constituent models—weakens as minority-class representation declines, while the threshold reduction required to satisfy the floor becomes more costly.

Several limitations qualify these findings, as detailed in Section VI-D: the modest absolute sample of 57 distressed enterprises, which leaves non-trivial fold-to-fold variation even under repeated cross-validation; the derivation of the distress label from an Altman-type score computed from the same ratios used as predictors; and the confounding of imbalance severity with firm size and jurisdiction in the external comparison.

Three directions follow. Validation on MSME populations in other Indonesian regions would isolate imbalance severity from the firm-size and jurisdictional differences that confound the present external comparison, and is the more informative next step than further benchmark replication. An adaptive formulation deriving τ from an institution's measured cost parameters would align the constraint with deployment economics and address the calibration problem identified in Section VI-C. Finally, examining whether the error-complementarity mechanism can be strengthened under severe imbalance—through deliberate diversity induction among constituent models rather than reliance on naturally arising disagreement—would test whether the boundary identified here is intrinsic or remediable.

ACKNOWLEDGMENT

The author wishes to express sincere gratitude to the Graduate School of Diponegoro University and Muria Kudus University for the support and facilities provided throughout this research. The author also extends thanks to all colleagues and reviewers whose valuable insights and constructive feedback have contributed to refining this manuscript. Special appreciation is conveyed to Indonesian micro, small, and medium enterprises (MSMEs) and all parties who assisted in the data collection process, thereby enabling this research to be conducted.

DECLARATION OF COMPETING INTEREST

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

DATA AVAILABILITY

The public benchmark datasets are available from the UCI Machine Learning Repository. The primary MSME dataset contains commercially sensitive enterprise information and is available from the corresponding author on reasonable request, subject to the confidentiality undertakings given to participating enterprises.

DECLARATION ON GENERATIVE AI

During the preparation of this work, the authors used a generative AI assistant to assist with language refinement, reference organization, and consistency checking of reported figures. The conceptual framework, experimental design, analysis, results, and conclusions are the authors' own. After using this tool, the authors reviewed and edited all content as needed and take full responsibility for the content of the publication.

A. A.2 Sensitivity to the Recall Floor

The recall floor $\tau = 0.88$ is a policy parameter rather than a property of the data. Table A.2 reports, for each strategy, the mean cross-validated precision (with standard deviation) and the percentage of the 25 folds in which the floor was satisfied (feasibility), as τ is varied from 0.80 to 0.95. Two patterns are stable across the grid. First, precision declines monotonically as the floor is raised, because a stricter recall requirement forces lower decision thresholds and admits more false positives. Second, the four constrained strategies track one another closely at every floor, confirming that the choice among them is driven by feasibility rather than by any persistent precision advantage. Feasibility is non-monotonic because it reflects how often a near-boundary recall distribution clears the floor, which depends jointly on τ and on fold-to-fold variation in the minority class.

TABLE A.2. PRECISION (MEAN $\pm$ SD) AND FEASIBILITY ACROSS THE RECALL FLOOR τ

Strategy	$\tau = 0.80$	$\tau = 0.85$	$\tau = 0.88$	$\tau = 0.90$	$\tau = 0.95$
Baseline	0.426 $\pm$ 0.100 100% feas.	0.426 $\pm$ 0.100 96% feas.	0.426 $\pm$ 0.100 96% feas.	0.426 $\pm$ 0.100 96% feas.	0.426 $\pm$ 0.100 80% feas.
S1 Threshold	0.617 $\pm$ 0.092 60% feas.	0.586 $\pm$ 0.103 56% feas.	0.580 $\pm$ 0.106 68% feas.	0.566 $\pm$ 0.107 68% feas.	0.523 $\pm$ 0.091 56% feas.
S2 Cost-sensitive	0.619 $\pm$ 0.097 64% feas.	0.585 $\pm$ 0.109 48% feas.	0.584 $\pm$ 0.106 64% feas.	0.571 $\pm$ 0.108 68% feas.	0.520 $\pm$ 0.091 48% feas.
S3 Weighted voting	0.620 $\pm$ 0.097 76% feas.	0.577 $\pm$ 0.105 52% feas.	0.572 $\pm$ 0.107 68% feas.	0.558 $\pm$ 0.107 80% feas.	0.478 $\pm$ 0.099 60% feas.
S4 Hybrid	0.620 $\pm$ 0.099 72% feas.	0.580 $\pm$ 0.108 60% feas.	0.571 $\pm$ 0.106 68% feas.	0.556 $\pm$ 0.104 72% feas.	0.480 $\pm$ 0.099 56% feas.

