Journal of Computer Science

Research Article

A Comparative Study of Resampling, Cost-Sensitive, and Ensemble Techniques for Handling Class Imbalance in Indonesian Financial Data

Gunawan Kurnia and Ditdit Nugeraha Utama

Department of Computer Science, Bina Nusantara Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article history
Received: 12-03-2025
Revised: 28-07-2025
Accepted: 30-09-2025

Corresponding Author:
Gunawan Kurnia
Department of Computer Science,
Bina Nusantara Graduate Program,
Master of Computer Science, Bina
Nusantara University, Jakarta,
Indonesia
Email: gunawan.kurnia@binus.ac.id

Abstract: Handling class imbalance is a critical challenge in machine learning applications, particularly in financial domains where minority instances often represent significant anomalies such as fraud or audit risks. Various oversampling and undersampling methods were tested, alongside cost-sensitive adjustments and ensemble models including Random Forest, AdaBoost, Gradient Boosting, and XGBoost. The evaluation, based on 10-fold stratified cross-validation and performance metrics such as F1-score, ROC-AUC, and confusion matrix, highlights the superiority of a hybrid approach combining Borderline SMOTE and XGBoost. This configuration achieved near-perfect performance with F1-scores of 0.99 for both classes, demonstrating excellent discrimination and minimal error rates. The findings underscore the importance of method integration in imbalanced data scenarios and offer practical insights for model selection in real-world financial risk modeling.

Keywords: Data Imbalanced, Predictive Modeling, Resampling, Cost Sensitive Learning, Ensemble Learning

Introduction

The objective of this study is to identify the most effective approach for addressing data imbalance in the Indonesian financial industry using techniques such as resampling, cost-sensitive learning, ensemble learning, and hybrid methods, specifically for anomaly detection and audit sampling. The scope of the research focuses on financial industry data with an imbalance ratio of 1:248. Limitations include the absence of extensive hyperparameter tuning and the restriction of data to a single industry domain.

Data imbalance occurs when the amount of the majority class data significantly exceeds that of the minority class within the dataset. This condition causes machine learning models to focus more on learning patterns from the majority class, while paying less attention to the minority class due to its limited representation. This study investigates several mitigation strategies, namely resampling techniques, cost-sensitive learning, and ensemble methods, on a highly imbalanced Indonesian financial dataset to determine the most effective approach.

Each technique provides distinct advantages and limitations in handling imbalanced data. Resampling consists of two main approaches: Oversampling and undersampling. Oversampling increases the number of minority class data, while undersampling reduces the number of majority class data. By applying these techniques, the dataset becomes more balanced, allowing the model to learn patterns from both classes more effectively, but it risks overfitting or information loss. Cost-sensitive learning, by contrast, embeds misclassification costs directly into the training process, encouraging the model to prioritize minority class accuracy (Elkan, 2001). Ensemble methods, such as Random Forest and Boosting, enhance model robustness by combining multiple learners, often improving performance on minority class predictions through their diversity and aggregated decision-making (Galar et al., 2012; Fernández et al., 2018).

By conducting this research, the study seeks to provide practical guidance on the most effective techniques for handling data imbalance in real-world scenarios in a financial context to figure out anomaly detection and audit assessment. The findings will offer valuable insights for researchers and industry practitioners, allowing them to streamline the data preprocessing and modelling stages by choosing the best-suited method for their specific datasets and applications. This contribution is expected to reduce the time and effort required to address data imbalance, enhancing the efficiency of machine learning workflows and improving the quality of predictions.

Related Work

Previous studies have discussed the use of techniques such as SMOTE, cost-sensitive learning, and ensemble learning methods like Random Forest and XGBoost (Mienye and Sun, 2021), are designed to combine the predictions of multiple base models (Breiman, 2001; Chen and Guestrin, 2016), this study offers a novel perspective by systematically combining these approaches and analyzing their collective impact on highly imbalanced financial datasets within the Indonesian context. Unlike prior research that often evaluates these methods in isolation or on synthetic datasets, our work emphasizes a comparative and integrative analysis, including an ablation study to assess the incremental contribution of each technique. Furthermore, by applying the methodology to real-world financial audit data with an extreme imbalance ratio (1:248), this study provides practical insights into handling imbalance in critical decision-making domains, which remains underexplored in existing literature.

Materials

This study utilizes a proprietary financial dataset obtained from an Indonesian financial institution. The dataset comprises 4,978 records, with only 20 records (0.4%) belonging to the minority class and 4,958 records (99.6%) to the majority class, reflecting an imbalance ratio of approximately 1:248.

The extreme imbalance of this dataset poses a significant challenge for predictive modeling, making it an ideal subject for evaluating various imbalance-handling techniques in this study.

The dataset used in this study contains various financial attributes that are critically relevant for anomaly detection purposes and audit assessment, as outlined in Table 1. These attributes include transactional records, customer identifiers, credit collectability, outstanding balances, risk indicators such as Non-Performing Loans (NPL), and Special Mention Accounts (SMA). These variables were carefully selected due to their significance in financial risk analysis. For instance, attributes like Outstanding Balance and NPL are directly associated with a customer's repayment capability and financial health, making them vital indicators in predictive modeling tasks.

Table 1: Data Parameter

No	Attribute	Data Type
1	Area	Categorical (String)
2	ID Customer	Categorical (String)
3	Branch	Categorical (String)
4	Credit Collectability	Numerical (Integer)
5	Outstanding Balance	Numerical (Float)
6	Non-Performing Loan (NPL)	Numerical (Float)
7	Special Mention Account (SMA)	Numerical (Float)
8	Outstanding Balance (Prior Year-End)	Numerical (Float)
9	Outstanding Growth (Year-End Reference)	Numerical (Float)
10	Outstanding Growth (Mid-Year Reference)	Numerical (Float)
11	Collateral Value	Numerical (Integer)
12	SMA under 6 Months	Numerical (Integer)
13	NPL within 1 Year	Numerical (Integer)

Related to risk classification and anomaly detection. Similarly, SMA and Outstanding Growth serve as early warning signals for potential financial distress, thus providing valuable information for audit reviews.

