

Beyond Accuracy: Leakage-Audited Machine Learning For Explainable Microcredit Risk Screening At Bank Mandiri

Puja Malika*

Information Systems Study Program, Universitas Bina Darma, Palembang, Indonesia

Ari Muzakir

Information Systems Study Program, Universitas Bina Darma, Palembang, Indonesia

Evi Yulianingsih

Information Systems Study Program, Universitas Bina Darma, Palembang, Indonesia

Maria Ulfa

Information Systems Study Program, Universitas Bina Darma, Palembang, Indonesia

Nia Oktaviani

Information Systems Study Program, Universitas Bina Darma, Palembang, Indonesia

*Corresponding author: Pujamulika7@gmail.com

Abstract

Accuracy can misrank models when at-risk borrowers are rare and missed risk is costlier than unnecessary review. This study evaluates a leakage-audited, explainable machine-learning framework for microcredit risk screening at Bank Mandiri and addresses prior work that rarely links temporal feature availability, class imbalance, and asymmetric error costs. A quantitative cross-sectional experiment used 1,718 anonymized active-facility records from four Palembang microbusiness units. An audit reduced 29 candidate fields to 13 operational predictors. Logistic Regression, Decision Tree, and Random Forest were evaluated through a stratified 80:20 split and training-only fivefold cross-validation. A tuned Decision Tree achieved the highest test accuracy (96.22%; 331/344) but detected 15 of 24 at-risk facilities. Random Forest at a training-selected 0.46 threshold achieved 95.35% accuracy and detected 18 of 24. The accuracy difference was not significant (exact McNemar $p = .607$). Cost analysis showed that thresholded Random Forest becomes preferable when a false negative costs more than twice a false positive. Account liquidity explained 88.72% of Decision Tree impurity reduction. The study contributes a transparent governance rule connecting leakage control, interpretability, and asymmetric screening costs while preserving human oversight.

846



Keywords: microcredit risk; data leakage; explainable machine learning; class imbalance; decision threshold

INTRODUCTION

Microcredit expands working-capital access for small enterprises, yet its operational economics create a difficult screening problem. Portfolios contain many relatively small facilities, monitoring capacity is finite, and deterioration is usually rare relative to current accounts. A decision-support model must therefore do more than classify most facilities correctly. It must identify the smaller at-risk group early enough for proportionate follow-up, avoid overwhelming staff with false alarms, and provide reasons that can be reviewed by officers, customers, auditors, and supervisors.

Machine-learning research has reported strong credit-risk performance from tree-based and ensemble methods. Logistic Regression remains a common benchmark, whereas Decision Tree and Random Forest can capture interactions and nonlinear boundaries that a single linear scorecard cannot represent (Dumitrescu et al., 2022; Noriega et al., 2023). Studies of small and micro enterprises likewise show that machine learning can improve risk differentiation when borrower and account information are available (Gu et al., 2024). In Indonesia, Wulansari and Purwitasari (2023) reported that Random Forest reached 88% accuracy on Kredit Usaha Rakyat data and identified payment-account availability and business sector as influential variables. These findings support data-driven screening, but they do not settle which model is operationally preferable.

Three methodological problems prevent a direct translation from benchmark accuracy to bank use. First, operational extracts often combine variables created before, during, and after the event being classified. Delinquency days, arrears, downgrade flags, or direct encodings of collectibility can allow a model to recover the target instead of learning an ex ante relationship. Such leakage produces impressive but unusable results and remains a major source of non-reproducibility in machine-learning science (Kapoor & Narayanan, 2023). Second, portfolio health creates class imbalance. In the present data, 93.0% of facilities were current and 7.0% were already non-current or at risk. A classifier that labels every facility current would consequently achieve 93.0% accuracy while detecting no risk. Third, model choice embeds an error-cost assumption. Accuracy weights an unnecessary review of a current facility and a missed at-risk facility equally, although their consequences are unlikely to be equal.



Governance intensifies these concerns. The Indonesian Financial Services Authority identifies risk management, prudence, accountability, and responsible development as minimum considerations for banking artificial intelligence (Otoritas Jasa Keuangan [OJK], 2025). In a high-stakes setting, interpretability is not cosmetic. A model must be contestable, monitored, and bounded by human authority. Rudin (2019) argues that inherently interpretable models should be preferred where they provide adequate performance, while recent credit-scoring research shows that interpretable trees can remain competitive with more complex methods (Tu & Wu, 2025). Threshold optimization offers another governance lever because it changes the decision boundary without retraining the underlying probability model (Kpatcha, 2025).

This study revises an accuracy-only comparison into a decision-oriented evaluation of microcredit risk screening. It uses internal, anonymized facility records from four Bank Mandiri microbusiness units in the Palembang District. The data are a single-date operational snapshot; accordingly, the task is explicitly defined as contemporaneous risk screening, not forecasting a future default. The at-risk class includes every collectibility category above current, including special-mention accounts, and should not be interpreted as the regulatory non-performing-loan definition.

