Equivalent: Jurnal Ilmiah Sosial Teknik
Volume 8, Issue 5, 2071-2083
e_ISSN: 2775-0833
https://jurnalequivalent.id/index.php/jequi/index
DOI: doi.org/10.59261/jequi.v8i5.467

Resampling And Cost-Sensitive Random Forest and Xgboost For Non-Performing Sharia Financing Prediction

Febriany Syafitri* Universitas Bina Nusantara, Indonesia

Tuga Mauritsius Universitas Bina Nusantara, Indonesia

***Corresponding author:**
Febriany Syafitri, Universitas Bina Nusantara, Indonesia. ✉ febriany.syafitri@binus.ac.id

Article Info:

Article history:
Received: September 15, 2026
Revised: October 20, 2026
Accepted: October 22, 2026
Available Online: October 31, 2026

Keywords:
Class Imbalance;
Cost-Sensitive Learning;
Ensemble Learning;
Machine Learning;
Non-Performing Financing;
Resampling;

This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author. Published by Politeknik Siber Cerdika International

Abstract

Background: Class imbalance presents a significant challenge in predicting non-performing financing (NPF) in Islamic financial institutions, potentially reducing the effectiveness of minority-class detection.
Objective: This study aims to compare six resampling techniques combined with Random Forest and XGBoost under standard and cost-sensitive learning to identify effective configurations for NPF prediction.
Methods: A quantitative experimental approach was applied to 8,814 historical Sharia financing records from 2016 to 2025. Six resampling techniques—SMOTE, ADASYN, Borderline-SMOTE, SMOTE-ENN, Distributed Gaussian (D-GA), and Distributed Poisson (D-PO)—were evaluated using stratified cross-validation. Model performance was assessed using the area under the precision–recall curve (AUPRC), macro F1 score, NPF precision, and NPF recall, followed by evaluation on a reported holdout dataset of 1,763 records.
Results: D-GA combined with cost-sensitive XGBoost achieved the highest cross-validation AUPRC (0.5503). On the holdout dataset, Borderline-SMOTE combined with standard XGBoost achieved the highest AUPRC (0.5942) and NPF precision (100.00%), although NPF recall remained 27.75%. SMOTE-ENN combined with cost-sensitive XGBoost achieved the highest NPF recall (43.54%), whereas SMOTE combined with standard Random Forest achieved the highest macro F1 score (0.7089).
Conclusion: Different configurations demonstrate distinct trade-offs between NPF detection and prediction reliability. The findings support objective-specific model selection, although inconsistent class distributions, unavailable baseline results, and data from a single institution limit generalizability.

To cite this article: Syafitri, F., & Mauritsius, T. (2026). Resampling and Cost-Sensitive Random Forest and XGBoost for Non-Performing Sharia Financing Prediction. *Equivalent: Jurnal Ilmiah Sosial Teknik,* 8(5), 2071-2083. https://doi.org/10.59261/jequi.v8i5.467

INTRODUCTION

Financing expands access to financial services and supports economic activity, while Islamic financial institutions must preserve portfolio quality as financing activity grows. Increasing financing volume enlarges the number and value of exposures requiring assessment and follow-up, even when the non-performing financing (NPF) ratio remains unchanged. Growth alone does not demonstrate worsening NPF; rather, it increases the scale of risk-monitoring demands. Historical repayment data may support more adaptive screening of nonlinear customer and financing patterns, but comparative evidence is needed before any advantage over conventional scoring can be established. The present study therefore examines the detection performance and false-alarm behavior of machine learning configurations rather than assuming that financing growth causes default. The Financial Services Authority of Indonesia (*Otoritas Jasa*

Keuangan [OJK]) reported financing of Rp 716.40 trillion in March 2026, representing a year-over-year increase of 9.82% (Otoritas Jasa Keuangan [OJK], 2026). These industry figures describe portfolio expansion; the study does not use them as evidence of increasing industry-wide NPF.

PT. XYZ is the anonymized provider of the historical consumer *Sharia* financing records analyzed in this study. Its internal product-level monitoring reported an NPF ratio of 13.2% in the third quarter of 2026. This portfolio indicator is separate from the 2016–2025 modeling dataset and does not represent the prevalence of the positive class in the modeling sample. Non-performing financing can increase impairment provisioning requirements and place pressure on financial performance (Shofiani, 2018). The company introduced a credit-scoring system in early 2026; however, neither its predictive performance metrics nor a matched before-and-after evaluation is available. The reported NPF ratio therefore motivates further evaluation but does not establish that the existing system is inaccurate or that machine learning will outperform it. The analysis concerns historical financing records rather than a verified count of unique customers. PT. XYZ remains pseudonymous, and no direct customer identifiers are disclosed. Access to the records is acknowledged in this manuscript, but the available study documentation does not establish the existence of a formal authorization instrument, a consent waiver, or ethics approval; none is claimed here.