In this study, a feature selection process was carried out to ensure that the selected independent variables have a significant influence on the predictive outcome. This process employed correlation analysis to identify highly correlated features; When two features exhibited strong correlations, one of them was removed to avoid redundancy. Furthermore, the feature selection process also incorporated insights from domain experts to ensure that the retained features are contextually relevant. As a result, only attributes with a strong influence on the target variable were retained for model training and evaluation. This rigorous feature selection ensures that the final model leverages high-quality inputs, thereby increasing its ability to detect minority class instances effectively in a highly imbalanced dataset.

Literature Review

Resampling

Numerous studies have investigated resampling strategies to address class imbalance in classification tasks. Two principal approaches, oversampling and undersampling, have emerged, each with distinct mechanisms and trade-offs:

1. Oversampling techniques aim to increase the representation of the minority class by generating new synthetic samples or duplicating existing ones. In this study, several oversampling methods were applied, including:

- The Synthetic Minority Oversampling Technique (SMOTE) is a well-known approach that generates artificial data points to balance class distributions

(Chawla et al., 2002; Fernández et al., 2018). It works by creating synthetic samples within the feature space, allowing models to better learn from the minority class and improve classification performance

- Random Oversampling, which duplicates minority instances and can be effective yet prone to overfitting (He et al., 2008)
- ADASYN is an oversampling technique designed to generate more synthetic samples in regions where the minority class is sparsely distributed, thereby improving the model's ability to learn difficult decision boundaries, making the model more sensitive to rare patterns., an extension of SMOTE that adapts to data distribution by generating more samples in regions with sparse minority representation (He et al., 2008)
- SMOTE Tomek Links and SMOTE ENN, which integrate oversampling with data cleaning methods through systematic removal of noisy or borderline majority samples, aim to improve class balance (Batista et al., 2004)
- Borderline-SMOTE focuses on generating synthetic samples near the decision boundary, where misclassification is most likely to occur, to enhance classification sensitivity (Han et al., 2005). This strategy allows the model to better learn the separation between classes in critical areas

2. Undersampling strategies. These techniques aim to reduce the number of majority class instances by selective elimination, with the objective of achieving a more balanced distribution across different classes. Techniques implemented in this study include:

- Cluster Centroids, which replace clusters of majority instances with their centroids to maintain representativeness (Yen and Lee, 2009)
- Tomek Links, a noise-reduction method that removes overlapping majority class instances
- Instance Hardness Threshold (IHT), which removes the majority of instances that are difficult to classify correctly
- NearMiss, which retains only the majority samples that are near the minority class, encourages better boundary learning

These techniques were chosen according to their representation in recent literature and their potential to enhance minority class detection. Their implementation in this study provides a comparative foundation for evaluating resampling efficacy under different scenarios of class imbalance.

Cost Sensitive Learning

Cost-sensitive learning has been recognized as a practical approach to handling data imbalance. This strategy incorporates varying misclassification costs, assigning higher penalties to errors involving the minority class (Elkan, 2001; López et al., 2012). Consequently, models are encouraged to improve their accuracy in identifying minority class instances, even under severe imbalance conditions (Ling and Sheng, 2011).

Ensemble Learning

Ensemble learning techniques have shown considerable efficacy in tackling class imbalance problems. By combining multiple weak or base learners, these methods enhance robustness and improve the detection of minority class instances (Galar et al., 2012). Specifically, ensemble methods such as Random Forest, AdaBoost, Gradient Boosting, and XGBoost are widely acknowledged for their strong performance on imbalanced datasets (Liu et al., 2022).

Hybrid Approach

Recent studies have highlighted the growing effectiveness of hybrid approaches that integrate resampling techniques with ensemble learning methods to improve model performance on imbalanced datasets. For instance, Khan et al. (2024) emphasized the potential of combining data augmentation and ensemble models in addressing class imbalance problems across various domains. Similarly, Fatih Gurcan and Soylu (2024) reported that the integration of advanced resampling techniques, particularly BorderLine SMOTE, with ensemble classifiers such as XGBoost, can significantly enhance predictive performance and robustness, especially in sensitive applications like cancer diagnosis.

Methods

This study uses a systematic approach to develop and evaluate a classification model based on financial data in the industry. The research design follows a sequential process, as illustrated in Figure 1, encompassing data collection, preprocessing, model development, and performance evaluation.

This figure illustrates the overall workflow, providing clarity on each stage from raw data to evaluation.

Data Collection

The initial phase involves gathering relevant data from the financial data industry. This step is crucial as it forms the foundation for subsequent analysis. The data collection process may involve extracting information from various sources within the organization, ensuring a comprehensive dataset that captures all relevant variables for the classification task.

In this research, we utilized a financial dataset extracted from the company database containing a total of 4.978 samples.

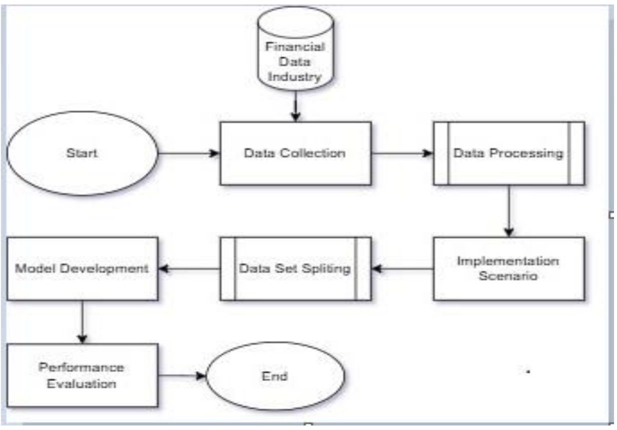

Fig. 1: Research Framework: Data Collection, Preprocessing, Modeling, and Evaluation Stages

The dataset exhibits a significant class imbalance, with the minority class comprising only 20 samples (0.4%) and the majority class encompassing 4958 samples (99.6%), yielding an imbalance ratio of approximately 1:248. This dataset will be used to determine review samples for audit purposes. The severe class imbalance presents a particular challenge for sample selection, as traditional random sampling methods would likely overlook the critical minority class instances. Therefore, our methodology incorporates specialized sampling techniques to ensure adequate representation of both classes in the review process, thereby enhancing the effectiveness of the examination procedures and the reliability of audit outcomes (Figure 2).

Data Preprocessing

The raw data underwent several preprocessing steps to ensure its quality and suitability for analysis. First, data cleaning was conducted by removing duplicate entries, handling missing values through imputation techniques, and correcting inconsistencies in the dataset to establish a solid foundation.