The study addresses three questions. First, which leakage-audited model maximizes overall accuracy on a held-out test set? Second, does the preferred model change when missed at-risk facilities are assigned a higher cost than unnecessary reviews? Third, which operational variables account for the model's class separation? The novelty lies in connecting four elements that are usually reported separately: temporal feature governance, imbalanced-class metrics, an interpretable model, and a transparent error-cost switching rule. This connection reframes model selection from “which accuracy is highest?” to “which documented error profile fits the bank's monitoring objective?”

LITERATURE REVIEW

Machine Learning in Credit-Risk Assessment

Credit-risk modeling has shifted from exclusively statistical scorecards toward comparative machine-learning pipelines. Logistic Regression remains attractive because coefficients are compact and directional, but the model imposes a linear relationship between predictors and the log odds of the outcome.



Dumitrescu et al. (2022) demonstrated that adding decision-tree effects can improve conventional logistic credit scoring by representing nonlinear structure. Decision Tree implements such structure directly through recursive partitions, while Random Forest averages many trees trained on bootstrapped samples to reduce variance (Breiman, 2001).

Recent reviews find no universally dominant algorithm. Noriega et al. (2023) documented frequent use of ensemble methods, yet performance varied with sample size, feature design, validation protocol, and class distribution. Suhadolnik et al. (2023) similarly showed that machine-learning gains depend on empirical context rather than algorithm labels alone. Gu et al. (2024) reported strong results for small- and micro-enterprise credit assessment after imbalance processing, whereas Akinjole et al. (2024) found benefits from ensemble-based default prediction. These studies establish the promise of nonlinear modeling but also reveal a benchmark culture in which datasets and metrics vary enough to make headline scores difficult to compare.

For Indonesian microcredit, public evidence is thinner. Within IJEVSS, Nahriana (2024) used NPL ratios to assess cooperative health, while Tahu (2026) examined macroeconomic stability in relation to banking performance, credit risk, and resilience. These studies establish credit risk as an institutional-health issue, but aggregate ratios and macro-level relationships do not resolve facility-level screening, temporal leakage, or asymmetric classification errors. Wulansari and Purwitasari (2023) applied Random Forest to Kredit Usaha Rakyat status prediction and found that payment-account and sector information were important. The study is substantively close to the present setting, but many credit-risk studies still rely on public consumer datasets whose borrower profiles, product rules, and collection processes differ from Indonesian bank operations. Internal data improve contextual relevance but impose stronger requirements for privacy, temporal validity, and reproducibility.

Leakage, Validation, and Operational Availability

Data leakage occurs when model development uses information that would not be available at the intended decision time or when data from validation/test observations influence training. Kapoor and Narayanan (2023) show that leakage can invalidate performance claims even when the algorithm and metric are otherwise correct. In banking extracts, the risk is concrete: arrears, delinquency days, current collectibility codes, and downgrade indicators can encode the outcome almost directly. A model trained with such fields may classify the present status perfectly but cannot provide an independent early-warning signal.



Leakage control therefore begins before algorithm selection. Each candidate field must be assigned a time of availability and a permitted use. Training-only preprocessing is equally important: imputation, standardization, category encoding, hyperparameter selection, and resampling must be fitted without test data. Cross-validation used for model selection must remain nested within the training partition because using the same observations both to choose and evaluate a configuration introduces selection bias (Varma & Simon, 2006). Pipelines in scikit-learn support this separation by applying fitted transformations consistently (Pedregosa et al., 2011).

The present study extends leakage control with an operational-availability test. A field can precede the target yet still be unsuitable if it is a near-direct proxy, has uncertain semantics, is nearly constant, or expands dimensionality without sufficient minority observations. Conversely, contemporaneous account-balance information can be admissible for pre-due monitoring if officers observe it before the repayment decision point. This distinction is essential: the validity of a variable depends on the intended use time, not only its database timestamp.

Class Imbalance and Decision Thresholds

Class imbalance changes what common metrics mean. Branco et al. (2016) explain that overall accuracy can be dominated by a majority class and should be supplemented by class-sensitive measures. Recall measures the proportion of at-risk facilities detected; precision measures how many flagged facilities are actually at risk; the F1 score summarizes their harmonic balance; balanced accuracy averages sensitivity and specificity; ROC-AUC measures ranking across thresholds; and PR-AUC focuses more directly on positive-class retrieval under low prevalence.

Class weighting and synthetic oversampling are frequently proposed responses. SMOTE creates minority observations between neighboring minority points (Chawla et al., 2002), and imbalanced-learn implements resampling within reproducible pipelines (Lemaître et al., 2017). However, balancing is not automatically beneficial. It changes the effective training distribution and may increase false positives when deployment prevalence remains low. Full 1:1 oversampling can also be aggressive when the original minority share is approximately 7%.



Decision-threshold adjustment offers a different mechanism. Instead of altering training data, it moves the probability cutoff that maps scores into actions. Kpatcha (2025) shows that threshold optimization can explicitly manage accuracy and fairness trade-offs in probability-of-default modeling. The same principle applies to monitoring costs: a lower threshold generally detects more at-risk facilities but sends more current facilities to review. The suitable point depends on the relative cost of false negatives and false positives, not on a universal threshold of 0.50.