Research on machine learning for default prediction includes systematic reviews and empirical comparisons of credit-risk models (Alvi et al., 2024; Bhandary & Ghosh, 2025; Zhang et al., 2025). Tree-based ensembles can model nonlinear interactions among predictors, whereas the effects of feature scaling depend on the algorithm and task (Pinheiro et al., 2025). In structured tabular data classification, Calvianto Lim (2026) and Liany & Oktavia (2024) demonstrated that ensemble tree-based models, particularly Random Forest, outperform single-tree baselines and XGBoost in predictive accuracy and stability when working with multi-attribute records. Similarly, Lianata et al. (2025) evaluated Random Forest and Gradient Boosting on imbalanced health datasets, highlighting that while hyperparameter tuning enhances overall precision and AUC, it may introduce a trade-off by slightly decreasing minority-class recall. Studies of sampling methods in diabetes classification and other health applications Hairani & Priyanto (2023) and Mishra et al. (2020), intrusion detection Amien et al. (2022), and regression Husain et al. (2025) address different outcomes and validation settings; their findings cannot be transferred directly to *Sharia* financing. Riantika et al. (2024) examined the effect of SMOTE-ENN on classification complexity. Borderline-SMOTE targets minority-class observations near class boundaries Han et al. (2005), whereas ADASYN assigns greater emphasis to difficult-to-learn minority-class regions. Cost-sensitive learning modifies the relative weights assigned to classification errors rather than altering the observed data distribution through resampling (Mienye & Sun, 2021). These methodological differences motivate a direct, application-specific comparison using both ranking and NPF detection metrics, rather than treating successful results in other sectors as proof of effectiveness in financial applications.

The research gap lies in the need for an application-specific comparison of resampling strategies and classifier behavior, rather than in an absence of prior credit-risk research. Xiao et al. (2021) compared combinations of resampling methods and classifiers for imbalanced credit scoring; their benchmark addresses model selection using credit data but does not establish performance for this institution's *Sharia* financing records. Zhao et al. (2024) examined resampling across credit-risk datasets and classifiers, demonstrating that effectiveness depends on dataset characteristics and the severity of class imbalance. Pan et al. (2020) proposed sampling approaches that combine SMOTE with Gaussian distributions for imbalanced classification; that methodological setting does not establish performance in *Sharia* financing. Collectively, these studies motivate a comparison of interpolation-based, boundary-focused, hybrid cleaning, and distribution-based generation methods within a single financing application. The present study applies six resampling strategies to Random Forest and XGBoost under standard and cost-sensitive learning, evaluating area under the precision–recall curve (AUPRC) alongside threshold-dependent NPF precision, recall, and Macro F1. Its contribution is an empirical evaluation using the reported *Sharia* financing records, without claiming to introduce a new resampling algorithm or establish universal superiority over previous approaches.

To address this application-specific gap, the study adopts a comparative experimental framework covering SMOTE, ADASYN, Borderline-SMOTE, SMOTE-ENN, Distributed Gaussian (D-GA), and Distributed Poisson (D-PO), with Random Forest and XGBoost evaluated under standard and cost-sensitive learning. These combinations yield 24 resampled configurations, while the no-resampling condition provides four additional baseline configurations. The reported comparison of tree-based models follows the no-scaling specification described in the Methods section. Threshold optimization is performed using out-of-fold predicted probabilities, and performance is compared using AUPRC, Macro F1, and class-specific precision and recall. The available results do not include measured training times, memory consumption, or baseline performance scores. Accordingly, computational efficiency and improvements over the no-resampling baseline have not been established. Accordingly, this study aims to (1) compare six resampling strategies for imbalanced Islamic financing classification; (2) examine Random Forest and XGBoost under standard and cost-sensitive learning across 24 resampled configurations; and (3) interpret the trade-offs among ranking performance, class balance, NPF detection, and false-positive control using AUPRC, Macro F1, NPF precision and recall, and confusion matrices. The four reported finalist configurations represent different evaluation priorities rather than an exhaustive ranking of all configurations. The practical objective of this study is to support early-risk screening and the prioritization of account reviews in Sharia financing. This objective aligns with the broader empirical evidence that machine learning frameworks can facilitate data-driven decision-making and improve resource allocation in operational risk surveillance (Mahendra & Oktavia, 2024). However, because complete baseline results and computational measurements are unavailable, this study does not quantify performance improvements over models trained without resampling or establish an operationally optimal model configuration. Because complete baseline results and computational measurements are unavailable, the study does not quantify improvements over models trained without resampling or establish an operationally optimal configuration.

METHOD

This quantitative study adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) to organize the analysis of PT XYZ's historical Islamic financing records from 2016 through 2025. The reported dataset contains 8,814 records and 19 original variables. Collectibility levels 1–2 were coded as performing financing (class 0), whereas levels 3–5 were coded as nonperforming financing (NPF; class 1). The preprocessing description reports 818 missing entries, which were treated using median imputation for numerical variables and mode imputation for categorical variables, followed by duplicate checking and removal. Identification and potential data leakage variables were excluded. Application dates were converted into year and month features, and ordinal, binary, and one-hot encoding were applied according to variable type. Numerical features were retained without scaling. The available documentation does not identify the variables containing missing entries, provide per-variable missing-value counts, or report the number of duplicates removed. The 8,814 records equal the reported 7,051 development records plus 1,763 holdout records; however, this arithmetic agreement does not establish an independently audited record count before and after cleaning. Following the data-partitioning protocol proposed by Mauritsius et al. (2025), synthetic resampling methods, including SMOTE, were applied exclusively to the training data within each cross-validation fold. Imputation and encoding parameters must be estimated using training data within each cross-validation fold to prevent data leakage, but the supplied documentation does not independently verify the order in which these operations were performed.

