Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15029
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Integrating SMOTE with XGBoost for Robust Classification on Imbalanced Datasets: A Dual-Domain Evaluation

Novriadi Antonius Siagian[1)*], Sardo P Sipayung[2)], Alex Rikki[3)], Nasib Marbun[4)]
[1,2,3)] Universitas Katolik Santo Thomas, Indonesia [4)] Universitas Negeri Manado, Indonesia
[1)] novriadi.antonius95@gmail.com, [2)] pinsarsiphom@gmail.com, [3)] alexrikisinaga@gmail.com , [4)] nasibmarbun@unima.ac.id

Submitted :Jun 30, 2025 | **Accepted** : Jul 14, 2025 | **Published** : Jul 16, 2025

Abstract: Class imbalance is one of the main challenges in classification problems, as it can reduce the model's ability to accurately identify minority classes and negatively impact the overall reliability of predictions. In response to this problem, this study proposes an integrated approach combining SMOTE and XGBoost to improve classification performance on imbalanced data. This approach aims to evaluate the impact of oversampling techniques on prediction accuracy and model sensitivity to class distribution. The evaluation was conducted using two public datasets representing different domains and different amounts of data, namely Spambase and Diabetes, to assess the effectiveness and generalization of the applied approach. The experimental results show that this integrated model consistently outperforms traditional comparison algorithms, with an F1 score of 0.94 and ROC-AUC of 0.98 on the Spambase dataset and ROC-AUC of 0.83 on the Diabetes dataset, with a good balance between precision and recall. The 10-fold cross-validation technique was applied to ensure objective performance estimates free from random data splitting bias. Additionally, this study highlights the importance of selecting appropriate evaluation metrics in the context of imbalanced data, as single accuracy often provides a misleading performance picture. This study makes a significant contribution by providing a benchmark for comparing the effectiveness of SMOTE-XGBoost integration using two different datasets, accompanied by rigorous cross-validation. These findings reinforce the position of integrating data preprocessing strategies and ensemble learning as a competitive and adaptive solution for addressing class imbalance challenges in data-driven classification systems.

Keywords: Smote; machine learning; Imbalanced Data; Classification; XGBoost

INTRODUCTION

In contemporary data-driven environments, data classification plays a central role in various intelligent systems. However, a major challenge often encountered in classification tasks is class imbalance, where the distribution of samples across classes is highly skewed. This

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15029
e-ISSN : 2541-2019
p-ISSN : 2541-044X

imbalance can significantly affect model performance by causing the classifier to prioritize the majority class while underrepresenting the minority class. As a result, models may demonstrate deceptively high accuracy while failing to detect rare but crucial instances, leading to what is widely recognized as the accuracy paradox (Kim & Kim, 2020) (S. Wang et al., 2021) (Elreedy & Atiya, 2019).

The challenge intensifies in large-scale datasets, where the degree of imbalance tends to escalate alongside data volume, making it increasingly difficult for machine learning models to accurately learn from the minority class (Elreedy & Atiya, 2019). Overlapping class boundaries, misclassification of minority samples, and deteriorating predictive performance are common consequences of improper class distribution. To address this, existing studies have introduced methods categorized into three main strategies: data-level approaches, algorithm-level modifications, and hybrid strategies (Xu et al., 2020).

Among data-level methods, SMOTE has emerged as one of the most prominent oversampling strategies. SMOTE creates synthetic instances of the minority class by interpolating between existing examples, thereby improving balance without introducing simple duplicates (Elreedy & Atiya, 2019). Numerous studies have reported its effectiveness in enhancing classification performance across domains such as student grade prediction (Abdul Bujang et al., 2023), fraud detection (Sun et al., 2020), and healthcare (Xu et al., 2020). To further improve sample quality, SMOTified-GAN, which combines SMOTE with Generative Adversarial Networks (GANs), has been introduced to generate more realistic minority class samples. Experimental results have shown that this integration can increase minority class quality by up to 9% (Sharma et al., 2022) (Dablain et al., 2023).

Despite these advances, traditional machine learning algorithms often lack robustness when facing severe class imbalance. As highlighted in recent research, their classification precision can significantly drop under such conditions (Sun et al., 2020). To provide more efficient training on imbalanced datasets, more complex models like Artificial Neural Networks (ANNs) and Deep Neural Networks (DNNs) have been improved with methods like batch normalizing, ReLU activation, and Borderline-SMOTE (Dablain et al., 2023). Using oversampling approaches with highly regularized ensemble-based classifiers is a huge step toward resolving class imbalance. Gradient boosting methods are great for working with sparse features and noisy datasets since they come with regularization algorithms built in (Chen & Guestrin, 2016) (C. Wang et al., 2020). In combination with DNNs, XGBoost has been used effectively in tasks such as network intrusion detection, forming hybrid models that involve feature normalization, selection, and deep classification layers (Devan & Khare, 2020). These models leverage the optimal aspects of both ensemble learning and deep learning to enhance generalization and predictive accuracy, even in the presence of imbalanced data (Nobre & Neves, 2019) (Zhang et al., 2018).