Interpretability, Governance, and Research Position

Credit screening is a high-stakes decision-support context. Opaque models can be difficult to validate, contest, and monitor even when they rank well. Rudin (2019) distinguishes inherently interpretable models from post-hoc explanations and argues for transparent models where the performance penalty is small. Tu and Wu (2025) demonstrate that interpretable classification trees can be designed specifically for credit scoring. Reviews published since then treat performance, fairness, and explainability as joint governance objectives rather than independent technical properties (Bahlool et al., 2026).

The literature leaves four linked gaps. Explicit temporal audits are still less common than generic feature selection; minority-class consequences are often hidden behind accuracy; threshold choice is seldom translated into a simple managerial rule; and studies using internal Indonesian microcredit data remain scarce. This study occupies that intersection. It does not claim a universally best algorithm. It determines when an accuracy-leading, interpretable tree should be preferred and when a thresholded ensemble's lower false-negative count justifies its additional false positives.

RESEARCH METHOD

Research Approach, Data Type, and Data Source

The research used a quantitative, retrospective, cross-sectional experimental design. The unit of analysis was an active microcredit facility, not an application decision or a unique person. The source was an authorized internal operational extract from four Bank Mandiri microbusiness units in the Palembang District dated 29 April 2026. Facilities had been originated between 2021 and 2026. The extract contained 1,718 rows and 60 original columns covering administration, identifiers, loan terms, account conditions, business categories, and credit quality.



The study used a census of all facilities present in the four-unit extract; no survey respondent or informant sampling was performed. The distribution was 811 facilities from Air Kumbang Banyuasin 1, 410 from Palembang Pusat Dagang 1, 271 from Palembang Pasar 16 Ilir 1, and 226 from Palembang Plaju 1. Direct identifiers, including customer name, account number, customer number, telephone number, and address, were removed before analysis. Results are reported only in aggregate.

Target Definition and Data Credibility

The target was constructed from the current collectibility field. Collectibility 1 was coded 0 (current); categories 2A, 2B, 2C, 3, 4, and 5 were coded 1 (at risk). This conservative definition supports early monitoring because it includes special-mention facilities rather than waiting until the regulatory non-performing threshold. The mapping was checked against observed delinquency-day ranges and the asset-quality categories in POJK No. 40/POJK.03/2019 (OJK, 2019). The final data contained 1,598 current facilities (93.0%) and 120 at-risk facilities (7.0%), a 13.3:1 imbalance ratio.

Data credibility was evaluated through four checks: target-category reconciliation; duplicate and type checks; missing-value profiling; and temporal feature audit. Numeric outliers were retained when values were plausible under product rules because deleting valid high-value facilities would narrow the population artificially. Missing account balance (18 records) and business sector (3 records) were handled inside the training pipeline. Zero values in effective interest rate represented structural non-availability rather than zero-cost credit; they were converted to missing, imputed from training data, and accompanied by an availability indicator.

Leakage Audit and Research Variables

Twenty-nine candidate operational fields were reviewed against the intended screening time. One field supplied the target. Thirteen fields were excluded because they encoded contemporaneous collectibility, status trajectories, observed delinquency, arrears, downgrade status, or a post-observation event. Three more were excluded because they were nearly constant, semantically uncertain, or excessively high-cardinality relative to the 96 at-risk training cases. Twelve source predictors remained, and one effective-rate availability indicator was derived, producing 13 final predictors: nine numeric and four categorical.



Table 1. Leakage and operational-availability audit

Audit group	Representative fields	Decision	Rationale
Target source	Current collectibility	Target only	Defines current versus at-risk status
Direct or post-event proxies	Numeric collectibility, downgrade status, delinquency days, principal/interest/penalty arrears, total arrears, first-payment-default status, last-payment date	Excluded	Available with or after deterioration; would recover the outcome rather than an independent signal
Status trajectory	Prior collectibility and prior-card status	Excluded conservatively	Strong near-target proxies; retained use would obscure whether other operational indicators add information
Origination terms	Credit limit, instalment, tenor, nominal/effective rate, product, due day, business sector	Included	Available from origination and interpretable operationally
Pre-due account conditions	Outstanding principal, account balance, sufficient-balance status, officer relationship status	Included	Observable for monitoring before the instalment decision point
Low-information or high-dimensional fields	Near-constant collection flag, uncertain due-date category, high-cardinality purpose code	Excluded	Weak variation, uncertain definition, or dimensionality disproportionate to minority cases

The numeric predictors were credit limit, outstanding principal, monthly instalment, tenor, nominal interest rate, effective interest rate, account balance, instalment due day, and effective-rate availability. The categorical predictors were product type, sufficient-balance status, officer relationship status, and business sector. One-hot encoding expanded the four categorical variables to 13 binary columns; together with the nine numeric columns, the model matrix had 22 columns.

Instruments, Processing, and Model Development

The research instruments were a data dictionary, the leakage-audit matrix, a reproducible Python analysis pipeline, and model-evaluation tables. Data processing used pandas and NumPy; modeling used scikit-learn. Numeric missing values were imputed with the training median and standardized. Categorical missing values were imputed with the training mode and one-hot encoded with unknown-category handling. Every fitted preprocessing parameter was learned from training observations only.