The reported data partition allocated 80% of the records to the development set (7,051 records) and 20% to an independent holdout set (1,763 records). Following the preprocessing and data-partitioning approach described by Ahmad et al. (2023), the dataset was divided into training and testing sets using an 80:20 ratio, while feature correlations were examined to identify and mitigate potential multicollinearity before model development. Stratified random sampling was the stated partitioning procedure. Seven conditions were considered: no resampling, Synthetic Minority Over-sampling Technique (SMOTE), SMOTE–Edited Nearest Neighbors (SMOTE-ENN), Adaptive Synthetic Sampling (ADASYN), Borderline-SMOTE, D-GA, and D-PO. Each

condition was paired with either standard or class-weighted Random Forest and either standard or cost-sensitive XGBoost. The methods specify that no feature scaling was applied in this comparison; the five scaling scenarios previously listed in the results are not supported by corresponding results and are therefore not treated as evaluated experiments. Resampling was reported to be restricted to the training portion of each cross-validation fold, with (k = 3), (5), and (10). The class totals displayed in Figures 1 and 2 are incompatible with a stratified split of a single, unchanged target cohort. Consequently, preservation of class prevalence between the development and holdout sets cannot be confirmed from the reported outputs.

For each model and cross-validation setting, predicted probabilities from the validation folds were combined to generate out-of-fold (OOF) predictions. Candidate decision thresholds within the reported interval of 0.01–0.99 were evaluated against the development-set labels, and the threshold that maximized the macro-averaged F1 score (Macro F1) was retained. Four selection criteria were considered: highest area under the precision–recall curve (AUPRC), NPF precision, NPF recall, and Macro F1. These criteria yielded four distinct configurations, as reported in Table 2. Each finalist was retrained on the complete development set and evaluated on the holdout set using the threshold selected during development; no test-set threshold tuning was reported. The selected cross-validation settings and decision thresholds were as follows: D-GA with cost-sensitive XGBoost ((k = 5), threshold = 0.84); Borderline-SMOTE with standard XGBoost ((k = 5), threshold = 0.79); SMOTE-ENN with cost-sensitive XGBoost ((k = 10), threshold = 0.89); and SMOTE with standard Random Forest ((k = 3), threshold = 0.41). The threshold increment, tie-breaking rule, random seeds, complete hyperparameter specifications, and generation parameters for D-GA and D-PO are not provided. Thus, although the stated selection rule is explicit, exact computational replication is not possible based on the available documentation. Furthermore, because the decision threshold was selected using OOF predictions, the resulting optimized OOF classification scores represent selection-stage estimates rather than a second, independent evaluation.

RESULTS AND DISCUSSION

Results

The Financing Dataset and Study Context

The empirical dataset comprises 8,814 historical Islamic financing records from PT. XYZ covering 2016–2025, with 19 original variables. Table 1 groups six demographic variables, six socioeconomic variables, four financing variables, and two application variables, together with the collectibility target. These groups account for 18 original predictors and one target variable; the number of encoded model inputs may differ after date transformation, one-hot encoding, and the exclusion of variables that may cause data leakage. Collectibility levels 1–2 denote performing financing, whereas levels 3–5 denote non-performing financing (NPF). The records describe the institution's consumer Sharia financing context. However, neither unique customer counts nor contract-level product details are provided, limiting the assessment of representativeness. The combination of numerical and categorical predictors makes neighborhood geometry and the plausibility of synthetic records relevant considerations in resampling. However, numerical distributions, outlier counts, and NPF-stratified associations are unavailable; therefore, no observed predictor effects can be inferred from the variable inventory alone.

Table 1
Dataset Characteristics

Variable Category	No. of Variables	Variables
Customer Demographics	6	Age, gender, marital status, education, province, and number of dependents
Socioeconomic Characteristics	6	Occupation, employment status, length of employment, salary, business turnover, and length of business operation

Variable Category	No. of Variables	Variables
Financing Characteristics	4	Financing tenor, product type, number of previous credits, and months since the last credit
Application Characteristics	2	Application period and source of funds
Target Variable	1	Collectability classification representing financing repayment performance

Preprocessing comprised missing-value treatment, duplicate checking, target coding, categorical encoding, date transformation, and data partitioning. The 818 reported missing entries were addressed through median imputation for numerical variables and mode imputation for categorical variables; however, variable-level missing-value counts and duplicate-removal totals are unavailable. Education was ordinally encoded, gender was binary encoded, and the remaining categorical predictors were one-hot encoded. The reported model comparison used unscaled numerical features, consistent with the described methodology; no comparison of scaling-specific performance is available. The stated partition comprised 7,051 development records and 1,763 hold-out records. Resampling was restricted to the training data, with the hold-out data reserved for final evaluation.