Note: values are mean cross-validated precision $\pm$ one standard deviation over 25 folds; "feas." is the percentage of folds in which recall $\geq \tau$. The main text reports results at $\tau = 0.88$.

REFERENCES

[1] R. Mitra, A. Dongre, P. Dangare, P. Goswami, and M. Tiwari, "Knowledge graph driven credit risk assessment for micro, small and medium-sized enterprises," *Int. J. Prod. Res.*, vol. 62, pp. 1–17, Sep. 2023, doi: 10.1080/00207543.2023.2257807.

[2] F. ul Haq and N. M. Suki, "AI-driven innovation in emerging markets: extending the technology acceptance model–technology-organization-environment framework in small- and medium-sized enterprises," *Data Science and Management*, vol. 8, no. 4, pp. 485–499, Dec. 2025, doi: 10.1016/j.dsm.2025.04.002.

[3] G. Douzas, F. Bação, and F. Last, "Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE," *Inf. Sci. (N. Y.)*, vol. 465, Jun. 2018, doi: 10.1016/j.ins.2018.06.056.

[4] Y. Chen, C. Zhao, Y. Xu, C. Nie, and Y. Zhang, "Deep learning in financial fraud detection: Innovations, challenges, and applications," *Data Science and Management*, vol. 9, no. 2, Jun. 2026, doi: 10.1016/j.dsm.2025.08.002.

[5] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," *Artif. Intell. Rev.*, vol. 57, Mar. 2024, doi: 10.1007/s10462-023-10652-8.

[6] G. Hou, D. Tong, S. Liew, and P. Choo, "Comparative Analysis of Resampling Techniques for Class Imbalance in Financial Distress Prediction Using XGBoost," *Mathematics*, vol. 13, p. 2186, Jul. 2025, doi: 10.3390/math13132186.

[7] G. Petrides and W. Verbeke, "Cost-sensitive ensemble learning: a unifying framework," *Data Min. Knowl. Discov.*, vol. 36, Jan. 2022, doi: 10.1007/s10618-021-00790-4.

[8] L. Dube and T. Verster, "Interpretability of the random forest model under class imbalance," *Data Science in Finance and Economics*, vol. 4, pp. 446–468, Jan. 2024, doi: 10.3934/DSFE.2024019.

[9] K. El Madou, S. Marso, M. El Kharrim, and M. El Merouani, "Evolutions in machine learning technology for financial distress prediction: A comprehensive review and comparative analysis," *Expert Syst.*, vol. 41, Oct. 2023, doi: 10.1111/exsy.13485.

[10] X. Zhu, Q. Chu, X. Song, P. Hu, and L. Peng, "Explainable prediction of loan default based on machine learning models," *Data Science and Management*, vol. 6, no. 3, pp. 123–133, Sep. 2023, doi: 10.1016/j.dsm.2023.04.003.

[11] B. Amirshahi and S. Lahmiri, "Bankruptcy prediction using optimal ensemble models under balanced and imbalanced data," *Expert Syst.*, vol. 41, no. 8, p. e13599, Aug. 2024, doi: https://doi.org/10.1111/exsy.13599.

[12] M. A. Chohan, T. Li, S. Ramakrishnan, and M. Sheraz, "Artificial Intelligence in Financial Risk Early Warning Systems: A Bibliometric and Thematic Analysis of Emerging Trends and Insights." [Online]. Available: www.ijacsa.thesai.org

[13] K. Yang et al., "Hybrid Classifier Ensemble for Imbalanced Data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1387–1400, 2020, doi: 10.1109/TNNLS.2019.2920246.

[14] P. P. Singh, F. I. Anik, R. Senapati, A. Sinha, N. Sakib, and E. Hossain, "Investigating customer churn in banking: A machine learning approach and visualization app for data science and management," *Data Science and Management*, vol. 7, no. 1, pp. 7–16, Mar. 2024, doi: 10.1016/j.dsm.2023.09.002.