Data were split 80:20 with stratification and `random_state = 42`. The training partition contained 1,374 facilities (1,278 current and 96 at risk); the test partition contained 344 facilities (320 current and 24 at risk). Logistic Regression represented a linear benchmark, Decision Tree represented an inherently interpretable nonlinear model, and Random Forest represented a bagged ensemble. Baseline and class-weighted configurations were fitted for all three. Grid search used stratified fivefold cross-validation within the training partition. The Decision Tree search considered `max_depth` of 3–5, `min_samples_split` of 10–30, `min_samples_leaf` of 5–20, and optional balanced class weights. The Random Forest search considered 150 or 250 trees, depth of 3–5, leaf size of 5–20, and optional balanced class weights. The Logistic Regression search considered `C` values of 0.01–10 and optional balanced weights.

Additional Random Forest scenarios evaluated SMOTE and a training-derived decision threshold. SMOTE used five neighbors and was applied only after preprocessing within the training data; the held-out test distribution remained unchanged. Threshold selection used out-of-fold probabilities from the training partition. Several cutoffs shared the maximum cross-validated accuracy in the independent re-execution; 0.46 was retained because it was the thesis-prespecified value and remained one of the maxima on the 0.01 grid. No test label was used to select the threshold.



Figure 1 condenses the separation between training-only model development and locked held-out evaluation. This boundary is central to the study because it prevents preprocessing, resampling, hyperparameter search, and threshold choice from learning from the final test labels.

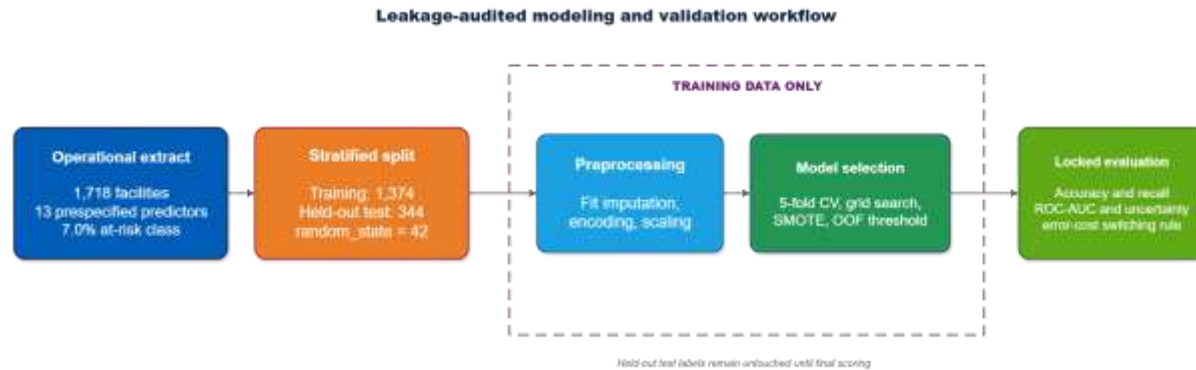


Figure 1. Leakage-audited modeling and validation workflow

Analysis Framework and Evaluation

Evaluation reported overall accuracy, balanced accuracy, at-risk precision, at-risk recall, F1, ROC-AUC, PR-AUC, and the confusion matrix. Wilson 95% confidence intervals were calculated for accuracy and at-risk recall. Because all models were tested on the same facilities, the accuracy difference between the two leading operational candidates was evaluated with an exact McNemar test.

The decision framework treated false-positive and false-negative costs symbolically rather than inventing monetary values. Let C_{FP} denote the cost of sending a current facility to unnecessary review and C_{FN} the cost of failing to flag an at-risk facility. If a model produces FP false positives and FN false negatives, its screening loss is:

$$L = FP \times C_{FP} + FN \times C_{FN}.$$

The accuracy-leading Decision Tree produced four false positives and nine false negatives, whereas Random Forest at threshold 0.46 produced ten false positives and six false negatives. Their losses are therefore $4C_{FP} + 9C_{FN}$ and $10C_{FP} + 6C_{FN}$. Solving the inequality between them yields a transparent switching rule: thresholded Random Forest has lower loss when $\frac{C_{FN}}{C_{FP}} > 2$; the models tie when the ratio equals 2; and Decision Tree is preferable when the ratio is below 2.

Credibility, Privacy, and Research Boundaries

Credibility was strengthened through a fixed seed, stratification, training-only transformations, training-only hyperparameter and threshold selection, a completely held-out test set, and an independent manuscript-stage re-execution. The re-execution reproduced the thesis's Decision Tree accuracy (96.22%), Random Forest threshold accuracy (95.35%), selected Decision Tree configuration, and feature-importance values. Access was institutionally authorized, and direct identifiers were removed. The data cannot be published because they remain subject to bank confidentiality and Indonesian personal-data obligations. The model is evaluated as decision support for staff prioritization, not as an autonomous approval, rejection, pricing, or enforcement mechanism.