In these situations, metrics like precision-recall, ROC-AUC, and F1-score are commonly used to evaluate model performance (Hand, 2009) (Verbakel et al., 2020). Conventional handling without addressing the imbalance can lead to misleading evaluations, especially when the imbalance ratio is high.

In conclusion, many methods have been proposed to fix class imbalance, including data sampling methods, cost-sensitive learning, and ensemble models. However, a full comparison of machine learning algorithms using standardized methods like SMOTE across different fields is still needed. This study aims to find out how well different machine learning models operate on SMOTE-balanced datasets by using different evaluation methods to find the best ways to deal with class imbalance in the real world.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i3.15029 p-ISSN : 2541-044X

LITERATURE REVIEW

Handling class imbalance has become a central concern in the development of machine learning classification models, as disproportionate class distributions can significantly degrade model performance, particularly in recognizing minority classes. One common approach to addressing class distribution imbalances is to add synthetic samples to minority classes through linear interpolation of the nearest neighbors. This strategy aims to create a more balanced representation of the data, thereby supporting a more equitable model learning process and enabling more effective generalization of patterns. The effectiveness of SMOTE has been demonstrated across various domains, including healthcare, social media, and financial prediction systems. For instance, studies in child stunting detection show that integrating SMOTE with the XGBoost model significantly improves classification performance, achieving an accuracy of 85.74%, a recall of 89.14%, and a ROC-AUC of 93.11%, indicating superior sensitivity to critical minority cases (Sugihartono et al., 2025). Similarly, research utilizing Random Forest for cardiovascular disease prediction reported improvements in accuracy, sensitivity, and specificity after applying SMOTE (Hasanah et al., 2024). In the domain of social media analysis, the combination of SMOTE with majority voting approaches yielded classification accuracy as high as 97% for sentiment analysis on Twitter data, with notable enhancements in recall and F1-score, reflecting increased sensitivity to minority classes (Suandi et al., 2024).

Beyond standard SMOTE, several variants have been developed to overcome limitations such as the generation of noisy or overlapping samples. Borderline-SMOTE, for example, focuses on synthesizing samples near class decision boundaries to improve classification in highly imbalanced datasets. Studies have shown that Borderline-SMOTE consistently outperforms standard SMOTE in terms of accuracy and F1-score (Ujaran et al., 2024). Other variants, such as SMOTE-ENN and SMOTE-Tomek Link, combine oversampling with noise-cleaning techniques to produce cleaner and more accurate data representations (Nemade et al., n.d.) (Husain et al., 2025). Furthermore, several methodologies integrate clustering algorithms with SMOTE to guarantee that the produced synthetic data aligns with the original data structure (Jiang et al., 2025).

Overall, comparative results from multiple studies suggest that SMOTE and its variants significantly enhance performance metrics, with balanced accuracy reaching up to 98.75% in certain Random Forest implementations, and consistently outperforming other oversampling methods such as ADASYN across different contexts. However, the effectiveness of SMOTE is highly dependent on the characteristics of the dataset and the classification algorithm used (Halim et al., 2023). Therefore, the selection of SMOTE variants and their strategic integration with other techniques such as ensemble learning, normalization, or clustering must be tailored to the specific requirements of each case. As such, SMOTE remains a vital component in the classification pipeline for imbalanced data.

METHOD

This study employs an experimental quantitative method to assess the effectiveness of different machine learning algorithms in categorizing imbalanced data that has been pre-processed using oversampling methods. The primary goal is to evaluate the performance of models before and after applying these methods using various assessment measures, including accuracy, F1-score, precision, recall, and ROC-AUC.

Dataset

The evaluation and comparison of models in this study were conducted using datasets ranging from small to large scale, all of which explicitly exhibit class imbalance characteristics,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15029
e-ISSN : 2541-2019
p-ISSN : 2541-044X

as shown in Table 1. The selection of datasets was based on varying class proportions that reflect real-world challenges in imbalanced data classification. All datasets were sourced from reputable open-access repositories, the UCI and Kaggle Machine Learning Repositories have both been widely used in previous studies on how to classify and manage imbalanced data.

Table 1. Description for datasets

Dataset	Features	Classess	Instances	Minority Class (%)	Description
Diabetes	8	2	768	34.90%	Predict based on diagnostic measurements whether a patient has diabetes.
Spam Base	57	2	4601	39.4%	Classification of Spam and Non-Spam Emails

Imbalance Class

Class imbalance is a prevalent issue across various domains, including chemical and biochemical engineering, financial management, information technology, cybersecurity, business analytics, agriculture, and emergency management systems (Douzas & Bacao, 2019). It occurs when the distribution of samples among classes is uneven, specifically in binary classification problems, where one class (the majority) significantly outnumbers the other (the minority) (Sun et al., 2020).