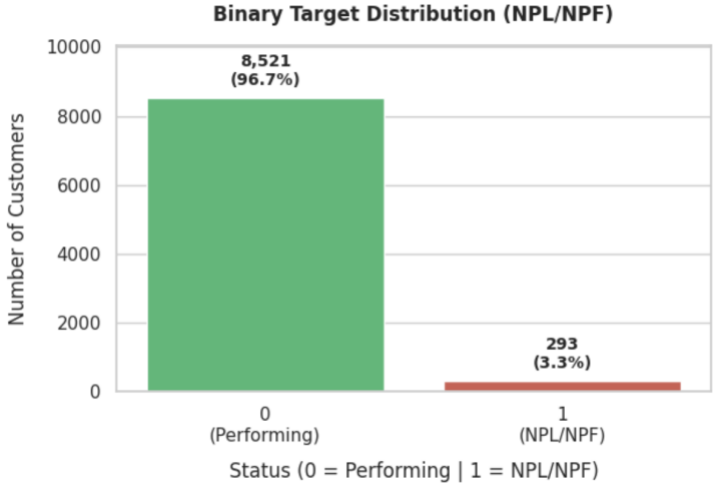


Figure 1. Bar Chart Class Imbalance

Figure 1 displays 8,521 performing records and 293 NPF records, totaling 8,814 records. These counts correspond to 96.68% and 3.32%, respectively, when rounded to two decimal places; the figure uses labels rounded to one decimal place. This imbalance can favor majority-class predictions, making accuracy alone inadequate for assessing NPF detection. Figure 2, however, contains 1,554 performing records and 209 NPF records in the test set, yielding an NPF prevalence of 11.85%, rather than approximately 3.32%. These displayed counts cannot both represent an unchanged target cohort under the stated stratified partition. Both original figures are retained, and no unobserved class counts are substituted. Hold-out findings are therefore interpreted conditionally on the displayed confusion matrices, with the discrepancy between the reported class distributions limiting comparability. The no-resampling condition was a planned comparator, but its model scores are absent from the supplied results. The confusion matrices describe the selected resampled configurations and cannot quantify improvement over the baseline; a majority-class-only reference is not equivalent to a fitted Random Forest or XGBoost baseline.

Model Development and Cross-Validation Results

Following the identification of substantial class imbalance, model development involved a systematic comparison of resampling strategies, classifier configurations, and cross-validation settings. Because conventional metrics may be dominated by the majority class, the area under the precision–recall curve (AUPRC) was used as a key evaluation metric because it focuses on the precision–recall relationship for the positive class. AUPRC provides a more informative assessment of minority-class identification under class imbalance. This metric is particularly relevant to the present study, in which the primary objective is to detect NPF observations that constitute a substantially smaller proportion of the data than the performing class (J. Hancock et al., 2022; J. T. Hancock et al., 2023).

The 7,051-record development set was evaluated using stratified k-fold cross-validation with k = 3, 5, and 10 (Abedin et al., 2026). The seven conditions comprised one no-resampling condition and six specified resampling methods. Resampling was restricted to the training folds, while the validation folds retained their original observed labels (Afane & Zhao, 2024; Toleva, 2026). Each condition was paired with either standard or class-weighted Random Forest and either standard or cost-sensitive XGBoost, with the latter using the class-imbalance ratio. Validation-fold probabilities were used to generate out-of-fold (OOF) predictions for AUPRC calculation and threshold selection based on Macro F1. NPF Precision and Recall were assessed at each selected threshold. Complete baseline scores are unavailable; therefore, Table 2 reports the four selected resampled configurations.

This procedure enabled the evaluation of minority-class ranking performance and threshold-dependent classification performance. Table 2 presents the four configurations selected according to the predefined criteria of highest AUPRC, NPF Precision, NPF Recall, and Macro F1. These configurations represent different performance priorities and were subsequently evaluated on the reported hold-out set.

Table 2
Performance of the Selected Model Configuration in Cross-Validation

k_fold	Resampling Method	Model Variant	Val AUPRC	Opt Thresh	Val Macro F1	Val Prec (NPF)	Val Rec (NPF)
5	D-GA	XGB Cost-Sensitive	0.5503	0.84	0.7022	0.8142	0.3142
5	Borderline-SMOTE	XGB Standard	0.5397	0.79	0.6980	0.9677	0.2867
10	SMOTE-ENN	XGB Cost-Sensitive	0.5301	0.89	0.6693	0.3876	0.4719
3	SMOTE	RF Standard	0.5385	0.41	0.7071	0.7588	0.3345

Table 2 reports four distinct selected configurations. During cross-validation, D-GA combined with cost-sensitive XGBoost achieved the highest reported AUPRC (0.5503), with k = 5, an optimal threshold of 0.84, a Macro F1 score of 0.7022, NPF Precision of 0.8142, and NPF Recall of 0.3142. This configuration represents the reported OOF AUPRC winner; it does not represent the hold-out AUPRC winner. In the hold-out results, Borderline-SMOTE combined with standard XGBoost instead achieved the highest reported AUPRC (0.5942). SMOTE combined with standard Random Forest achieved the highest reported Macro F1 score at both evaluation stages: 0.7071 during cross-validation and 0.7089 on the hold-out set. These stage-specific results are consistent with the abstract and should not be interchanged.

Borderline-SMOTE combined with standard XGBoost achieved the highest NPF Precision (0.9677), with an optimal threshold of 0.79, an AUPRC of 0.5397, a Macro F1 score of 0.6980, and NPF Recall of 0.2867. This pattern indicates that the model was highly selective when predicting NPF, producing highly reliable positive predictions while leaving many actual NPF cases unidentified. In contrast, SMOTE-ENN combined with cost-sensitive XGBoost achieved the highest NPF Recall (0.4719), with an optimal threshold of 0.89, an AUPRC of 0.5301, a Macro F1 score of 0.6693, and NPF Precision of 0.3876. The higher Recall indicates greater sensitivity to the minority class, although the lower Precision reflects the trade-off associated with identifying more NPF cases.

SMOTE combined with standard Random Forest achieved the highest Macro F1 score (0.7071), with an optimal threshold of 0.41, an AUPRC of 0.5385, NPF Precision of 0.7588, and NPF Recall of 0.3345. This result indicates more balanced performance between the performing and NPF classes than that of configurations that primarily emphasized either Precision or Recall. Overall, these findings demonstrate that resampling and classifier selection influence model prediction behavior under class imbalance. The results reveal distinct strengths in probability ranking, NPF Precision, NPF Recall, and overall class balance. Accordingly, the four configurations were retained as representative candidates based on their respective strengths for subsequent evaluation.