RESULTS**Data Profile and Preprocessing Outcomes**

The target distribution confirmed severe imbalance: 1,598 current and 120 at-risk facilities. Stratification preserved the 7.0% minority share in both partitions, leaving only 24 at-risk facilities in the test set. This count matters for interpretation: one reclassified at-risk facility changes recall by 4.17 percentage points, and one additional correct prediction changes overall accuracy by 0.29 percentage points.

No facility was removed during preprocessing. Eighteen account-balance values were median-imputed, three sector values were mode-imputed, and 1,493 structural zero effective-rate values were handled with imputation plus an availability indicator. The final model matrix contained 22 columns and no missing values. The leakage audit removed outcome-bearing fields before any split or transformation, preventing the models from using observed arrears or collectibility proxies.

Model Performance

Table 2 reports the configurations needed to understand the model-selection conflict. Accuracy alone ranked the tuned Decision Tree first at 96.22%. It correctly classified 331 of 344 test facilities. The model was shallow (maximum depth five, minimum leaf size five, and minimum split size ten), and its fivefold training accuracy was 95.56%, close to the held-out value. This indicates controlled complexity rather than an unrestricted tree's memorization.



Table 2. Held-out performance of decision-relevant configurations

Model	Accuracy	Balanced accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression, baseline	0.9273	0.5563	0.4286	0.1250	0.1935	0.9253
Logistic Regression, class weighted	0.8314	0.9094	0.2927	1.0000	0.4528	0.9249
Decision Tree, baseline	0.9506	0.8385	0.6296	0.7083	0.6667	0.8385
Decision Tree, tuned	0.9622	0.8063	0.7895	0.6250	0.6977	0.9518
Random Forest, baseline	0.9477	0.7984	0.6250	0.6250	0.6250	0.9609
Random Forest, threshold 0.46	0.9535	0.8594	0.6429	0.7500	0.6923	0.9609

Note. Precision, recall, and F1 refer to the at-risk class. The threshold changes class decisions but not ROC-AUC because underlying Random Forest probabilities are unchanged.

Class weighting exposed the inadequacy of an accuracy-only narrative. Weighted Logistic Regression detected all 24 at-risk facilities but incorrectly flagged 58 current facilities; accuracy fell to 83.14% and precision to 29.27%. The result is not “worse” under every objective: it maximizes recall, but its review burden is large. Conversely, baseline Logistic Regression achieved 92.73% accuracy while detecting only three at-risk facilities. These two configurations demonstrate that accuracy and risk detection can move in opposite directions.

The tuned Decision Tree and thresholded Random Forest formed the practical frontier. Decision Tree had higher accuracy and precision, while thresholded Random Forest had higher balanced accuracy and recall. Their F1 scores were nearly equal (0.6977 versus 0.6923), and Random Forest had the higher ROC-AUC (0.9609 versus 0.9518). Figure 2 visualizes why no single “best” model exists without a declared objective.