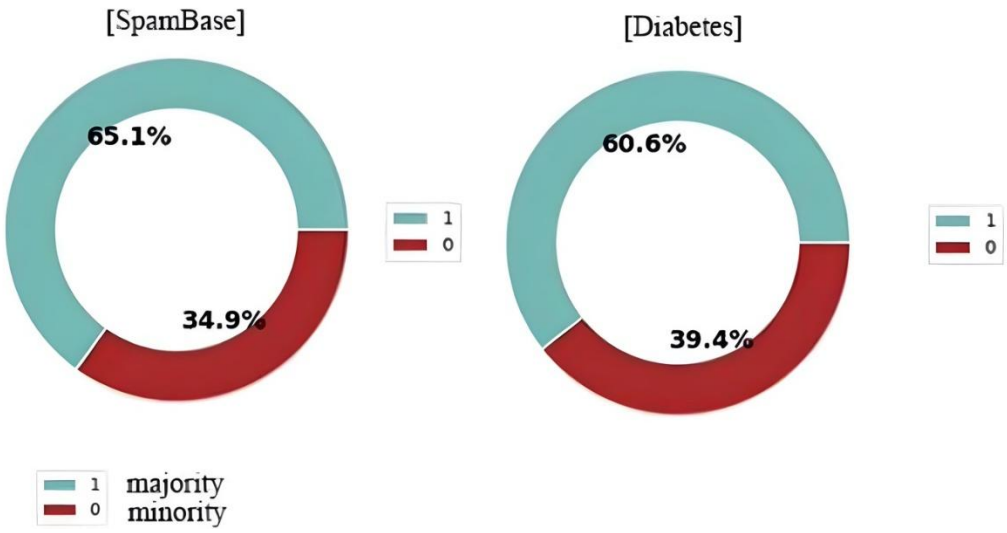

Figure 1. Imbalance Data

In Figure 1, the datasets used in this study, namely SpamBase and Diabetes, clearly exhibit imbalanced class distributions, with the minority class comprising a significantly smaller proportion of the total instances.

Synthetic Minority Oversampling Technique (SMOTE)

Oversampling is a technique employed to rectify class imbalance by augmenting the sample size of the minority class to equal that of the majority class. A prevalent method is closest neighbor oversampling, wherein supplementary samples from the minority class are produced

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i3.15029 p-ISSN : 2541-044X

based on the distribution of its nearest neighbors, utilizing the Euclidean distance metric (Adeoti Babajide Ebenezer1, 2021). This technique is considered efficient, easy to implement, and capable of expanding the training dataset size while increasing the number of features available for model learning, thereby improving the overall accuracy of the classification model (Sharma et al., 2022) (Xu et al., 2020).

SMOTE is an essential element in the oversampling technique. This technique entails creating synthetic samples from the minority class by linear interpolation between a selected minority data point x_i and one of its nearest neighbors $\bar{x}$, both belonging to the same class. The mathematical formulation for generating new samples is as follows:

$$x_{\text{syn}} = x_i + (\bar{x} - x_i) \times \text{rand}(0,1) \quad (1)$$

Here, x_i is the selected minority sample, $\bar{x}$ is one of its k-nearest neighbors in the minority class, and rand(0,1) denotes a random number uniformly sampled from the interval (0, 1). This interpolation strategy enables SMOTE to generate diverse yet realistic synthetic samples, thereby enhancing the classifier's ability to learn minority class patterns more effectively.

An enhanced version of SMOTE, incorporating normal distribution principles, aims to generate synthetic samples that better reflect the statistical characteristics of the minority class. This approach improves the representational consistency of the synthetic data with the original data distribution and reduces the risk of introducing noise or class overlap. The methodological steps of this approach are outlined as follows (S. Wang et al., 2021),

In the preprocessing stage prior to applying SMOTE, normalization was performed on numerical features to ensure that all attributes were scaled uniformly. This step is crucial because SMOTE generates synthetic data based on distance calculations between data points, and variations in scale can lead to unrepresentative interpolations. Among the available normalization methods, Min-Max Normalization was selected due to its simplicity and effectiveness in preserving the proportional relationships between features without altering the original data distribution (S. Wang et al., 2021). Unlike Z-score normalization, which is more appropriate for normally distributed data, or Robust Scaler, which is designed for datasets with significant outliers (Sun et al., 2020), Min-Max was deemed more suitable in this study since the data had been pre-cleaned and contained no extreme values. By ensuring consistent feature scaling, the synthetic samples produced by SMOTE become more stable and representative, thereby supporting a more reliable classification process.

The first step in enhancing SMOTE with a normal distribution-based approach involves standardizing each feature of the minority class samples to ensure consistent scale and comparability. This standardization is performed using min–max normalization, which transforms the values of each feature into the range [0,1]. The normalization formula is defined as:

$$x_{ij}^{*} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \quad (2)$$

In this expression, x_{ij} denotes the value of the j feature for the I sample; x_j^{min} and x_j^{max} refer to the minimum and maximum values of the j feature across all samples belonging to the minority class. This normalization ensures that all features contribute equally in subsequent stages and prevents dominance by features with larger numeric ranges, thereby improving the quality of synthetic sample generation.

After the normalization process is completed, the next step is to calculate the centroid (mean vector) of the normalized minority class samples. This centroid represents the central tendency of the minority class distribution and serves as a primary reference point in the generation of synthetic samples. The calculation is formulated as follows:

$$\bar{x}' = \frac{1}{n} \sum_{i=1}^{n} x_i^{*} \quad (3)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i3.15029 p-ISSN : 2541-044X

Where n denotes the total number of minority class samples, and x_i^* represents the normalized feature vector of the *i* sample. By averaging all normalized minority class samples, this centroid provides a general representation of the feature distribution within the class and offers a stable foundation for distribution estimation and the generation of more realistic and targeted synthetic data.