Holdout Test and Model Generalization

Four selection criteria produced four unique configurations: D-GA with cost-sensitive XGBoost for OOF AUPRC, Borderline-SMOTE with standard XGBoost for NPF Precision, SMOTE-ENN with cost-sensitive XGBoost for NPF Recall, and SMOTE with standard Random Forest for Macro F1. Table 3 lists these selected configurations rather than presenting a separate set of performance scores. Each configuration was evaluated on the reported 1,763-record hold-out set, which was excluded from model fitting, resampling, candidate selection, and threshold optimization. The discrepancy in class counts described above limits confirmation that this test set belongs to the same stratified cohort represented in Figure 1.

Table 3

Configurations Selected for Hold-Out Evaluation by Cross-Validation Criterion

Best Model Criteria	Resampling	Model
Highest AUPRC	D-GA	XGBoost Cost-Sensitive
Highest NPF Precision	Borderline-SMOTE	XGBoost Standard
Highest NPF Recall	SMOTE-ENN	XGBoost Cost-Sensitive
Highest Macro F1	SMOTE	Random Forest Standard

The hold-out evaluation revealed distinct predictive behaviors across the four selected configurations. D-GA combined with cost-sensitive XGBoost achieved an AUPRC of 0.5587, a Macro F1 score of 0.6954, NPF Precision of 88.41%, and NPF Recall of 29.19%. Borderline-SMOTE combined with standard XGBoost achieved the highest AUPRC (0.5942) and NPF Precision (100.00%), but its Recall was only 27.75%, indicating a highly selective prediction pattern. Borderline-SMOTE focuses synthetic sample generation on minority-class observations near the class boundary, which can strengthen learning in difficult regions but may not improve coverage across the entire minority class (Yang et al., 2022). Thus, its perfect Precision should be interpreted alongside its low Recall: all predicted NPF cases were correct, but many actual NPF cases remained undetected.

SMOTE-ENN combined with cost-sensitive XGBoost achieved the highest NPF Recall (43.54%), with an AUPRC of 0.5344, a Macro F1 score of 0.6770, and NPF Precision of 42.72%. SMOTE-ENN combines minority-class oversampling with Edited Nearest Neighbors (ENN) to remove potentially ambiguous observations, while cost-sensitive learning assigns greater importance to minority-class misclassification (Ling & Sheng, 2008; Mapfumo & Shongwe, 2026). These mechanisms are consistent with the model's greater sensitivity to NPF, although the higher Recall was accompanied by lower Precision and, consequently, more false-positive

predictions. Meanwhile, SMOTE combined with standard Random Forest achieved the highest Macro F1 score (0.7089), with an AUPRC of 0.5669, NPF Precision of 84.81%, and NPF Recall of 32.06%, indicating the most balanced performance across the two classes among the four reported finalists. However, because complete no-resampling baseline scores are unavailable, these results cannot be interpreted as evidence of improvement over the unresampled Random Forest model.

Compared with cross-validation, the selected configurations generally retained their characteristic prediction patterns on the hold-out data, although their absolute performance changed. For example, Borderline-SMOTE combined with standard XGBoost increased its AUPRC from 0.5397 to 0.5942 while maintaining high Precision and low Recall. AUPRC evaluates the quality of positive-class ranking across classification thresholds and should therefore be interpreted jointly with Precision and Recall, particularly under severe class imbalance. Overall, no configuration was uniformly superior: Borderline-SMOTE combined with standard XGBoost performed best in terms of AUPRC and Precision, SMOTE-ENN combined with cost-sensitive XGBoost performed best in terms of NPF Recall, and SMOTE combined with standard Random Forest performed best in terms of Macro F1. These results demonstrate distinct trade-offs among NPF detection, prediction reliability, and overall class balance, which are further examined through the confusion matrices in Figure 2.

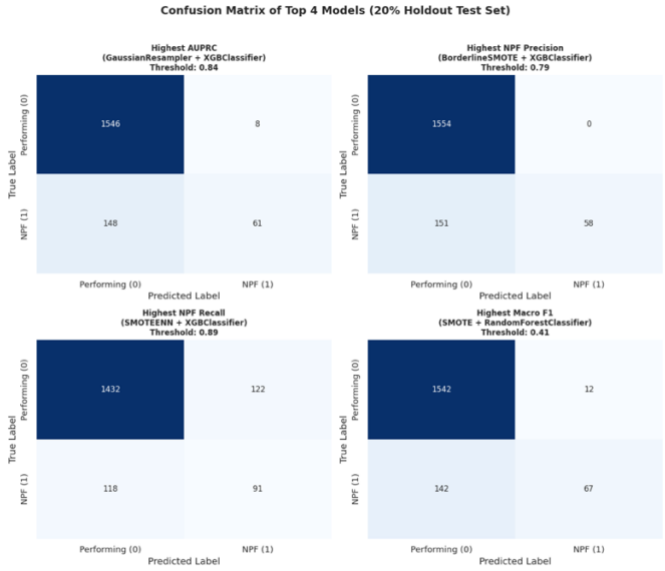


Figure 2. Confusion Matrices of the Four Selected Configurations

Based on Figure 2, D-GA combined with cost-sensitive XGBoost produced TN = 1,546, FP = 8, FN = 148, and TP = 61. Borderline-SMOTE combined with standard XGBoost produced TN = 1,554, FP = 0, FN = 151, and TP = 58. SMOTE-ENN combined with cost-sensitive XGBoost produced TN = 1,432, FP = 122, FN = 118, and TP = 91. SMOTE combined with standard Random Forest produced TN = 1,542, FP = 12, FN = 142, and TP = 67. All four matrices total 1,763 records and reproduce the reported hold-out Precision, Recall, and Macro F1 values. The corresponding false-negative rates are 70.81%, 72.25%, 56.46%, and 67.94%, respectively; the corresponding specificities are 99.49%, 100.00%, 92.15%, and 99.23%. These values are arithmetic derivations from the displayed counts rather than additional observations.