[15] M. Aly SharafEldin, A. M. Idrees, and S. Ouf, "A Proposed Framework for Loan Default Prediction Using Machine Learning Techniques." [Online]. Available: www.ijacsa.thesai.org

[16] C. Zhang, K. C. Tan, H. Li, and G. S. Hong, "A Cost-Sensitive Deep Belief Network for Imbalanced Classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 109–122, 2019, doi: 10.1109/TNNLS.2018.2832648.

[17] A. Michiels and V. Molly, "Financing Decisions in Family Businesses: A Review and Suggestions for Developing the Field," *Family Business Review*, vol. 30, pp. 369–399, Nov. 2017, doi: 10.1177/0894486517736958.

[18] P. Rau and T. Yu, "A survey on ESG: investors, institutions and firms," *China Finance Review International*, vol. 14, pp. 3–33, Jan. 2023, doi: 10.1108/CFRI-12-2022-0260.

[19] M. A. Umar, Z. Chen, K. Shuaib, and Y. Liu, "Effects of feature selection and normalization on network intrusion detection," *Data Science and Management*, vol. 8, no. 1, pp. 23–39, Mar. 2025, doi: 10.1016/j.dsm.2024.08.001.

[20] S. Nobili and M. L. Drudi, "A Liquidity Risk Early Warning Indicator for Italian Banks: A Machine Learning Approach," *SSRN Electronic Journal*, Jan. 2021, doi: 10.2139/ssrn.3891566.

[21] W. Qiu, S. Rudkin, and P. Dlotko, "Refining understanding of corporate failure through a topological data analysis mapping of Altman's Z-score model," *Expert Syst. Appl.*, vol. 156, p. 113475, 2020, doi: https://doi.org/10.1016/j.eswa.2020.113475.

- [22] M. Clintworth, D. Lyridis, and E. Boulougouris, "Financial risk assessment in shipping: a holistic machine learning based methodology," *Maritime Economics & Logistics*, vol. 25, Jan. 2021, doi: 10.1057/s41278-020-00183-2.
- [23] B. R. Gunnarsson, S. vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, "Deep learning for credit scoring: Do or don't?," *Eur. J. Oper. Res.*, vol. 295, no. 1, pp. 292–305, 2021, doi: <https://doi.org/10.1016/j.ejor.2021.03.006>.

APPENDIX A. REPRODUCIBILITY AND THRESHOLD-SENSITIVITY

B. A.1 Experimental Configuration

All experiments were conducted in Python 3.13 with scikit-learn 1.9.1, XGBoost 3.4.1, NumPy 2.5.3 and SciPy 1.18.1. The evaluation protocol was repeated stratified five-fold cross-validation with five repeats (25 folds in total) over the full set of 500 enterprises, so that every one of the 57 distressed cases appeared in a held-out fold five times. Feature selection, class-weight calibration, ensemble-weight selection, and decision-threshold selection were all performed strictly within the training partition of each fold using an inner three-fold cross-validation; no held-out fold influenced any modelling choice. Table A.1 records the fixed settings needed to reproduce the reported results.

TABLE A.1. FIXED EXPERIMENTAL SETTINGS

Component	Setting
Evaluation protocol	Repeated stratified K-fold, 5 splits x 5 repeats (25 folds)
Sample	Full dataset, n = 500 (57 distressed, 11.4%)
Random seed	42 (fold generation and all estimators)
Inner selection CV	3-fold stratified, within each training partition
Base learner 1	RandomForest: n_estimators=300, max_depth=15, class_weight='balanced'
Base learner 2	XGBoost: scale_pos_weight set to the training imbalance ratio
Meta-learner	Logistic regression (stacking)
Features retained	20, selected within training folds
Recall floor (main)	tau = 0.88
Threshold grid	scanned on training-fold predictions only
Ensemble-weight grid	0.50 to 0.90 in steps of 0.10 (S3/S4)
AUC-PR scorer	sklearn named scorer 'average_precision'
Software	Python 3.13; scikit-learn 1.9.1; XGBoost 3.4.1; NumPy 2.5.3; SciPy 1.18.1