Enhancing SMOTE Using Euclidean Weighting for Imbalanced Classification Dataset

Nur Ghaniaviyanto Ramadhan[1], Warih Maharani[2], Alfian Akbar Gozali[3], Adiwijaya[4,*]

[1,2]*Department of Data Science, School of Computing, Telkom University, Bandung, Indonesia*

[3]*Department of Software Engineering, School of Applied Science, Telkom University, Bandung, Indonesia*

[4]*Department of Informatics, School of Computing, Telkom University, Bandung, Indonesia*

(Received: January 20, 2025; Revised: February 28, 2025; Accepted: May 30, 2025; Available online: July 19, 2025)

Abstract

Class imbalance is a significant challenge in machine learning classification tasks because it often causes models to be biased toward the majority class, resulting in poor detection of minority classes. This study proposes a novel enhancement to the Synthetic Minority Over-sampling Technique (SMOTE) by incorporating Euclidean distance-based feature weighting, called Weighted SMOTE. The key idea is to improve the quality of synthetic minority samples by calculating feature importance using a Random Forest model and assigning higher weights to the most relevant features. The objective of this research is to generate more representative synthetic data, reduce model bias, and increase predictive accuracy on highly imbalanced datasets. Experiments were conducted on four benchmark datasets from the KEEL Repository with imbalance ratios ranging from 0.013 to 0.081. The proposed Weighted SMOTE combined with an ensemble voting classifier (Random Forest, AdaBoost, and XGBoost) demonstrated significant improvements compared to standard SMOTE and models without resampling. For example, on the Zoo-3 dataset, the Balanced Accuracy Score (BAS) increased from 75% to 90%, while the F1-score improved from 48% to 94%. On the Cleveland-0_vs_4 dataset, precision improved from 83% to 91% and recall remained high at 99%. Statistical testing using the Wilcoxon signed-rank test confirmed these improvements with p-values < 0.05 for key metrics. The findings show that the proposed method effectively balances sensitivity and precision, generates more meaningful synthetic samples, and reduces the risk of overfitting compared to conventional oversampling. The novelty of this work lies in integrating Euclidean-based feature weighting into the SMOTE process and validating its performance on multiple domains with varying feature types and imbalance ratios. These results indicate that the proposed Weighted SMOTE approach contributes a practical solution for improving classification performance and model stability on severely imbalanced data.

Keywords: Imbalanced Data, Weighted SMOTE, Euclidean Distance, Feature Weighting, Ensemble Voting

1. Introduction

Class imbalance is one of the major challenges encountered in classification tasks within the domain of data mining [1]. Class imbalance transpires when the quantity of samples in one class markedly surpasses that of the other classes [2]. This condition is commonly found in various domains, such as the detection of chronic diseases (e.g., cardiovascular disease, stroke, diabetes, cancer, and hypertension) [3], [4], [5] financial fraud detection [6], and spam email classification [7]. In some of these studies, it has been shown that the SMOTE oversampling algorithm is superior to other oversampling algorithms such as SMOTE-Tomek, and Random Oversampling. Current machine learning models tend to be biased toward the majority class [8], resulting in poor performance when identifying the minority class [9].

To address the class imbalance problem, various techniques have been developed, one of which is the resampling method. Resampling is generally categorized into two main approaches: undersampling and oversampling. Undersampling diminishes the quantity of samples in the majority class, whereas oversampling amplifies the quantity of samples in the minority class [10]. However, oversampling may lead to the loss of important information due to data removal, whereas conventional oversampling methods, such as Random Oversampling (ROS), may increase the risk

*Corresponding author: Adiwijaya (adiwijaya@telkomuniversity.ac.id)