Discussion

Research Gap and Contribution

The empirical contribution of this study is a common comparison of resampling methods and tree-based ensemble configurations in the reported Sharia financing application. Xiao et al. (2021) examined combinations of resampling methods and classifiers for imbalanced credit scoring, while Zhao et al. (2024) emphasized that sampling effectiveness varies across datasets and classifiers. Lestari et al. (2025) further found that tree-based ensemble methods, particularly

XGBoost and Random Forest, can model complex nonlinear relationships and temporal patterns in structured tabular data, achieving better performance than conventional linear regression models. The different winning configurations observed here are consistent with this context dependence: D-GA achieved the highest OOF AUPRC, whereas Borderline-SMOTE achieved the highest hold-out AUPRC and Precision, and SMOTE-ENN combined with cost-sensitive XGBoost detected more NPF cases. Y. Xie et al. (2022) provide methodological context for Gaussian-based oversampling; however, their findings do not validate the specific D-GA or D-PO implementation used in this study. . Albright Wiladi & Mauritsius (2026) further demonstrated that ensemble learning methods, including Gradient Boosting and Random Forest, can identify complex relationships among features in business datasets while supporting model interpretability through SHAP-based feature importance analysis. The comparison is empirical rather than a contribution of a new algorithmic or conceptual framework. Because complete configuration results, synthetic-data generation parameters, and a reconciled cohort description are unavailable, the study cannot establish a causal advantage for any resampling mechanism or demonstrate general superiority across Sharia financing institutions.

Resampling Behavior

The observed performance profiles are consistent with differences in how sampling methods reshape minority-class neighborhoods. Borderline-SMOTE concentrates synthetic sample generation near class boundaries, whereas SMOTE-ENN combines interpolation with the removal of locally inconsistent observations. Gaussian-based approaches, by contrast, model a probability distribution for sample generation (W. Xie & Huang, 2025; Y. Xie et al., 2022). Given the mixture of demographic, income, and financing predictors and the small NPF group, boundary-focused sampling may produce selective predictions, whereas data cleaning and cost weighting may broaden minority-class detection. These explanations are plausible interpretations rather than measured feature-level explanations. The hold-out confusion matrices illustrate the resulting trade-off: Borderline-SMOTE produced no false positives but 151 false negatives, whereas SMOTE-ENN produced 122 false positives and 118 false negatives. Without feature distributions, outlier diagnostics, or matched ablation results, the performance differences cannot be attributed to any particular predictor, synthetic-data generation mechanism, or cost weight in isolation.

Cost-Sensitive Behavior

Cost-sensitive learning addresses class imbalance by assigning different importance to classification errors, thereby creating a trade-off between false negatives and false positives. Its effectiveness depends on the dataset, classification strategy, and the consequences of, or uncertainty associated with, misclassification costs (Araf et al., 2024; Komisarenko & Kull, 2025; Pes & Lai, 2021). In this study, the effect was particularly evident in the SMOTE-ENN and cost-sensitive XGBoost configuration, which achieved the highest NPF Recall (43.54%) but lower Precision (42.72%). This suggests that, when combined with SMOTE-ENN, cost-sensitive learning shifted the model toward broader minority-class recognition while increasing false-positive predictions. Such a trade-off is relevant to credit-risk prediction, where greater detection of problematic accounts may require accepting additional false alarms. Cost sensitivity should therefore be viewed as a mechanism for adjusting error preferences rather than as a guarantee of uniformly better performance.

Evaluation Metrics and Threshold Selection

AUPRC evaluates positive-class ranking, whereas Precision, Recall, and Macro F1 depend on the selected decision threshold. The thresholds reported in Table 2 range from 0.41 to 0.89, indicating that the selected candidates use different classification rules. Borderline-SMOTE combined with standard XGBoost achieved 100% Precision on the displayed hold-out confusion matrix but missed 151 of the 209 NPF records; this represents selective detection rather than perfect performance. SMOTE-ENN combined with cost-sensitive XGBoost reduced the number of missed NPF records to 118 while generating 122 false-positive predictions. Credit-risk comparisons by Zhao et al. (2024) support the joint evaluation of these competing objectives. A

false negative can delay account review, collection action, or provisioning assessment, whereas a false positive can consume review resources or subject a performing customer to unnecessary scrutiny. Maximizing Macro F1 weights the class-specific F1 scores equally and does not account for the institution's financial losses or its minimum acceptable NPF Recall. Operational use therefore requires thresholds to be selected on development data according to an explicit Recall target, review capacity, or validated error costs, followed by separate validation. Monetary benefits are not estimated in this study.

Generalization to Independent Data

The cross-validation and hold-out results show different metric rankings: the reported AUPRC winner changes, whereas SMOTE combined with standard Random Forest retains the highest Macro F1 among the four finalists. These observations describe the supplied model outputs but do not establish stable generalization to a new institution or time period. The hold-out set originates from the same reported source and does not constitute external validation. Moreover, the discrepancy in NPF prevalence between Figures 1 and 2 prevents the two evaluation stages from being treated as a fully reconciled assessment of the same cohort. Robust claims about generalization require consistent class definitions, documented preprocessing, complete baseline scores, and validation using later periods and data from other institutions.