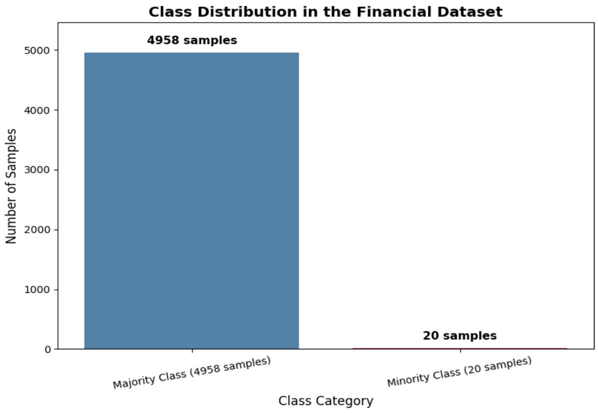

Fig. 2: Class Distribution

Subsequently, normalization was applied to scale numerical features, ensuring they contributed equally to the clustering algorithm without bias from differences in measurement units or ranges.

In the third step, Categorical (non-numeric) variables were encoded into a numerical format using the one-hot encoding method, transforming them into machine-readable inputs.

Finally, Feature selection was conducted to determine the attributes that contribute most substantially to the target variable. This process helps improve model performance by reducing dimensionality and eliminating irrelevant features. This essential process decreased dimensionality, enhanced computational efficiency, and significantly improved the model's interpretability. By focusing on the most relevant features, the model achieved better generalization on unseen data. Multiple techniques were employed in this selection process, including correlation analysis to identify redundant variables and domain knowledge to determine contextual relevance. Through systematic evaluation, redundant or irrelevant features were eliminated, retaining only the most powerful predictors that demonstrated strong statistical relationships with the target outcome.

Method and Implementation Scenario

This phase focuses on executing specific techniques designed to handle data characteristics, especially the class imbalance concern, and to strengthen model performance. The research adopts three fundamental strategies: resampling, cost-sensitive learning, and ensemble learning or a hybrid approach, which combines these strategies to provide comprehensive solutions for imbalanced data challenges.

Resampling Techniques

Resampling methods are applied to adjust the proportion distribution of each class, aiming to mitigate the negative effects of class imbalance on model performance. Two main approaches are implemented:

1) Oversampling the Minority Class:

- Random Oversampling: Randomly duplicating instances from the minority class
- Synthetic Minority Over-sampling Technique (SMOTE): Creating synthetic examples of the minority class based on feature space similarities between existing minority instances

Impact of Synthetic Samples on Model Generalization. While oversampling techniques like SMOTE improve class balance, they also carry the risk of introducing synthetic instances that do not fully represent the true data distribution. This may lead to overfitting or reduced generalization, particularly when synthetic samples overlap excessively with majority class regions (Elreedy and Atiya, 2019). Therefore, careful application and evaluation of oversampling methods are necessary to prevent noise amplification and ensure robust model performance, especially in real-world deployments where unseen data may differ from the training set (Khushi et al., 2021):

2) Under-sampling the Majority Class:

- Random undersampling: Randomly removed instances from the majority class
- Tomek Links: Removing the majority class instances that form Tomek links with minority class instances, focusing on cleaning the class boundaries

The effectiveness of each resampling technique is evaluated to determine the optimal approach for balancing the dataset while preserving important information.

Cost Sensitive Learning

Cost-sensitive learning techniques are implemented to address class imbalance by assigning different misclassification costs to different classes. This approach aims to make the model more sensitive to the minority class without altering the original data distribution.

Key Implementations Include:

1) Weighted Classification:

- Assigning higher weights to minority class instances during model training
- Adjusting the class weight parameter in algorithms that support it

2) Threshold:

- Adjusting the classification threshold to favor the minority class, based on a cost matrix or the class distribution

3) Cost-Sensitive Random Forest:

- Implementing Random Forest that considers misclassification costs when making splitting decisions

The cost matrix is carefully designed based on domain knowledge and the specific implications of different types of misclassifications in the business context.

In this study, cost-sensitive learning techniques were initially applied to the imbalanced dataset without incorporating any resampling methods. This approach was intended to evaluate whether the models could adequately handle the severe imbalance by relying solely on misclassification cost adjustments.

Ensemble Learning

Ensemble learning methods are employed to enhance model performance and robustness, particularly when dealing with imbalanced datasets. In this study, both bagging-based and boosting-based ensemble techniques are implemented as follows:

1) Bagging-based Methods:

- Random Forest: An ensemble of decision trees, each trained on a bootstrap sample of the data to reduce variance and prevent overfitting
- Balanced Random Forest: A modified version of Random Forest that draws balanced bootstrap samples to better handle class imbalance

2) Boosting-based Methods:

- AdaBoost (Adaptive Boosting): Builds an ensemble by sequentially training weak learners, placing greater emphasis on instances that were previously misclassified
- Gradient Boosting Machines (GBM): Construct models in a stage-wise manner, optimizing a loss function using gradient descent
- XGBoost: A sophisticated variant of gradient boosting that incorporates regularization techniques and efficient handling of sparse data for improved accuracy and speed

Before incorporating any resampling techniques, ensemble learning methods such as Random Forest, AdaBoost, Gradient Boosting, and XGBoost were directly applied to the original imbalanced dataset to observe their baseline performance.

In addition, this study also experimented with threshold adjustment to further optimize classification performance, especially for the minority class, by shifting the default decision boundary to better reflect the data distribution.

Hybrid Approach

To enhance the performance of the classification model in a more adaptive and data-driven manner, this study also considers a hybrid strategy by integrating the three primary approaches: Resampling, cost-sensitive learning, and ensemble learning. Rather than combining these methods arbitrarily, the integration is conducted based on performance insights derived from prior individual evaluations.

Specifically, the model first evaluates each standalone technique, resampling, cost-sensitive learning, and ensemble learning independently on the original dataset. The performance of each method is assessed using key evaluation metrics, with particular focus on the F1-score for the minority class (class 1), which is the primary indicator of effectiveness in addressing class imbalance (He and Garcia, 2009).

After identifying the method that yields the highest F1-score for class 1, the selected technique is then combined with the others in a structured manner.

Evaluation Metrics and Rationale

The F1-score is calculated as the harmonic mean of Precision and Recall, using the formula:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Assessment between false positives and false negatives.