The third step in the normal distribution-based SMOTE enhancement involves estimating the standard deviation of each normalized feature in the minority class dataset. This step is essential to quantify the spread or dispersion of data points around the centroid, thus capturing the variability inherent in each feature. The standard deviation for the j feature is calculated using the following formula:

$$\sigma_j = \frac{1}{n}\sum_{i=1}^{n} \left(x_{ij}^* - \bar{x}_j'\right)^2 \quad (4)$$

Where x_{ij}^* represents the normalized value of the j feature for the i sample, and $\bar{x}_j'$ denotes the mean of the j feature across all normalized minority class samples. The resulting standard deviation values are used to guide the synthesis process by controlling the magnitude of variation around the centroid. This ensures that newly generated synthetic samples maintain statistical consistency with the original feature distribution and reduces the risk of generating unrealistic or noisy data points.

The fourth step in the development of SMOTE based on a normal distribution involves synthesizing new samples that incorporate the statistical dispersion of the minority class. In this stage, synthetic samples (p_i) are generated by adding noise drawn from a standard normal distribution to the centroid of the normalized minority class data. This process is formulated as:

$$p_i = \bar{x}' + \sigma \cdot N(0,1) \quad (5)$$

Where $\bar{x}'$ represents the centroid vector of the standardized minority class data, σ is the vector of standard deviations for each feature, and $N(0,1)$ denotes a vector of random values drawn from the standard normal distribution. By introducing controlled noise around the centroid, the newly generated samples remain within a relevant range of the feature space while maintaining statistical consistency with the original data distribution. This approach enables the creation of more diverse and realistic synthetic samples, while also minimizing the risks of class overlap or the generation of non-representative instances.

This study uses the 10-fold cross-validation technique on oversampled training data to obtain a more reliable performance evaluation and reduce bias due to random data division. In this approach, the training data is divided into ten stratified and balanced subsets; each subset is used in turn as validation data, while the other nine subsets serve as training data. This process is repeated ten times so that each subset is used exactly once as test data. The average accuracy is calculated to represent the overall performance of the model, while the standard deviation is used to measure the consistency of performance across folds. The choice of K = 10 is based on empirical results showing that this configuration provides an optimal balance between bias and variance in classification tasks (James et al., 2021).

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an advanced development of the gradient boosting technique that utilizes an ensemble learning approach to build highly accurate predictive models. In the process, XGBoost gradually produces a series of weak learners, which are simple models that have slightly better classification capabilities than random guesses, generally in the form of decision trees with limited depth. Each model is built sequentially, with subsequent iterations focused on correcting the prediction errors of the previous model. This goal is achieved by giving greater weight to training samples that were previously misclassified.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i3.15029 p-ISSN : 2541-044X

Through this approach, the overall performance of the model improves progressively as the number of iterations increases. (Zhang et al., 2018) (C. Wang et al., 2020).
Formula Extreme Gradient Boosting (XGBoost)

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \quad (6)$$

where fk denotes the k decision tree (serving as a weak learner), K represents the total number of iterations or trees in the ensemble, and F is the function space comprising all decision tree models. The vector x_i refers to the input features of the i sample being processed. This approach constructs the model in an additive manner, where each function fk is responsible for correcting the prediction errors made in the previous iteration, thereby progressively improving the model's accuracy at each stage.

Objective function

Used to measure and minimize prediction errors while controlling model complexity.

$$L(t) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (7)$$

Represents the loss function, which quantifies the discrepancy between the true label yi and the prediction $\hat{y}_i^{(t-1)}$ from the previous iteration, incremented by the output of the newly added decision tree $f_t(x_i)$. The function $l(\cdot)$ can take the form of squared error for regression or log-loss for classification tasks, depending on the problem domain. The second component, $\Omega(f_t)$, denotes the regularization term, which controls the complexity of the model.

Performance Evaluation

In classification tasks involving imbalanced datasets, relying solely on accuracy as an evaluation metric does not adequately represent the model's total performance. Consequently, evaluation is conducted by incorporating numerous supplementary metrics to present a more comprehensive assessment of the model's capacity to identify both minority and majority classes. This multi-metric methodology facilitates a more precise and insightful performance evaluation (Verbakel et al., 2020).

Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

TP represents True Positive, TN denotes True Negative, FP signifies False Positive, and FN indicates False Negative. Accuracy quantifies the ratio of correct predictions to the total instances; nevertheless, it can be deceptive in imbalanced scenarios, since the model may exhibit bias towards the majority class.

Precision

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

Precision denotes the ratio of true positive predictions to the total positive predictions, illustrating the model's efficacy in accurately detecting minority class cases while minimizing false positives.

Recall

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

This metric measures the model's ability to correctly detect all actual positive instances. In imbalanced data scenarios, recall is particularly important since it reflects how well the model captures the minority class.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15029
e-ISSN : 2541-2019
p-ISSN : 2541-044X

F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

The F1-score is particularly significant when a trade-off exists between precision and recall, rendering it an essential metric for imbalanced datasets.