CONCLUSION

The reported comparison indicates that resampling methods and standard versus cost-sensitive ensemble configurations produce different probability-ranking, detection, and selectivity profiles in Sharia financing prediction. Among the four finalists, D-GA combined with cost-sensitive XGBoost achieved the highest OOF AUPRC; Borderline-SMOTE combined with standard XGBoost achieved the highest hold-out AUPRC and Precision; SMOTE-ENN combined with cost-sensitive XGBoost achieved the highest NPF Recall; and SMOTE combined with standard Random Forest achieved the highest Macro F1 at both evaluation stages. These findings address the comparative objectives by demonstrating that the preferred configuration depends on the risk-screening objective. High Precision alone does not ensure adequate detection, and Macro F1 optimization does not establish the most appropriate operational threshold. The contribution is a context-specific empirical comparison limited by data from a single institution, incomplete baseline and computational-cost reporting, missing preprocessing and synthetic-data generation details, and inconsistent class prevalence between the original distribution figure and the hold-out confusion matrices. No comparative superiority over the unresampled baseline or readiness for deployment has been established. Future research should reconcile the cohort and target definitions, report complete baseline and computational-timing results, and validate the models using later financing periods and external institutions, with Recall or cost constraints specified during development.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to PT. XYZ for providing access to the historical Islamic financing dataset used in this research. The authors also thank University of Bina Nusantara for its academic support throughout the completion of this study.

AUTHOR CONTRIBUTION STATEMENT

Febriany Syafitri conceptualized the study, developed the research framework, designed the methodology, conducted the data analysis, interpreted the findings, and drafted the original manuscript. Tuga Mauritsius contributed to the literature review, validated the research design, provided critical revisions to the manuscript, and approved the final version for publication. Both authors reviewed and approved the final manuscript.

REFERENCES

Abedin, T., Xu, H., & Uddin, S. (2026). The impact of K selection in K-fold cross-validation on bias and variance in supervised learning models. *Scientific Reports*, *16*(1), 6084. https://doi.org/10.1038/s41598-026-37247-x

Afane, K., & Zhao, Y. (2024). Selecting Classifiers and Resampling Techniques for Imbalanced Datasets: A New Perspective. *Procedia Computer Science*, *246*, 1150–1159. https://doi.org/10.1016/j.procs.2024.09.539

Ahmad, M., Prabowo, H., Warnars, H. L. H. S., & Gaol, F. L. (2023). A Machine Learning Approach for Model Selection of Social Aid Beneficiaries. *Journal of System and Management Sciences*, *13*(6), 230–243. https://doi.org/10.33168/JSMS.2023.0614

Albright Wiladi, J., & Mauritsius, T. (2026). INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Forecasting the Demand of a Specific Water Pump Product with a 3-day Lead Time. *JOIV : Int. J. Inform. Visualization, 10*. www.joiv.org/index.php/joiv

Alvi, J., Arif, I., & Nizam, K. (2024). Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management. *Heliyon*, *10*(21), e39770. https://doi.org/10.1016/j.heliyon.2024.e39770

Amien, J. Al, Yoze Rizki, & Mukhlis Ali Rahman Nasution. (2022). Implementasi Adasyn Untuk Imbalance Data Pada Dataset UNSW-NB15 Adasyn Implementation For Data Imbalance on UNSW-NB15 Dataset. *Jurnal CoSciTech (Computer Science and Information Technology)*, *3*(3), 242–248. https://doi.org/10.37859/coscitech.v3i3.4339

Araf, I., Idri, A., & Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: a review. *Artificial Intelligence Review*, *57*(4), 80. https://doi.org/10.1007/s10462-023-10652-8

Bhandary, R., & Ghosh, B. K. (2025). Credit Card Default Prediction: An Empirical Analysis on Predictive Performance Using Statistical and Machine Learning Methods. *Journal of Risk and Financial Management*, *18*(1), 23. https://doi.org/10.3390/jrfm18010023

Calvianto Lim, R. (2026). Comparative Analysis of Decision Tree, Random Forest, and XGBoost for Student Category Prediction. *International Journal of Computer Science and Humanitarian AI*, *3*(1), 9–19. https://doi.org/10.21512/ijcshaijournal.v3i1.14936

Hairani, H., & Priyanto, D. (2023). A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data. *International Journal of Advanced Computer Science and Applications*, *14*(8). https://doi.org/10.14569/IJACSA.2023.0140864

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning* (pp. 878–887). https://doi.org/10.1007/11538059_91

Hancock, J., Khoshgoftaar, T. M., & Johnson, J. M. (2022). Informative Evaluation Metrics for Highly Imbalanced Big Data Classification. *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1419–1426. https://doi.org/10.1109/ICMLA55696.2022.00224

Hancock, J. T., Khoshgoftaar, T. M., & Johnson, J. M. (2023). Evaluating classifier performance with highly imbalanced Big Data. *Journal of Big Data*, *10*(1), 42. https://doi.org/10.1186/s40537-023-00724-5

Husain, G., Nasef, D., Jose, R., Mayer, J., Bekbolatova, M., Devine, T., & Toma, M. (2025). SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models. *Algorithms*, *18*(1), 37. https://doi.org/10.3390/a18010037

Komisarenko, V., & Kull, M. (2025). Cost-sensitive classification with cost uncertainty: do we need surrogate losses? *Machine Learning*, *114*(6), 132. https://doi.org/10.1007/s10994-024-06634-8

Lestari, D. R., Bangun, E. A. S., Gaol, F. L., & Matsuo, T. (2025). Machine Learning-Based Forecasting of Agricultural Commodity Prices Using Ensemble Models. *Journal of Human, Earth, and Future*, *6*(4), 887–899. https://doi.org/10.28991/HEF-2025-06-04-09