DOI: https://doi.org/10.47738/jads.v6i3.798

of overfitting as a result of duplicating minority class samples [11]. Due to the inherent drawbacks of basic resampling approaches, such as overfitting in ROS and information loss in undersampling, more advanced techniques have been explored to better address class imbalance without compromising data quality. One such widely adopted method is the SMOTE. The SMOTE is one of the most commonly used oversampling methods for addressing class imbalance problems [12]. SMOTE generates synthetic samples by performing linear interpolation between existing minority class samples [13]. Although effective in improving model performance on the minority class, SMOTE has several limitations, such as the suboptimal selection of synthetic samples and the potential generation of samples that do not align with the original data distribution. Additionally, the quality of resampled data may degrade when minority samples are located too far from their nearest neighbours or when neighbouring samples overlap with those from other classes [14], [15]. Several studies have proposed enhancements to SMOTE, such as Borderline-SMOTE, Adaptive Synthetic Sampling (ADASYN), and Safe-Level SMOTE. However, these methods still face challenges in preserving the intrinsic distribution characteristics of the minority class [16], [17].

One approach to enhance the effectiveness of SMOTE is by applying Euclidean distance-based weighting in the synthetic sample generation process [9]. Euclidean weighting helps in selecting more representative pairs of minority samples, allowing the distribution of the generated samples to better reflect the natural pattern of the data. This approach is expected to improve SMOTE by producing a more balanced dataset without compromising the essential characteristics of the minority class.

In order to address the limitations of conventional SMOTE, this study focusses on the creation and assessment of an improved SMOTE technique that employs Euclidean-based feature weighting. The suggested approach specifically seeks to address SMOTE's drawback of treating every feature equally throughout the synthetic sample creation process, which may lead to samples that are overlapping or poorly representative. In contrast to other weighted methods like FW-SMOTE, which use mutual information to determine feature relevance, our technique combines a normalised Euclidean distance framework with Random Forest-based feature importance. This method guarantees that more important contributions to the interpolation process are provided by more influential features. The suggested method's performance is assessed by looking at how well it can increase classification accuracy, particularly on the minority class, on a number of imbalanced datasets with different imbalance ratios. The study's findings should lead to more efficient and useful approaches to processing imbalanced data in fields including text classification, anomaly detection, and chronic disease prediction.

2. Literature Review

One of the key obstacles to creating precise and trustworthy classification models is still the problem of unbalanced data. When one class is substantially more numerous than the others, biassed model performance results, especially when the minority class is not well recognised. In real-world applications where precise identification of uncommon occurrences is essential, such as financial fraud detection [13] and chronic disease prediction [6], such imbalance commonly occurs. Alkhawaldeh et al. [14] have pointed out that imbalance frequently leads models to overlook minority events, which increases misclassification in critical applications like anomaly detection and disease diagnosis.

Researchers have looked into a number of resampling strategies to address this problem, including oversampling strategies like Chawla's SMOTE [18]. SMOTE enriches the minority class by using linear interpolation to create fresh minority samples. Nevertheless, research like [19] has shown that traditional SMOTE has drawbacks, such as the propensity to generate overlapping or unrepresentative synthetic samples, particularly in areas where class boundaries are ambiguous.

A number of SMOTE variations, such as SMOTE-Tomek, Borderline-SMOTE, and ADASYN, have been proposed to overcome these restrictions. Even while these techniques are better, they still have problems including overfitting and producing noisy samples that don't accurately represent the minority class's distribution. Furthermore, these methods frequently fail to take into consideration the significance of specific traits while creating synthetic samples.

Feature-Weighted SMOTE (FW-SMOTE) is a new approach that enhances the generating process by including feature relevance [20]. By applying statistical methods such as mutual information, Fisher score, or L1-regularization to assign weights to features, FW-SMOTE enables more significant characteristics to have a bigger impact during interpolation. The lack of model-based importance in this approach, however, may limit its applicability to a variety of datasets.

Simultaneously, ensemble learning methods have been applied to enhance classification resilience on unbalanced datasets, including voting ensembles. Compared to single classifiers, these methods generate more dependable results by combining predictions from models such as Random Forest, XGBoost, and AdaBoost [21]. Ensemble approaches boost performance, however they usually only work at the prediction level and don't raise the calibre of the training data.