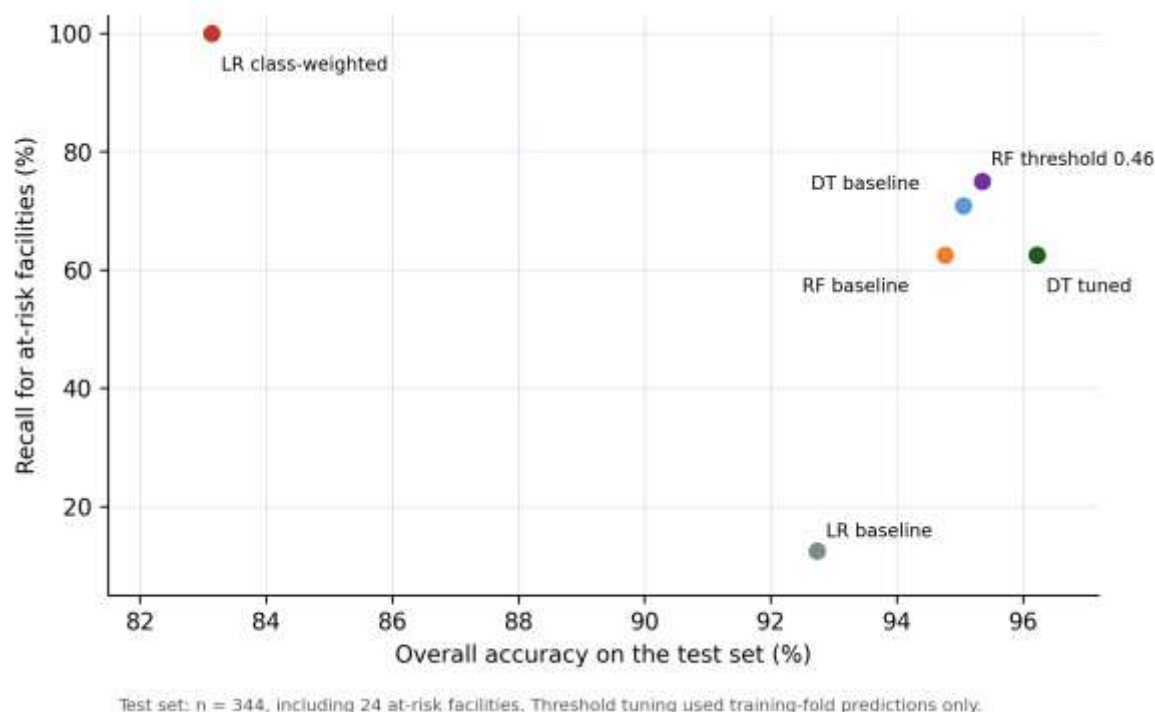


Figure 2. Overall accuracy and recall for selected model configurations

Error Profiles, Uncertainty, and Cost Switching

Table 3 isolates the two leading operational candidates. Decision Tree's 96.22% accuracy corresponds to four false positives and nine false negatives. Random Forest at 0.46 exchanged six additional false positives for three fewer false negatives. Thus, the 0.87-point accuracy advantage of Decision Tree represents three net correct classifications, not categorical superiority.

Table 3. Paired error profiles and uncertainty of the leading candidates

Measure	Tuned Decision Tree	Random Forest, threshold 0.46
Correct classifications	331/344	328/344
Accuracy (95% Wilson CI)	0.9622 (0.9364–0.9778)	0.9535 (0.9258–0.9712)

Measure	Tuned Decision Tree	Random Forest, threshold 0.46
At-risk facilities detected	15/24	18/24
Recall (95% Wilson CI)	0.6250 (0.4271–0.7884)	0.7500 (0.5510–0.8800)
False positives	4	10
False negatives	9	6
Symbolic error cost	$4C_{FP} + 9C_{FN}$	$10C_{FP} + 6C_{FN}$

Figure 3 makes the operational exchange explicit: the thresholded Random Forest detects three additional at-risk facilities but sends six additional current facilities to review.

Operational error trade-off on the held-out test set (n = 344)

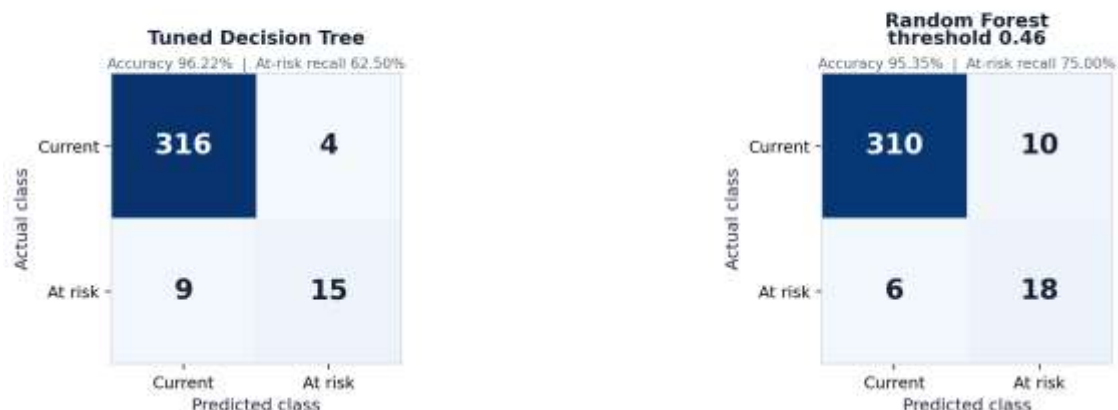


Figure 3. Confusion matrices of the two leading operational candidates

The accuracy intervals overlap substantially. In the paired comparison, Decision Tree alone was correct on nine discordant facilities and thresholded Random Forest alone was correct on six; the exact McNemar test was not significant ($p = .607$). The evidence therefore does not support a claim that the observed accuracy difference represents a stable population-level advantage.

The symbolic cost calculation produces a sharper conclusion. At equal error costs, Decision Tree is preferable because it makes fewer total mistakes. If one missed at-risk facility costs exactly twice one unnecessary review, both models have equal screening loss. For any ratio above two, thresholded Random Forest has lower loss. The data do not identify the bank's actual ratio; management must estimate it from review time, collection capacity, exposure, and expected loss. The contribution is the explicit boundary, not an unsupported monetary claim.

Interpretability and Influential Variables

The tuned Decision Tree formed a depth-five structure with 16 leaves. Account balance occupied the root and contributed 83.49% of total Gini-impurity reduction. Sufficient-balance status contributed another 5.23%; together, the two liquidity-related features accounted for 88.72%. Outstanding principal contributed 4.37%, agriculture sector 2.50%, plantation sector 2.05%, monthly instalment 1.62%, and service sector 0.75%. Figure 4 summarizes this concentration.