Lianata, I. W., Nicholas, K., Nugroho, D., Nathanael, Y., Christopher, N., & Irwansyah, E. (2025). Implementation of Random Forest Algorithm in Handling Imbalanced Data: A Study on

Default Models and Hyperparameter Tuning. *International Journal of Computer Science and Humanitarian AI*, *2*(2), 69–74. https://doi.org/10.21512/ijcshaijournal.v2i2.14417

Liany, D., & Oktavia, T. (2024). Comparative Evaluation of Data Mining Classification Techniques for Assistant Recruitment in a Computer Laboratory. *Journal of Logistics, Informatics and Service Science*, *11*(6), 184–197. https://doi.org/10.33168/JLISS.2024.0611

Ling, C. X., & Sheng, V. S. (2008). Cost-Sensitive Learning and the Class Imbalance Problem Motivation and Background. *Encyclopedia of Machine Learning*.

Mahendra, K., & Oktavia, T. (2024). DATA-DRIVEN INSIGHTS INTO FISHING PATTERNS USING VMS AND MACHINE LARNING. *Journal of Theoretical and Applied Information Technology*, *102*(22). www.jatit.org

Mapfumo, I., & Shongwe, T. (2026). Performance Evaluation of Machine Learning and Deep Learning Models for Credit Risk Prediction. *Journal of Risk and Financial Management*, *19*(3). https://doi.org/10.3390/jrfm19030210

Mauritsius, T., Haris, W., Atmaja, K., Isa, M., Bin Man, M., Oktavia, T., Retnowardhani, A., Komsiyah, S., Atallah, B., & Parikesit, B. (2025). An analytics framework for predicting electric vehicle charging station demand using spatial and interpretive methods. *Decision Analytics Journal*, *17*, 100648. https://doi.org/10.5281/zenodo.1

Mienye, I. D., & Sun, Y. (2021). Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, *25*. https://doi.org/10.1016/j.imu.2021.100690

Mishra, S., Mallick, P. K., Jena, L., & Chae, G.-S. (2020). Optimization of Skewed Data Using Sampling-Based Preprocessing Approach. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.00274

Pan, T., Zhao, J., Wu, W., & Yang, J. (2020). Learning imbalanced datasets based on SMOTE and Gaussian distribution. *Information Sciences*, *512*. https://doi.org/10.1016/j.ins.2019.10.048

Pes, B., & Lai, G. (2021). Cost-sensitive learning strategies for high-dimensional and imbalanced data: a comparative study. *PeerJ Computer Science*, *7*, e832. https://doi.org/10.7717/peerj-cs.832

Pinheiro, J. M. H., Oliveira, S. V. B. de, Silva, T. H. S., Saraiva, P. A. R., Souza, E. F. de, Godoy, R. V., Ambrosio, L. A., & Becker, M. (2025). The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks. *IEEE Access*, *13*. https://doi.org/10.1109/ACCESS.2025.3635541

Riantika, I., Sartono, B., & Anwar Notodiputro, K. (2024). Effectiveness of SMOTE-ENN to Reduce Complexity in Classification Model. *Indonesian Journal of Statistics and Its Applications*, *8*(1). https://doi.org/10.29244/ijsa.v8i1p70-82

Shofiani, P. (2018). Pengaruh Pembiayaan, Non Performing Finance (NPF) dan Dewan Pengawas Syariah Terhadap Manajemen Laba Menggunakan Cadangan Kerugian Penurunan Nilai (CKPN) Perbankan Syariah di Indonesia. *CIMAE: Conference on Islamic Management Accounting and Economics*, *1*.

Toleva, B. (2026). Is There a Preferred Cross-Validation Type for Train/Test Split in Imbalanced Classification. *Applied Computational Intelligence and Soft Computing*, *2026*(1). https://doi.org/10.1155/acis/6832228

Xiao, J., Wang, Y., Chen, J., Xie, L., & Huang, J. (2021). Impact of resampling methods and classification models on the imbalanced credit scoring problems. *Information Sciences*, *569*, 508–526. https://doi.org/10.1016/j.ins.2021.05.029

Xie, W., & Huang, X. (2025). Oversampling Algorithm Based on Improved K-Means and Gaussian Distribution. *Information*, *17*(1), 28. https://doi.org/10.3390/info17010028

Xie, Y., Qiu, M., Zhang, H., Peng, L., & Chen, Z. (2022). Gaussian Distribution Based Oversampling for Imbalanced Data Classification. *IEEE Transactions on Knowledge and Data Engineering*, *34*(2), 667–679. https://doi.org/10.1109/TKDE.2020.2985965

Yang, C., Dong, Y., Lu, J., & Peng, Z. (2022). Solving Imbalanced Data in Credit Risk Prediction: A Comparison of Resampling Strategies for Different Machine Learning Classification Algorithms, Taking Threshold Tuning into Account. *2022 5th International Conference on Machine Learning and Machine Intelligence*, 30–40. https://doi.org/10.1145/3568199.3568204

Zhang, X., Zhang, T., Hou, L., Liu, X., Guo, Z., Tian, Y., & Liu, Y. (2025). Data-Driven Loan Default Prediction: A Machine Learning Approach for Enhancing Business Process Management. *Systems*, *13*(7), 581. https://doi.org/10.3390/systems13070581

Zhao, Z., Cui, T., Ding, S., Li, J., & Bellotti, A. G. (2024). Resampling Techniques Study on Class Imbalance Problem in Credit Risk Prediction. *Mathematics*, *12*(5), 701. https://doi.org/10.3390/math12050701