ROC-AUC

Assesses the model's ability to differentiate between categories. An AUC approaching 1 indicates robust discriminative ability, whereas an AUC near 0.5 implies performance comparable to random chance.

RESULT

This section delineates the outcomes of the implementation and assessment of the classification method utilizing the XGBoost algorithm, which has been refined by the SMOTE oversampling methodology. The results are systematically presented, beginning with a flowchart that illustrates the overall research process, followed by a comparative analysis of model performance.

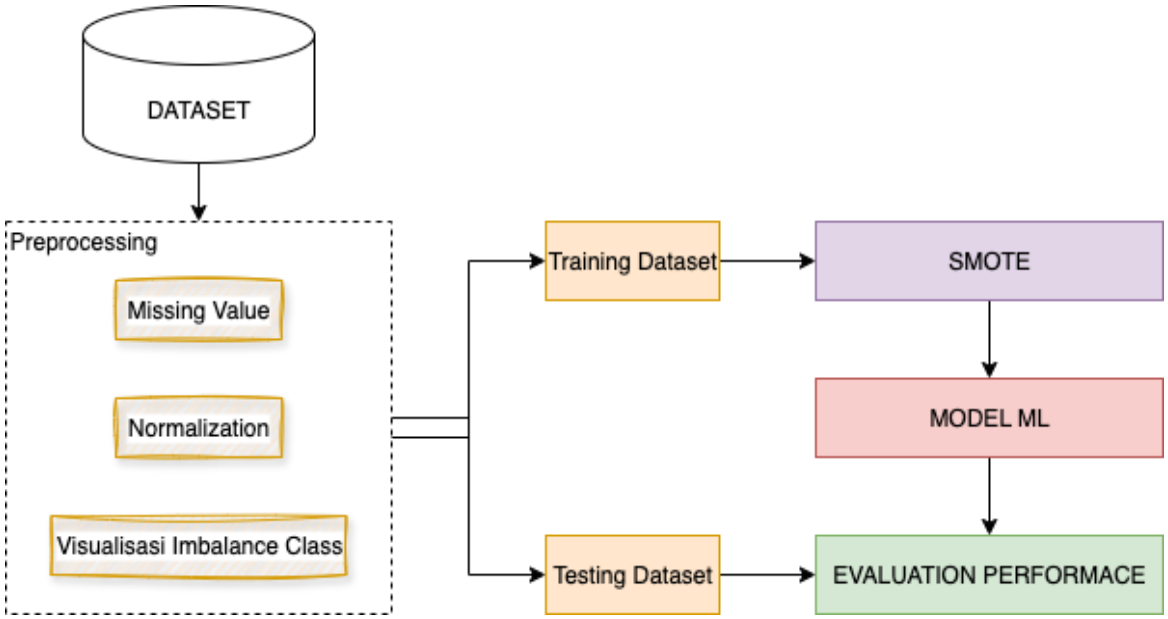

Figure 2. Flow

The research stages commence with dataset collecting, succeeded by a data preprocessing step. This phase comprises three primary steps: addressing missing values, executing normalization to standardize all features, and displaying class distribution to detect any data imbalance.

After preprocessing is done, the dataset is divided into two parts: the training dataset and the testing dataset. Then, the SMOTE method is used to oversample the training dataset. This is done to balance the quantity of samples in the minority class. We use the balanced dataset to train a classification model that uses the XGBoost technique.

The concluding stage is to assess the model's efficacy utilizing the testing dataset. This assessment utilizes various essential parameter metrics for a thorough evaluation of the model's proficiency in properly classifying unbalanced data.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15029
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Tabel 2. Evaluation By Dataset

Dataset	Accuracy	Precision	Recall	F1-score	Support
Diabetes	0.75	0.79	0.76	0.77	154
Spambase	0.94	0.94	0.94	0.94	921

The evaluation results indicate that the XGBoost algorithm exhibits competitive classification performance on two separate datasets. The Diabetes dataset yielded a model accuracy of 0.76, with precision at 0.79, recall at 0.76, and an F1-score of 0.77, demonstrating a favorable equilibrium between sensitivity and precision among imbalanced data. In contrast, the model attained optimal performance on the Spambase dataset, with accuracy and all major evaluation metrics reaching 0.94. This highlights XGBoost's strong capability in consistently handling class distinction, particularly in large-scale datasets. These findings affirm the robustness of XGBoost in managing class distribution complexity while maintaining high predictive accuracy (Liu et al., 2020).