The use of model-driven feature importance to direct resampling is a more recent innovation. Features are frequently ranked according to relevance using Random Forest, which is well-known for its resilience in evaluating feature contribution to prediction [22]. By giving influential features more weights during interpolation, these importance scores can be incorporated into SMOTE. A cumulative thresholding strategy, in which the top 80% of cumulative feature importance is deemed relevant, is frequently used to accomplish this [23]. This method preserves all dimensions in the synthetic data while lessening the impact of noisey or irrelevant elements.

SMOTE has to be improved, according to a number of earlier studies. According to Ramadhan et al. [13], SMOTE performed better on datasets related to chronic diseases, but it lost its reliability when feature distributions were erratic. Generalisation is hampered by class overlap problems and sensitivity to outliers, as noted by Alkhawaldeh et al. [14]. Fahrudin et al. [15] investigated attribute weighting, however instead of incorporating feature importance from model insights, they only used statistical weights.

To overcome these drawbacks, the approach presented in this work combines Euclidean distance weighting and Random Forest-based feature importance to direct SMOTE interpolation. This method makes it possible to create synthetic samples that are more feature-aware, representative, and in line with the minority class's underlying distribution. Additionally, the suggested approach is tested alongside ensemble voting, which has been demonstrated in [15] and [21] to improve overall performance measures, especially recall, precision, and Balanced Accuracy Score (BAS), when combined with strong resampling techniques.

Deep generative models, in particular Generative Adversarial Networks (GANs), have emerged as potent instruments for creating synthetic data as a result of recent developments in the field of imbalanced learning. Studies like those by [24] have shown that conditional GANs (cGANs) outperform oversampling variants in complicated datasets and can represent the underlying distribution of minority classes, particularly in high-dimensional spaces. These models are appropriate for oversampling situations where linear interpolation might not be adequate since they adaptively learn the data manifold.

Transformer-based designs have lately been used for feature importance analysis and weighting in addition to generative models. According to research by [25], transformers can dynamically assign relevance scores to input features across instances by utilising self-attention techniques. These techniques allow for context-aware feature selection, which is especially helpful for datasets with textual imbalances or temporal series. The incorporation of transformer-driven dynamic weighting may be a promising avenue for future improvements of the Weighted SMOTE architecture, even though our method uses Random Forest-based static importance.

In addition to model improvements, recent research has also emphasized the importance of statistical significance testing to validate the effectiveness of SMOTE and its variants. Rather than relying solely on point estimates such as accuracy or F1-score, studies have begun employing non-parametric tests like the Wilcoxon signed-rank test or Friedman test to determine whether observed performance differences are statistically meaningful across multiple datasets and classifiers [26].

3. Methodology

The overall flow of the proposed Weighted SMOTE methodology is illustrated in figure 1, which outlines the stages from dataset preprocessing to synthetic sampling and ensemble classification. The process begins with an imbalanced dataset. In the second step, the system identifies the most influential features within the dataset. This analysis helps determine which features contribute the most to the model’s predictions. The third step involves calculating feature weights using a mathematical formula. These weights are then integrated into the SMOTE algorithm. In the fourth step, Weighted SMOTE is applied to generate synthetic samples based on the computed feature weights, making the synthetic data more representative of the original data distribution [9]. The fifth step involves applying an ensemble voting model to classify the imbalanced dataset. Finally, an analysis is conducted to compare the performance results across the different models.

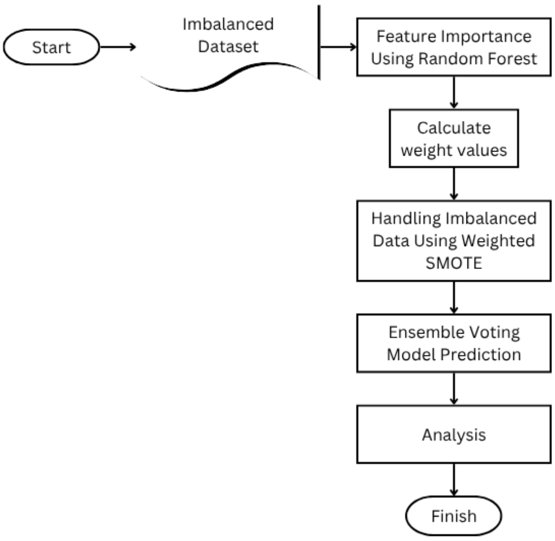
Start
Imbalanced Dataset
Feature Importance Using Random Forest
Calculate weight values
Handling Imbalanced Data Using Weighted SMOTE
Ensemble Voting Model Prediction
Analysis
Finish