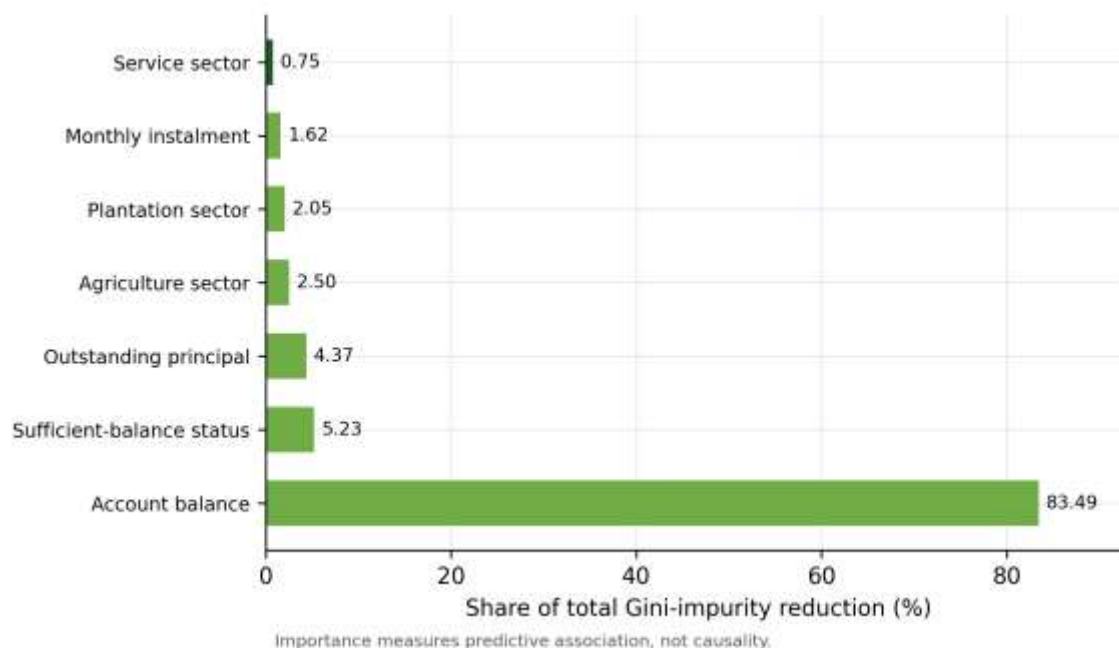


Figure 4. Gini-based feature importance of the tuned Decision Tree

The dominance of liquidity is consistent with descriptive patterns. At-risk facilities clustered near low account balances, while facilities with sufficient balances were rarely at risk at the snapshot date. Similar emphasis on payment-account information was reported in Indonesian Kredit Usaha Rakyat data (Wulansari & Purwitasari, 2023). Nevertheless, impurity importance is associative, can favor variables with more splitting opportunities, and does not establish that raising a balance would cause risk status to improve. It shows that current liquidity organizes the observed screening distinction.

DISCUSSION

Strength and Meaning of the Findings

The study is strongest where it moves beyond headline accuracy. The tuned Decision Tree achieved the best accuracy and precision, while thresholded Random Forest detected three additional at-risk facilities. Training-only threshold selection, a locked test set, confidence intervals, and an exact paired test make this comparison more credible than an unqualified leaderboard. The symbolic cost boundary adds practical value: model choice changes when one missed at-risk facility costs more than two unnecessary reviews.

The evidence supports a conditional decision, not a universal winner. Only 24 at-risk facilities were available in the test set, the recall intervals were wide, and the accuracy difference was not significant. Class weighting also achieved perfect recall by creating 58 false positives, confirming that imbalance correction can transfer error rather than remove it. This agrees with Branco et al. (2016) and Gu et al. (2024): class distribution, algorithm behavior, and operational objectives must be evaluated together.

Relationship to Previous Research and Novelty

Previous IJEVSS studies treated credit risk at broader levels. Nahriana (2024) used NPL ratios to assess cooperative health, whereas Tahu (2026) examined macroeconomic relationships with banking performance, NPL, and resilience. Wulansari and Purwitasari (2023) moved closer to facility-level Indonesian microcredit by applying Random Forest to Kredit Usaha Rakyat data. The present study extends this literature by auditing when predictors become available, testing minority-class errors, and translating competing error profiles into a transparent selection rule.



The study contributes a methodological audit and a managerial decision rule. It does not establish intrinsic algorithmic superiority. Threshold adjustment raised Random Forest recall from 62.5% to 75.0% without changing its ranking ability, consistent with Kpatcha's (2025) treatment of the decision boundary as a policy variable. The novelty lies in linking that boundary to leakage control, interpretable modeling, and an explicit cost ratio in an internal Indonesian microcredit setting.

Interpretability and Operational Implications

The shallow Decision Tree offers an operational strength: its 16 leaves can be inspected, challenged, and translated into officer-facing rules. This supports Rudin (2019) and Tu and Wu (2025), who favor inherently interpretable models when performance remains competitive in high-stakes decisions. Under constrained review capacity, the tree is defensible; under an early-warning objective with higher missed-risk costs, Random Forest at 0.46 is preferable.

Account balance and sufficient-balance status produced 88.72% of the tree's impurity reduction. This concentration supports short-horizon monitoring and exposes the model's main weakness. Liquidity can change quickly, provides limited lead time, and is associative rather than causal. Officers must treat the model as a triage tool and validate each flag. Governance teams should monitor data freshness, distribution shift, review burden, and customer treatment.