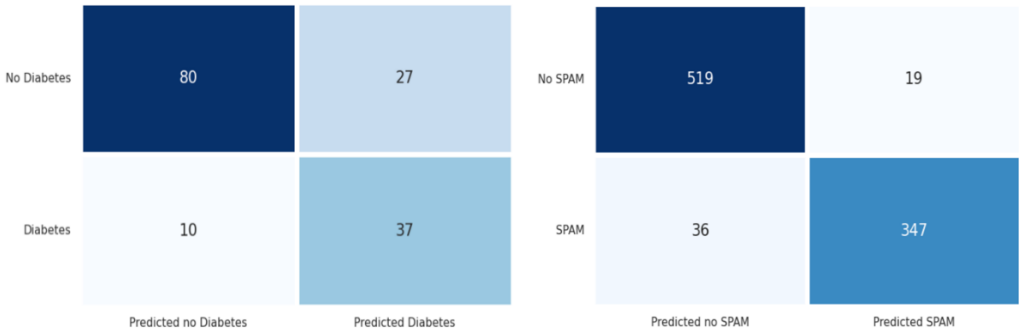

Figure 3. Confusion matrix

In the Spambase dataset, XGBoost demonstrated strong classification performance, accurately detecting 519 true negatives and 347 genuine positives, with minimal classification error rates. There are 19 false positives and 36 false negatives. This distribution highlights the model's ability to maintain a fair balance between sensitivity and specificity, as well as its effectiveness in managing a broad and complex class distribution. On the other hand, despite the smaller size of the diabetes dataset and the more complex clinical classification challenges it presents, the model still achieves competitive performance. The model demonstrates excellent sensitivity toward the minority class, with 80 true negatives, 37 true positives, 27 false positives, and 10 false negatives, while maintaining overall prediction accuracy.

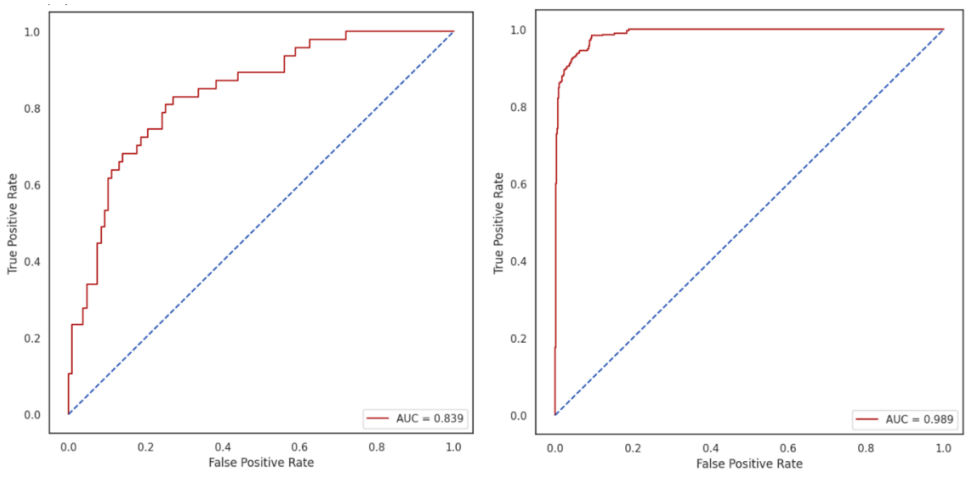

Figure 4. Curva Roc

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025
DOI : https://doi.org/10.33395/sinkron.v9i3.15029
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Where Figure 4. Roc curve analysis confirms the discriminative ability of the model across the evaluated datasets. For the Diabetes dataset, an AUC value of 0.839 indicates a reasonably good capability in distinguishing between positive and negative classes, although there is still room for improvement, particularly in addressing class imbalance. In contrast, the Spambase dataset achieves an AUC of 0.989, which is very close to 1, reflecting near-perfect class separation. According to ROC theory, such high AUC values demonstrate strong model performance in minimizing false positives while maximizing true positive detections. Overall, these results validate the reliability and effectiveness of the XGBoost model, especially when supported by SMOTE-based oversampling techniques.

Tabel 3. Performance Comparison of ML Models

DATASET	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Diabetes	DecisionTree	0.72	0.54	0.68	0.60	0.71
	SVM	0.73	0.55	0.68	0.61	0.72
	GaussianNB	0.75	0.59	0.70	0.64	0.74
	Logistic Regression	0.74	0.56	0.66	0.61	0.72
	XGBoost	0.75	0.79	0.76	0.77	0.83
SpamBase	DecisionTree	0.90	0.88	0.90	0.89	0.91
	SVM	0.69	0.64	0.65	0.64	0.69
	GaussianNB	0.80	0.70	0.93	0.70	0.83
	Logistic Regression	0.91	0.89	0.92	0.90	0.92
	XGBoost	0.94	0.94	0.94	0.94	0.98

In the Diabetes dataset, the XGBoost model showed enhanced classification efficacy relative to alternative machine learning techniques, including Decision Tree, Support Vector Machine (SVM), Gaussian Naive Bayes, and Logistic Regression. This model attained an accuracy of 0.79 and a ROC-AUC of 0.83, the highest among all comparative models. The results demonstrate that our method is not only more precise in identifying positive cases (diabetic patients) but also exhibits superior discriminative capability in differentiating between classes, a vital factor in medical data analysis with imbalanced class distributions.

The XGBoost model demonstrated superior performance in the Spambase dataset, achieving a score of 94% across all major measures, including accuracy, precision, recall, and F1-score, along with a ROC-AUC value of 0.98, indicative of near-optimal performance. This accomplishment exceeds the AUC values of 0.92 and 0.91 attained by the logistic regression and decision tree models. In contrast, algorithms like SVM and Gaussian Naive Bayes exhibited diminished precision and recall metrics, indicating constraints in managing high-dimensional and unbalanced datasets. These findings underscore that the integration of XGBoost with SMOTE-based oversampling approaches can yield more dependable predictions and enhanced generalization in practical classification contexts.