Figure 1. Proposed Methodology System

3.1. Dataset

The imbalanced datasets used were obtained from the KEEL repository [27]. A dataset is considered imbalanced based on its Imbalance Ratio (IR) value. The IR values of the datasets vary within the range of 0.013 to 0.081. The closer the IR value is to zero, the more severely imbalanced the dataset is. The IR can be calculated using Formula (1) [28].

$$IR = \frac{Number\ Class\ of\ Minority}{Number\ Class\ of\ Majority} \tag{1}$$

In the datasets used in this study, the minority class is labeled as positive, while the majority class is labeled as negative. Table 1 provides a description of the imbalanced datasets utilized in this research.

Table 1. KEEL Repository Dataset (Imbalanced)

Dataset	Number of Features	Number of Minority Label	Number of Majority Label	IR
Zoo-3	16	5	96	0.052
Cleveland-0_vs_4	14	13	160	0.081
Dermatology-6	35	20	338	0.059
Kddcup-buffer_overflow	42	30	2203	0.013

In the Zoo-3 dataset, all features are binary numerical (0/1), except for the "Legs" feature, which ranges from 0 to 8. In the Cleveland dataset, all features are numerical with a float64 data type. In the Dermatology dataset, most features are discrete numerical values (ranging from 0 to 3), except for the "age" feature, which is continuous. In the KDDCup dataset, most of the features are binary numerical, with fewer discrete numerical features.

In addition to the IR, the selected datasets represent a range of domains and feature complexities. Zoo-3 originates from zoological classification and contains primarily binary features. Cleveland-0_vs_4 is a medical dataset with continuous clinical attributes. Dermatology-6 involves discrete dermatological measurements, while KDDCup-buffer_overflow represents a cybersecurity intrusion detection problem with high-dimensional binary inputs. This variety enables a broader evaluation of the proposed method across different application contexts.

3.2. Features Importance

At this section, feature importance is calculated using the Random Forest algorithm (figure 2). Random Forest is capable of intrinsically evaluating the importance of each feature during the model training process [22]. This enables the identification of the most influential features for classification or prediction, thereby facilitating better data understanding and model simplification [29]. This process is useful for assigning weights to the dataset features, where more significant features are given higher weights compared to less significant ones. As a result, no features are eliminated from the dataset. However, despite its advantages, Random Forest-based feature importance is known to have several limitations. It may produce unstable importance rankings across different runs, exhibit bias toward continuous or high-cardinality features, and be affected by feature correlation, leading to misleading attributions in some cases [20], [30]. Nonetheless, Random Forest was selected in this study for its empirical robustness and its ability to model nonlinear relationships, particularly in imbalanced and high-dimensional datasets.