To ensure a comprehensive evaluation of model performance on imbalanced datasets, this study employs multiple metrics: Accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). Relying solely on accuracy may lead to misleading conclusions, as it tends to favor the majority class in highly imbalanced scenarios (Powers, 2020). Therefore, precision and recall are critical for measuring the correctness of positive predictions and the model's sensitivity toward true positives, respectively.

The F1-score, as the harmonic mean of precision and recall, provides a balanced assessment, particularly when both false positives and false negatives are significant (Ganganwar, 2012). In this study, the author specifically highlights the F1-score in the classification report to evaluate model performance, as it offers a more comprehensive reflection of effectiveness in handling imbalanced data.

Moreover, the ROC curve and its AUC offer a comprehensive measure of a model's discriminative ability across different thresholds, making them essential for evaluating performance beyond simple accuracy. As highlighted by Powers (2020), the AUC remains a reliable indicator even in imbalanced data settings, providing insights into the trade-off between sensitivity and specificity.

Cross Validation

To verify the robustness and generalizability of our result, we implemented stratified 10-fold cross-validation across all combinations of resampling techniques, cost-sensitive learning methods, and ensemble learning.

Stratified 10-fold cross-validation is employed in the model validation process to divide the dataset into 10 balanced parts, which is critical for obtaining reliable performance metrics in imbalanced classification problems. The choice of 10 folds is based on the statistical bias-variance trade-off analysis by Kohavi (1995), who demonstrated that 10 folds offer an optimal balance between computational efficiency and reliable performance estimation. This has been further supported in imbalanced data settings by Bischl et al. (2007).

Results and Discussion

It should be noted that the F1-score range of 0.63–0.87 refers to the performance of individual methods such as SMOTE, ADASYN, or Random Oversampling (Tables 2–4). Meanwhile, the score of 0.99 represents the best result obtained from the hybrid approach (Table 5), combining Borderline SMOTE with XGBoost.

Besides confusion matrices, the evaluation of classifier performance across different techniques was further supported by Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC), as depicted in Figures 3-7. These visualizations provide a threshold-independent measure of model performance, particularly important in imbalanced settings where simple accuracy is misleading. ROC curves plot the true positive rate (sensitivity) against the false positive rate, while the AUC quantifies the overall ability of the model to distinguish between classes.

As shown in Figures 3 and 4, which present the ROC curves for various oversampling techniques, models trained with SMOTE, ADASYN (He et al., 2008), are an oversampling method that generates additional synthetic samples strategically. These samples are concentrated in areas where the minority class is located, making the model more sensitive to rare patterns.

Table 2: F1 Score Resampling

Model	F1-Score Class 0	F1-Score Class 1
OverSampling		
SMOTE	0.77	0.75
Random Oversampler	0.69	0.64
ADASYN	0.77	0.75
SMOTE Tomek	0.77	0.75
SMOTE ENN	0.78	0.76
Borderline SMOTE	0.85	0.86
UnderSampling		
Cluster Centroid	0.74	0.78
Tomek Links	1.00	0.00
IHT	0.99	0.10
NearMiss	0.73	0.74

Table 3: F1 Score Cost Sensitive Learning

Model	F1-Score Class 0	F1-Score Class 1
CSL	0.99	0.00

Table 4: F1 Score Ensemble Learning

Model	F1-Score Class 0	F1-Score Class 1
AdaBoost	1.00	0.01
Grad Boosting	0.99	0.03
XGBoost	0.99	0.05