DISCUSSIONS

This study focuses on addressing classification challenges in imbalanced datasets by integrating the SMOTE with the XGBoost algorithm. Evaluations were conducted on two distinct datasets, Diabetes and Spambase, representing the domains of clinical health and digital communication, respectively. Based on experimental outcomes and metric-based evaluations, several key discussion points are elaborated below:

Effectiveness of SMOTE and XGBoost Integration

*name of corresponding author

The application of SMOTE during the preprocessing stage significantly improved the representation of minority classes, particularly when combined with XGBoost. On the Spambase dataset, XGBoost achieved an F1-score of 0.94 and a ROC-AUC of 0.98, surpassing other models such as Decision Tree (F1: 0.89, AUC: 0.91) and SVM (F1: 0.64, AUC: 0.69). These findings suggest that the proposed integration not only improves class balance but also enhances the model's discriminative power in recognizing minority class patterns.

Model Generalization Across Domains

XGBoost also demonstrated consistent performance on the smaller and more complex Diabetes dataset, with a recall of 0.76 and an ROC-AUC of 0.83, outperforming SVM (recall: 0.68, AUC: 0.72) and logistic regression (recall: 0.66, AUC: 0.72). This suggests the model's robustness in generalizing across domains, even under data-limited and clinically sensitive scenarios.

ROC Curve Analysis and Classification Accuracy

The ROC curves revealed strong discriminative ability of the XGBoost model in distinguishing between positive and negative classes. In the Spambase dataset, the curve closely approached the ideal upper-left corner (AUC: 0.989), indicating an optimal trade-off between true positive and false positive rates. This visualization supports the validity of the previously calculated numerical metrics.

Model Comparison and Performance Stability

Compared to models such as Gaussian Naive Bayes and Decision Tree, XGBoost consistently provided the best balance among accuracy, precision, recall, and F1-score across both datasets. For example, GaussianNB exhibited high recall but low precision (Spambase Recall: 0.93, Precision: 0.70), suggesting a tendency toward false positives. The decision tree, on the other hand, displayed signs of overfitting on certain dataset structures, as indicated by its inconsistent cross-dataset performance.

Limitations and Opportunities for Improvement

While SMOTE improves model performance, it relies on linear interpolation among nearest neighbors, which can result in overly homogeneous synthetic samples. This may lead to overfitting, particularly on datasets with complex structures or high variability. To mitigate this, future studies could explore advanced resampling techniques such as Borderline-SMOTE, ADASYN, or hybrid methods like SMOTE-ENN to enhance the diversity of synthetic data while maintaining classification accuracy.

Contribution to the Literature on Imbalanced Data Classification

This study validates the efficacy of XGBoost as a classification model that achieves superior performance, particularly when utilized on data that is imbalance by suitable oversampling methods. The primary methodological contribution is the demonstration that the predictive quality for minority classes is influenced not only by the algorithm's efficacy but also by data preprocessing procedures that address class distribution imbalances. In this context, approaches like SMOTE are essential for achieving a more balanced data representation, thereby enhancing XGBoost's capacity to effectively differentiate between classes. This study underscores the significance of employing evaluation metrics that are responsive to class distribution.

*name of corresponding author

CONCLUSION

This study integrates the SMOTE with the XGBoost algorithm to address classification challenges on datasets with imbalanced class distributions. Evaluations were conducted across two distinct domains, diabetes and spambase, to assess the model's consistency and generalizability.

This study's findings demonstrate that the integration of the SMOTE technique with the XGBoost model markedly enhances classification performance, especially for the F1-score and ROC-AUC metrics, which are responsive to class distribution. The created model attained an AUC value of 0.98 on the Spambase dataset and 0.83 on the Diabetes dataset, indicating its proficiency in properly identifying occurrences from minority classes. In comparison to conventional machine learning methods like decision trees, support vector machines (SVM), and logistic regression, the XGBoost-based method regularly exhibits enhanced performance for overall accuracy and sensitivity to class imbalance.

Moreover, the use of 10-fold cross-validation improved the reliability of performance estimation and mitigated potential bias due to random data partitioning. This study also underscores the importance of selecting appropriate evaluation metrics, as accuracy alone may be misleading in imbalanced data scenarios.

SMOTE's generation of synthetic samples via linear interpolation between minority neighbors can produce points that invade majority class regions, leading to label noise and ambiguity, especially in datasets with overlapping or non-linear class boundaries. The assumption of uniform interpolation fails in the presence of minority class sub-clusters (small disjuncts), causing synthetic samples to blur distinct clusters and reduce intra-class variability representation. As a result, models may fit superficial patterns rather than the true decision boundary, increasing overfitting risk and diminishing generalization capabilities. (Elreedy & Atiya, 2019). Therefore, future work is encouraged to explore more advanced techniques such as Borderline-SMOTE or SMOTE-ENN. In conclusion, integrating SMOTE with XGBoost constitutes a robust and adaptable approach for real-world classification tasks, particularly those requiring high sensitivity to minority instances. Future research may benefit from incorporating more advanced oversampling strategies to address the current limitations in synthetic diversity.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to our co-authors and colleagues for their valuable contributions to this article. We hope that the insights presented herein will be beneficial to readers and practitioners alike, and may serve as a useful reference for institutions seeking to implement similar approaches.