Critical Appraisal and Limitations

The authors strengthened the design through the leakage audit, training-only preprocessing, held-out evaluation, and explicit human-oversight boundary. These features make the results reproducible within the studied extract and prevent outcome-bearing fields from inflating performance. The study also avoids inventing monetary costs, which keeps the recommendation transparent and auditable.

Its generalizability remains limited. The data represent one date, four units in one district, and contemporaneous status rather than future default. The small positive test count limits statistical power; facility rows may cluster within borrowers; and no external, temporal, calibration, or demographic-fairness validation was available. The 0.46 threshold cannot serve as a permanent deployment setting. Prospective multi-period validation, probability calibration, fairness review, drift monitoring, and periodic re-estimation of prevalence and error costs are required before operational adoption.



CONCLUSION

This study shows that the model with the highest microcredit screening accuracy is not automatically the model with the most suitable risk-control profile. After auditing 29 candidate fields and retaining 13 operational predictors, a tuned, depth-five Decision Tree achieved the highest held-out accuracy at 96.22% and produced directly inspectable rules. Yet it detected 15 of 24 at-risk facilities. Random Forest at a training-selected threshold of 0.46 achieved slightly lower accuracy at 95.35% while detecting 18 of 24. The observed accuracy difference was not statistically significant.

The research question is therefore answered conditionally. Decision Tree is preferable when total mistakes, interpretability, and review efficiency dominate. Thresholded Random Forest is preferable when the cost of missing an at-risk facility exceeds twice the cost of an unnecessary review. Account balance and sufficient-balance status supplied 88.72% of the tree's impurity reduction, making current liquidity the central screening signal but not a causal explanation.

The practical impact is a transparent governance rule for allocating monitoring attention. The model should remain advisory, with officers retaining authority and validating each flagged case. Future research should use multi-period, multi-region data; model the time to deterioration; estimate actual asymmetric costs; evaluate probability calibration and fairness; and test performance prospectively before operational use.

REFERENCES

- Akinjole, A., Shobayo, O., Popoola, J., Okoyeigbo, O., & Ogunleye, B. (2024). Ensemble-based machine learning algorithm for loan default risk prediction. *Mathematics*, 12(21), 3423. <https://doi.org/10.3390/math12213423>
- Bahloul, R., Hewahi, N., & Elmedany, W. (2026). Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review. *Journal of Risk and Financial Management*, 19(2), 104. <https://doi.org/10.3390/jrfm19020104>
- Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2), Article 31. <https://doi.org/10.1145/2907070>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>



- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178–1192. <https://doi.org/10.1016/j.ejor.2021.06.053>
- Gu, Z., Lv, J., Wu, B., Hu, Z., & Yu, X. (2024). Credit risk assessment of small and micro enterprise based on machine learning. *Heliyon*, 10(5), e27096. <https://doi.org/10.1016/j.heliyon.2024.e27096>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- Kpatcha, E. (2025). Balancing fairness and accuracy in machine learning-based probability of default modeling via threshold optimization. *Journal of Risk and Financial Management*, 18(12), 724. <https://doi.org/10.3390/jrfm18120724>
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17), 1–5. <https://www.jmlr.org/papers/v18/16-365.html>
- Nahriana. (2024). Analysis of the health assessment level of cooperatives in view of NPL and investment levels in the MSME sector during the COVID-19 pandemic: Case study of the Berkah Savings and Loans Cooperative in Makassar City. *International Journal of Education, Vocational and Social Science*, 3(1), 47–58.
- Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine learning for credit risk prediction: A systematic literature review. *Data*, 8(11), 169. <https://doi.org/10.3390/data8110169>
- Otoritas Jasa Keuangan. (2019). Peraturan Otoritas Jasa Keuangan Nomor 40/POJK.03/2019 tentang penilaian kualitas aset bank umum. <https://www.ojk.go.id>
- Otoritas Jasa Keuangan. (2025). Tata kelola kecerdasan artifisial perbankan Indonesia. <https://www.ojk.go.id/id/Publikasi/Roadmap-dan-Pedoman/Perbankan/>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>



- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Suhadolnik, N., Ueyama, J., & da Silva, S. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management*, 16(12), 496. <https://doi.org/10.3390/jrfm16120496>
- Tahu, G. P. (2026). The influence of macroeconomic stability on financial performance, credit risk, and the resilience of the Indonesian banking sector. *International Journal of Education, Vocational and Social Science*, 5(3), 2985–3000.
- Tu, J., & Wu, Z. (2025). Inherently interpretable machine learning for credit scoring: Optimal classification tree with hyperplane splits. *European Journal of Operational Research*, 322(2), 647–664. <https://doi.org/10.1016/j.ejor.2024.10.046>
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, 91. <https://doi.org/10.1186/1471-2105-7-91>
- Wulansari, & Purwitasari, D. (2023). Algoritma Random Forest pada prediksi status Kredit Usaha Rakyat untuk mengurangi non-performing loan rate. *Journal of Intelligent System and Computation*, 5(2), 109–114. <https://doi.org/10.52985/insyst.v5i2.358>