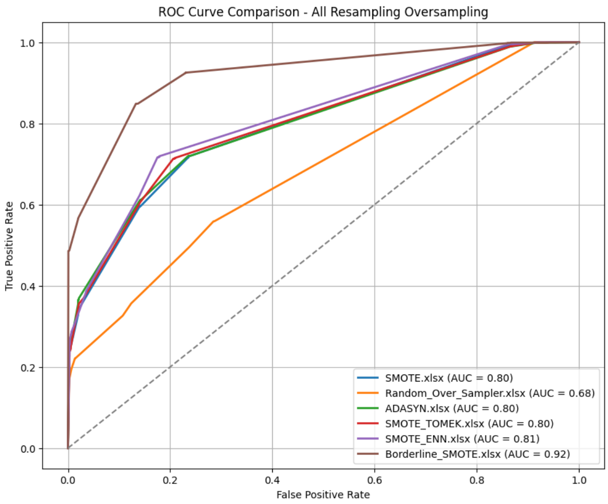

Fig. 3: ROC Curve and AUC – Oversampling

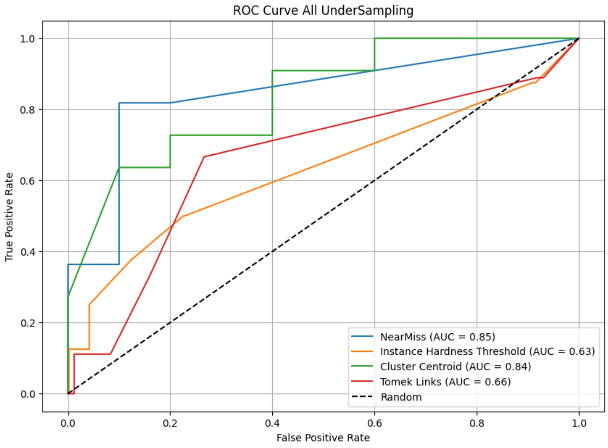

Fig. 4: ROC Curve and AUC – Oversampling

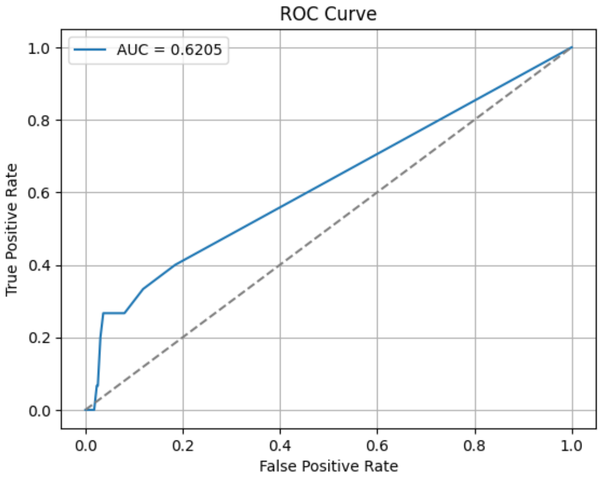

Fig. 5: ROC Curve and AUC – Cost Sensitive Learning

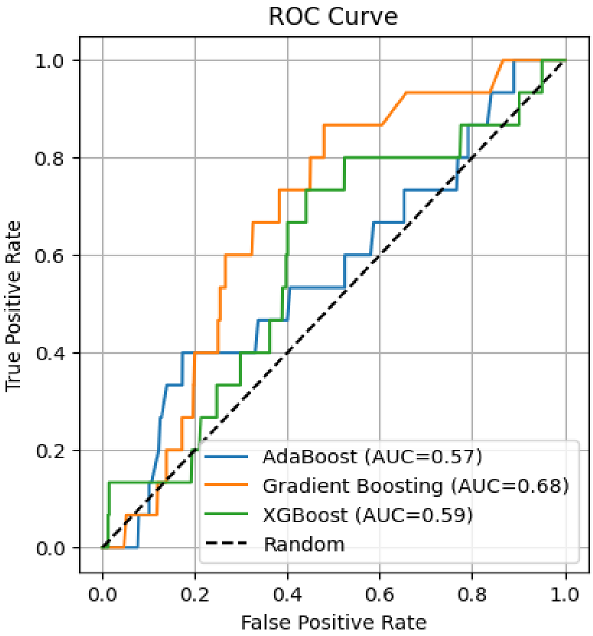

Fig. 6: ROC Curve and AUC – Ensemble Learning

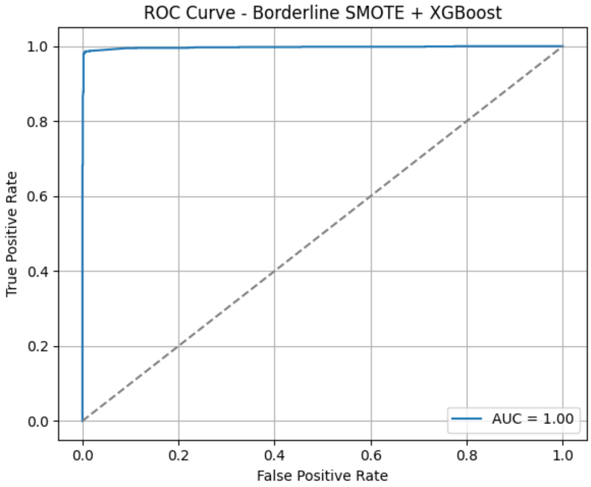

Fig. 7: ROC Curve and AUC – Hybrid Approach

SMOTE-ENN exhibits moderate discriminative ability with AUC values noticeably higher than random guessing (AUC > 0.5), but still not optimal. The Borderline SMOTE configuration once again outperformed the others with a ROC curve that nearly reaches the top-left section of the plot, showing strong sensitivity and minimal false positive occurrences. This is reflected in its AUC value close to 1.0, which aligns with its superior F1-score and confusion matrix in earlier figures. These results confirm that the enhanced synthetic samples generated near decision boundaries effectively help the classifier distinguish minority samples.

Figure 5 illustrates the ROC curve for the Cost-Sensitive Learning (CSL) approach. The ROC curve lies near the diagonal, indicating that the model behaves almost like random guessing. This is in line with the earlier finding from the confusion matrix (Figure 10) and the F1-score of 0.00 (Table 3), confirming that CSL alone was not effective in separating class 1 from class 0 under severe imbalance.

Meanwhile, Figure 6 shows the ROC curves for ensemble learning, such as AdaBoost, Gradient Boosting, and XGBoost, without any resampling. Although these models produced smoother curves than CSL, their AUC values remained low, especially for AdaBoost and Gradient Boosting. XGBoost showed slightly better discrimination but still failed to significantly separate minority instances from majority ones. These curves reflect the models' tendency to prioritize the majority class and overlook minority instances, consistent with their confusion matrices in Figure 11 and low F1-scores in Table 4.

The most notable performance appears in Figure 7, which displays the ROC curve of the hybrid model combining Borderline SMOTE with XGBoost. The curve in this figure almost perfectly hugs the top-left corner, and the corresponding AUC is nearly 1.0. This confirms the model's near-perfect discriminative ability, capable of identifying class 1 with minimal error. This is also supported by the F1-score of 0.99 for both classes (Table 5) and the well-balanced confusion matrix shown in Figure 12. Together, the results from Figures 3-7 reinforce the effectiveness of the hybrid strategy, which integrates data-level and algorithm-level techniques to maximize predictive performance.

The results obtained from this study provide a comprehensive understanding of how different techniques resampling, cost-sensitive learning, ensemble methods, and their hybrid combination—perform in addressing extreme class imbalance. To ensure reliable evaluation, a stratified 10-fold cross-validation was applied, maintaining the original imbalance ratio of 1:248 across all folds. The performance was primarily evaluated using the F1-score metric, supported by ROC-AUC curves and confusion matrices.

Table 2 presents the F1-scores of both majority (class 0) and minority (class 1) classes for various resampling methods. Among the oversampling techniques, Borderline SMOTE clearly outperformed others with an F1-score of 0.86 for class 1, indicating a significant improvement in minority class detection.

This success stems from its strategy of generating synthetic instances near the decision boundary, allowing the classifier to better capture subtle distinctions.

Table 5: F1 Hybrid Approach

Model	F1-Score Class 0	F1-Score Class 1
Borderline SMOTE + XGBoost	0.99	0.99

Other methods, such as SMOTE, ADASYN (He et al., 2008), are oversampling methods that generate more synthetic samples in areas where the minority class is sparsely represented, making the model more sensitive to rare patterns. SMOTE-ENN, and SMOTE-Tomek showed moderate gains (F1-scores ranging between 0.74–0.76) but failed to achieve the same level of precision. Conversely, undersampling methods like Tomek Links and Instance Hardness Threshold (IHT) performed very poorly on class 1, with F1-scores of 0.00 and 0.10, respectively, as these methods eliminated too many majority class samples or failed to capture sufficient minority class structure.