REFERENCES

Abdul Bujang, S. D., Selamat, A., Krejcar, O., Mohamed, F., Cheng, L. K., Chiu, P. C., & Fujita, H. (2023). Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review. In *IEEE Access* (Vol. 11, pp. 1970–1989). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2022.3225404

Adeoti Babajide Ebenezer1, B. O. K. (PhD) 2 , O. M. I. (2021). A Comprehensive Analysis of Handling Imbalanced Dataset. *International Journal of Advanced Trends in Computer Science and Engineering*, *10*(2), 454–463. https://doi.org/10.30534/ijatcse/2021/031022021

*name of corresponding author

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503

Devan, P., & Khare, N. (2020). An efficient XGBoost–DNN-based classification model for network intrusion detection system. *Neural Computing and Applications*, *32*(16), 12499–12514. https://doi.org/10.1007/s00521-020-04708-x

Douzas, G., & Bacao, F. (2019). Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. *Information Sciences*, *501*, 118–135. https://doi.org/10.1016/j.ins.2019.06.007

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

Halim, A. M., Dwifebri, M., & Nhita, F. (2023). Handling Imbalanced Data Sets Using SMOTE and ADASYN to Improve Classification Performance of Ecoli Data Sets. *Building of Informatics, Technology and Science (BITS)*, *5*(1). https://doi.org/10.47065/bits.v5i1.3647

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

Hasanah, U., Soleh, A. M., & Sadik, K. (2024). Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models. *Jurnal Matematika, Statistika Dan Komputasi*, *21*(1), 88–102. https://doi.org/10.20956/j.v21i1.35552

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). https://doi.org/10.3390/a18010037

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition*.

Jiang, J., Zhang, C., Ke, L., Hayes, N., Zhu, Y., Qiu, H., Zhang, B., Zhou, T., & Wei, G. W. (2025). A review of machine learning methods for imbalanced data challenges in chemistry. In *Chemical Science* (Vol. 16, Issue 18, pp. 7637–7658). Royal Society of Chemistry. https://doi.org/10.1039/d5sc00270b

Kim, B., & Kim, J. (2020). Adjusting decision boundary for class imbalanced learning. *IEEE Access*, *8*, 81674–81685. https://doi.org/10.1109/ACCESS.2020.2991231

Liu, Y., Li, X., Chen, X., Wang, X., & Li, H. (2020). High-Performance Machine Learning for Large-Scale Data Classification considering Class Imbalance. *Scientific Programming*, *2020*. https://doi.org/10.1155/2020/1953461

Nemade, B., Bharadi, V., Alegavi, S. S., & Marakarkandy, B. (n.d.). International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons. In *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE* (Vol. 2023, Issue 9s). www.ijisae.org

Nobre, J., & Neves, R. F. (2019). Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets. *Expert Systems with Applications*, *125*, 181–194. https://doi.org/10.1016/j.eswa.2019.01.083

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 3, July 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i3.15029 p-ISSN : 2541-044X

Sharma, A., Singh, P. K., & Chandra, R. (2022). SMOTified-GAN for Class Imbalanced Pattern Classification Problems. *IEEE Access*, *10*, 30655–30665. https://doi.org/10.1109/ACCESS.2022.3158977

Suandi, F., Anam, M. K., Firdaus, M. B., Fadli, S., Lathifah, L., Yumami, E., Saleh, A., & Hasibuan, A. Z. (2024). *Enhancing Sentiment Analysis Performance Using SMOTE and Majority Voting in Machine Learning Algorithms* (pp. 126–138). https://doi.org/10.2991/978-94-6463-620-8_10

Sugihartono, T., Wijaya, B., Marini, Alkayes, A. F., & Anugrah, H. A. (2025). Optimizing Stunting Detection through SMOTE and Machine Learning: a Comparative Study of XGBoost, Random Forest, SVM, and k-NN. *Journal of Applied Data Sciences*, *6*(1), 667–682. https://doi.org/10.47738/jads.v6i1.494

Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Information Fusion*, *54*, 128–144. https://doi.org/10.1016/j.inffus.2019.07.006

Ujaran, K., Ridwan, K., Heni Hermaliani, E., Ernawati, M., & Author, C. (2024). Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada. In *Computer Science (CO-SCIENCE* (Vol. 4, Issue 1). http://jurnal.bsi.ac.id/index.php/co-science

Verbakel, J. Y., Steyerberg, E. W., Uno, H., De Cock, B., Wynants, L., Collins, G. S., & Van Calster, B. (2020). ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *Journal of Clinical Epidemiology*, *126*, 207–216. https://doi.org/10.1016/j.jclinepi.2020.01.028

Wang, C., Deng, C., & Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognition Letters*, *136*, 190–197. https://doi.org/10.1016/j.patrec.2020.05.035

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-03430-5

Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics*, *107*. https://doi.org/10.1016/j.jbi.2020.103465

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost. *IEEE Access*, *6*, 21020–21031. https://doi.org/10.1109/ACCESS.2018.2818678

*name of corresponding author