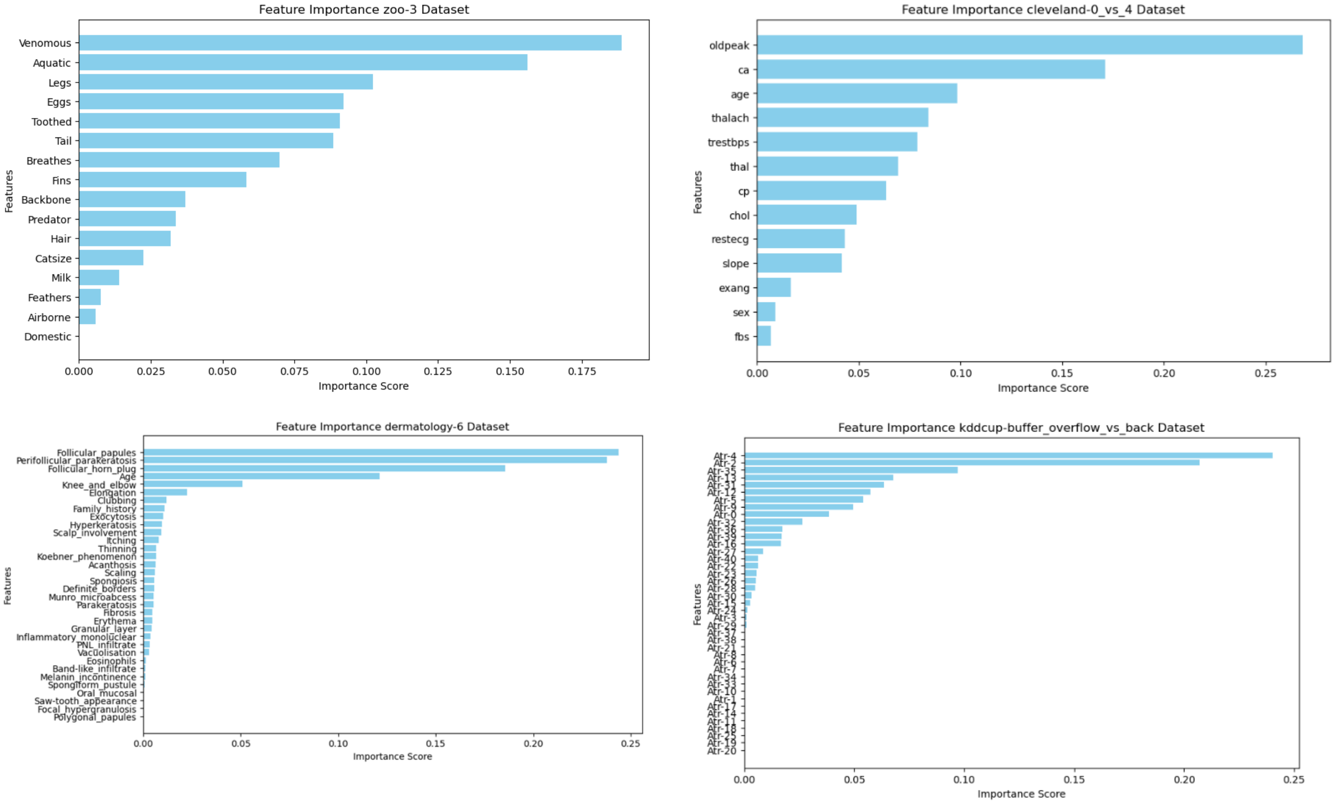

Figure 2. Features Importance Using Random Forest

The determination of significant and non-significant features is conducted using the variance threshold method via cumulative importance, also known as feature selection by cumulative contribution. This technique is categorized as a threshold-based feature selection method [31]. This technique selects features based on their cumulative contribution to the total feature importance, under the assumption that a small subset of features accounts for the majority of the model's predictive power [23].

The process is outlined as follows: Calculating Cumulative Importance. The importance values of each feature are summed cumulatively. This helps identify how many features are needed to reach X% of the total importance. Determining the Threshold. The threshold used set at 80% of the total importance. Features with the highest importance that cumulatively account for 80% of the total importance are considered the most significant. Filtering Features. The top 80% of features will be assigned higher weights compared to the remaining 20%, ensuring that all features in the dataset are retained. While no features are discarded in this study, the influence of less relevant or potentially noisy features is minimized through a weighting scheme based on feature importance. Features falling outside the top 80%

cumulative importance are assigned lower weights, which diminishes their contribution in the distance calculation and synthetic sample interpolation.

This study used an 80% cumulative importance threshold in this analysis to determine which traits were important and which were not. This threshold selection is based on a well-accepted approach in the literature on feature selection, which holds that the majority of the predictive value can be captured by the top 80% of feature contributions [23], [31]. While this fixed threshold may not be universally optimal across all datasets, it offers a practical balance between model simplicity and effectiveness. Figure 2 is the calculation of the value of the importance feature in each dataset used. Feature correlation and relevance are reflected through the feature importance scores shown in figure 3, which serve as the basis for determining feature weights in the Weighted SMOTE process. Meanwhile, features that are given greater weight based on cumulative feature importance can be seen in table 2.