This observation is reinforced by the confusion matrices in Figure 8 and 9, where oversampling, particularly Borderline SMOTE resulted in increased True Positives (TP) for class 1 with minimal false negatives (FN), while undersampling methods misclassified nearly all minority instances, resulting in zero or near-zero TP values. Although some undersampling approaches preserved class balance, they often did so at the expense of losing valuable information.

Table 3 summarizes the results for Cost-Sensitive Learning (CSL). Despite assigning higher misclassification costs to the minority class, the model achieved an F1-score of 0.00 for class 1.

The corresponding confusion matrix in Figure 10 shows that the classifier misclassified all 20 minority instances as the majority class. This demonstrates that, under conditions of extreme imbalance, cost-sensitive learning alone is insufficient to achieve meaningful recall or precision for the minority class, even when penalization mechanisms are introduced during model training. Moving to Table 4, which reports F1-scores for ensemble learning methods (without any resampling), it is evident that these algorithms also struggled with class 1 detection.

XGBoost slightly outperformed the others with an F1-score of 0.05, followed by Gradient Boosting (0.03) and AdaBoost (0.01). As shown in Figure 11, their confusion matrices confirm that the majority of class 1 instances were misclassified, reflecting a strong bias toward class 0. Although ensemble methods offer better generalization through aggregation of weak learners, this experiment highlights that ensemble methods alone are not sufficient to overcome the representational limitations inherent in extremely imbalanced datasets.

In contrast, Table 5 demonstrates the outstanding performance of the hybrid approach combining Borderline SMOTE with XGBoost. This configuration achieved F1-scores of 0.99 for both class 0 and class 1, indicating an almost perfect balance between precision and recall for both classes. The confusion matrix in Figure 12 shows that nearly all minority instances were correctly identified, with very few misclassifications. Similarly, the ROC curve in Figure 7 displays an AUC close to 1.0, reflecting exceptional discriminative power across thresholds. This outcome supports the study's central finding: That a hybrid strategy integrating both data-level (resampling) and algorithm-level (ensemble learning) interventions is the most effective solution to handle extreme imbalance.

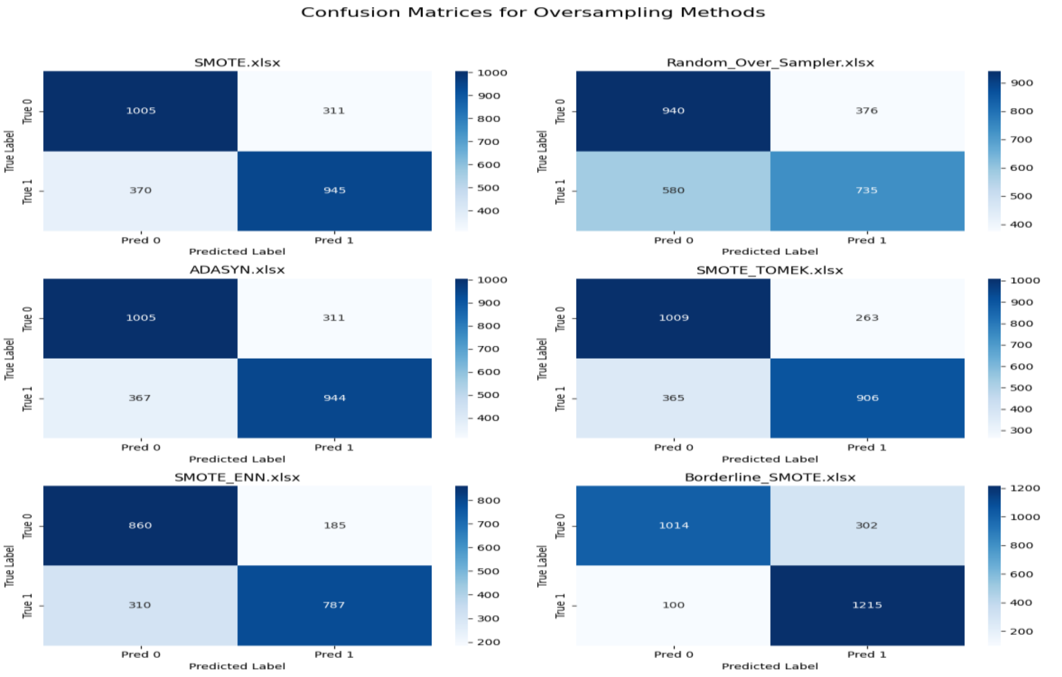

Fig. 8: Confusion Matrices for Oversampling Methods

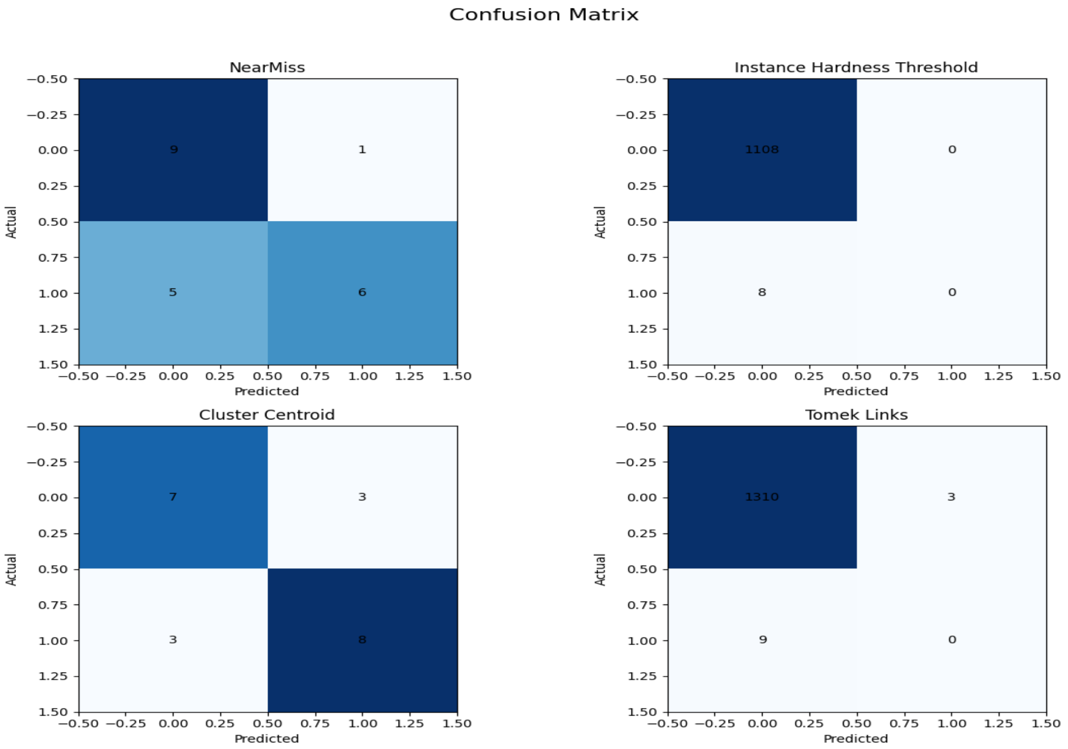

Fig. 9: CM Undersampling

The earlier figures also substantiate these findings. Figures 3 to 6 compare the ROC curves of various standalone techniques, where Borderline SMOTE and hybrid models clearly outperformed others. The AUC values for cost-sensitive learning and standalone ensemble methods were low, often close to the random baseline, whereas Figure 7 (Hybrid) displayed a curve hugging the top-left corner, indicative of near-perfect classification. These visualizations underscore the critical importance of combining well-selected resampling with a powerful classifier to improve both sensitivity and specificity in highly skewed datasets.

In conclusion, the comparative analysis of all four scenarios resampling (Table 2), cost-sensitive learning (Table 3), ensemble learning (Table 4), and hybrid (Table 5) clearly points to the hybrid model as the most effective strategy for addressing severe class imbalance. Not only does it deliver superior F1-scores and AUC, but its confusion matrix also confirms reliable and balanced performance across both classes. This integrated method provides an operationally viable solution for real-world tasks such as anomaly detection, audit sampling, and financial risk classification, where minority instances often represent critical outcomes.

The confusion matrices presented in Figures 8 through 12 provide detailed insights into the classification performance of each technique, particularly in handling the minority class (class 1), which is critically underrepresented in the dataset. As illustrated in Figure 8, oversampling techniques generally improved the model's ability to detect minority instances. Among them, Borderline SMOTE exhibited the most balanced confusion matrix, correctly identifying nearly all class 1 samples (high true positives) while maintaining a low false positive rate.

This corresponds to the highest F1-score of 0.86 for class 1 among all oversampling methods. Other methods, such as SMOTE, ADASYN (He et al., 2008), are oversampling methods that generate more synthetic samples in areas where the minority class is sparsely represented, making the model more sensitive to rare patterns. SMOTE-Tomek and SMOTE-ENN also increased the number of true positives for class 1 compared to the original dataset, but still suffered from a moderate number of false negatives. This indicates that while synthetic oversampling helped, only Borderline SMOTE effectively generated samples that accurately reflect the decision boundary between the two classes.

In contrast, Figure 9, which visualizes the confusion matrices for undersampling techniques, reveals a significant performance drop in identifying class 1. Techniques like Tomek Links and Instance Hardness Threshold (IHT) failed to classify any class 1 instances correctly, resulting in zero true positives and all 20 minority instances misclassified as the majority class (false negatives). This outcome is consistent with the F1-score of 0.00 shown in Table 2. While Cluster Centroid and NearMiss retained a few true positives, they still lost valuable majority class information and provided no substantial improvement compared to oversampling methods.

Moving to Figure 10, the confusion matrix for Cost-Sensitive Learning (CSL), the model again misclassified all class 1 instances as class 0. This resulted in no true positives, which confirms the extremely poor recall and precision for the minority class and explains the F1-score of 0.00 reported in Table 3. Despite CSL theoretically assigning higher penalty costs for minority class misclassification, it proved ineffective when applied in isolation under such an extreme imbalance (1:248).

Figure 11 shows the confusion matrices for ensemble methods (AdaBoost, Gradient Boosting, and XGBoost) applied directly to the imbalanced dataset without resampling. These models performed well on the majority class (class 0), achieving high true negatives. However, their ability to correctly classify class 1 remained poor. For example, XGBoost managed to classify only 1 out of 20 minority instances correctly, while AdaBoost and Gradient Boosting performed even worse. The confusion matrices reveal severely skewed classification patterns, aligning with the low F1-scores for class 1 in Table 4.

Finally, Figure 12 represents the confusion matrix of the hybrid approach combining Borderline SMOTE with XGBoost, which clearly stands out. The matrix shows a nearly perfect classification of both classes, with almost all class 1 instances correctly predicted as positive (high TP) and minimal false positives.

This performance aligns with the F1-score of 0.99 for both classes reported in Table 5 and supports the earlier claim of "near-perfect" classification. The model not only succeeded in detecting minority class instances but also preserved high accuracy in the majority class, making it the most effective strategy for addressing extreme data imbalance.

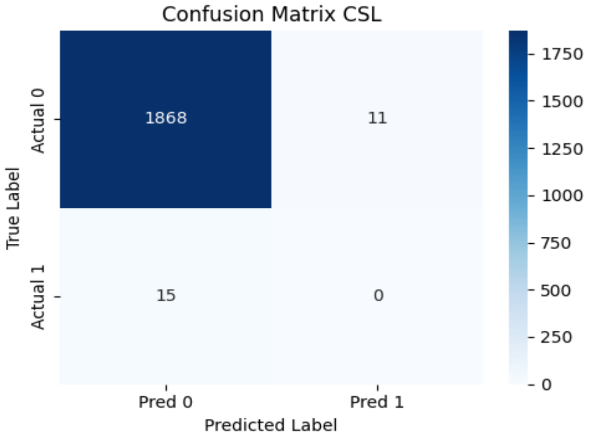

Fig. 10: CM CSL

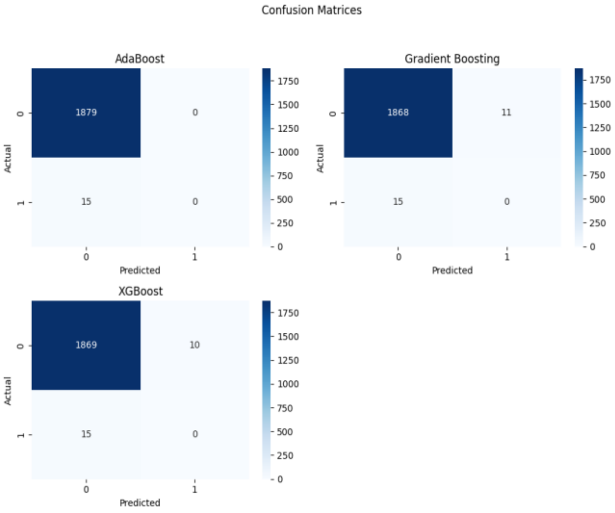

Fig. 11: CM Ensemble Learning

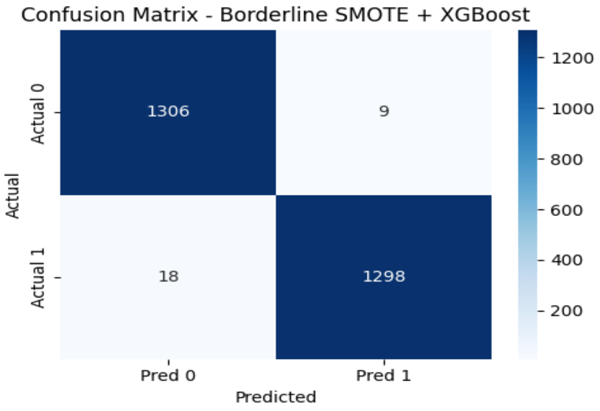

Fig. 12: CM Borderline SMOTE + XGBoost

Discussion

Although this study thoroughly explores resampling, cost-sensitive learning, and ensemble methods, it acknowledges a limitation in the comprehensive tuning of hyperparameters, particularly for tree-based ensemble algorithms such as Random Forest and XGBoost. While default parameters offer a baseline, prior studies suggest that hyperparameter optimization, for example, adjusting the number of estimators, learning rate, maximum depth, and subsample ratios, can substantially influence model performance, especially on imbalanced datasets (Bergstra and Bengio, 2012; Probst et al., 2019). Future studies are encouraged to employ comprehensive hyperparameter optimization methods, including approaches like grid search or Bayesian optimization, to enhance model robustness and predictive reliability, and to further enhance the predictive capability and generalizability of the models.

Future studies may also consider exploring alternative classification algorithms beyond tree-based ensembles. Algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Logistic Regression with advanced regularization, and deep learning models like neural networks could offer different perspectives in handling imbalanced data.

Conclusion

This study concludes that addressing extreme class imbalance effectively requires an integrated hybrid approach, rather than relying on a single technique. Based on results obtained from 10-fold stratified cross-validation, the most reliable and accurate model configuration was the combination of Borderline SMOTE and XGBoost, as seen in Table 5. This hybrid model consistently achieved F1-scores of 0.99 for both majority and minority classes, significantly outperforming other individual approaches (Tables 2–4).

Figure 11 confirms that the model accurately detects both classes with minimal errors. The hybrid model not only improves the representation of minority instances but also ensures decision boundaries are effectively learned through iterative boosting. This synergy addresses both data imbalance and model optimization, making it especially valuable for real-world applications like fraud detection, financial audit sampling, and risk classification.

Further supporting its effectiveness, the ROC curve in Figure 6 and its corresponding AUC demonstrate excellent discriminatory power between the classes. This is especially important in imbalanced datasets where accuracy alone can be misleading. AUC provides a threshold-independent metric, ensuring the model's performance remains stable even under shifting operational conditions.

Acknowledgment

We extend our appreciation to BINUS (Bina Nusantara University) for their help and encouragement, which played a vital role in fostering collaboration and enhancing the depth of our research.

Funding Information

This research was self-funded by the authors, reflecting their independent commitment to advancing scientific inquiry.

Authors Contributions

Gunawan Kurnia: Conducted the research, performed the analysis, and drafted the manuscript.

Ditdit Nugeraha Utama: Supervised the research and reviewed the manuscript.

Ethics

The submitted manuscript constitutes original work by its authors and has not been published or is under consideration elsewhere. Each author has reviewed and approved their content, confirming its accuracy and adherence to academic guidelines. The study and its dissemination were carried out in strict adherence to ethical principles, with no conflicts of interest or ethical concerns arising at any stage. Moreover, the research fully conformed to the ethical guidelines prescribed by Bina Nusantara University, underscoring our commitment to responsible research.

References

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Bischl, B., Bartz-Beielstein, T., & Preuss, M. (2007). Experimental research in evolutionary computation. *Proceedings of the 9th Annual Conference Companion on Genetic and Evolutionary Computation*, 3001–3320. https://doi.org/10.1145/1274000.1274102

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 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

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Elkan, C. (2001). The foundations of cost-sensitive learning. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI’01*, 973–978.

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463–484. https://doi.org/10.1109/tsmcc.2011.2161285

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42–47.

Gurcan, F., & Soylu, A. (2024). Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. *Cancers*, *16*(19), 3417. https://doi.org/10.3390/cancers16193417

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *Advances in Intelligent Computing*, *3644*, 878–887. https://doi.org/10.1007/11538059_91

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceeding of the IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/ijcnn.2008.4633969

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244*, 122778. https://doi.org/10.1016/j.eswa.2023.122778

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access*, *9*, 109960–109975. https://doi.org/10.1109/access.2021.3102399

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137–1145.

Ling, C. X., & Sheng, V. S. (2011). Cost-Sensitive Learning. *Encyclopedia of Machine Learning*, 231–235. https://doi.org/10.1007/978-0-387-30164-8_181

Liu, L., Wu, X., Li, S., Li, Y., Tan, S., & Bai, Y. (2022). Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. *BMC Medical Informatics and Decision Making*, *22*(1), 82. https://doi.org/10.1186/s12911-022-01821-w

López, V., Fernández, A., Moreno-Torres, J. G., & Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications*, *39*(7), 6585–6608. https://doi.org/10.1016/j.eswa.2011.12.043

Mienye, I. D., & Sun, Y. (2021). Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. In *Informatics in Medicine Unlocked* (Vol. 25, p. 100690). https://doi.org/10.1016/j.imu.2021.100690

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, *2*(1), 37–63. https://doi.org/https://doi.org/10.48550/arXiv.2010.1606

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest [dataset]. In *WIREs Data Mining and Knowledge Discovery* (Vol. 9, Issue 3, p. e1301). https://doi.org/10.1002/widm.1301

Yen, S.-J., & Lee, Y.-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, *36*(3), 5718–5727. https://doi.org/10.1016/j.eswa.2008.06.108