Table 2. Significant Features All Dataset

Dataset	Feature	Importance	Cumulative Importance
Zoo-3	venomous	0.188992	0.188992
	aquatic	0.156031	0.345024
	legs	0.102234	0.447257
	eggs	0.092079	0.539336
	toothed	0.090790	0.630126
	tail	0.088624	0.718750
	breathes	0.069906	0.788657
Cleveland-0_vs_4	oldpeak	0.268352	0.268352
	ca	0.171163	0.439515
	age	0.098401	0.537916
	thalach	0.084404	0.622319
	trestbps	0.078966	0.701286
	thal	0.069160	0.770446
Dermatology-6	Follicular_papules	0.243640	0.243640
	Perifollicular_parakeratosis	0.237713	0.481353
	Follicular_horn_plug	0.185680	0.667034
	age	0.121163	0.788196
KDDCup-buffer_overflow	Atr-4	0.240158	0.240158
	Atr-2	0.207189	0.447347
	Atr-35	0.097218	0.544565
	Atr-13	0.067681	0.612247
	Atr-31	0.063522	0.675769
	Atr-12	0.057264	0.733033
	Atr-5	0.054010	0.787042

3.3. Calculate Weight Value

The weight values are calculated using the following formula [9]. In this study, there are differences in terms of determining significant features. The difference is that in this study significant features were determined using feature importance, on the contrary, in the previous study the determination was based on an expert [9].

$$\mathrm{W\alpha = K\,x\,\alpha} \tag{2}$$

$$\mathrm{W\beta = K\,x\,\beta} \tag{3}$$

$$\left(\sum_{\mathrm{SFFI}} \mathrm{x\ W\alpha}\right) + \left(\sum_{\mathrm{Non-SFFI}} \mathrm{x\ W\beta}\right) \tag{4}$$

∑ Significant Features by Features Importance (SFFI) denotes the count of features in the GCU dataset that are considered significant based on features importance. ∑ Non-Significant Features by features importance refers to the count of features in the GCU dataset that are deemed non-significant according to features importance assessment. K is a fundamental constant used to normalize the total weight to 1. Before computing the values of Wα and Wβ, the value of K must first be determined. α is a constant that establishes the relative weight differentiation for significant features, whereas β is a constant that determines the relative weight differentiation for non-significant features. Wα represents the weight assigned to significant features, while Wβ represents the weight assigned to non-significant features. The weights assigned to significant and non-significant features are computed based on the following formulation: First, a normalization constant K is calculated as:

$$\mathrm{K} = \frac{1}{\alpha\,.\,\mathrm{S} +\ \beta\,.\,\mathrm{N}} \tag{5}$$

S is the number of significant features (top 80% cumulative importance), N is the number of non-significant features (bottom 20%), α and β are scaling constants (e.g., α=3, β=1). Then, the weights are defined as formula (2) and (3). These weights ensure that the total contribution of all features is normalized to 1:

$$\mathrm{S\,.\,W\alpha + N\,.\,W\beta = 1} \tag{6}$$

This mechanism emphasizes features deemed more predictive (via feature importance) while preserving all features in the synthetic sample generation process.

3.4. Handling Imbalanced Data

Class imbalance in a dataset can cause machine learning models to be more accurate in classifying the majority class but less effective in recognizing the minority class [32]. Therefore, this study applies the SMOTE enhanced with Euclidean distance-based weighting (Euclidean Weighting).

To address data imbalance, this study compares three main approaches: Without SMOTE: The model is trained directly on the imbalanced dataset without applying any resampling techniques. Standard SMOTE: The conventional SMOTE technique is used to generate synthetic samples based on linear interpolation between existing minority class samples. Weighted SMOTE: A novel approach proposed in this study, which applies SMOTE with Euclidean weighting [9]. This technique calculates feature weights based on the Euclidean distance between minority samples and their nearest neighbours, ensuring that the generated synthetic samples better represent the distribution of the minority class.

The Euclidean weighting formula is defined as follows:

$$Euclidean\ distance\ i,j = \sqrt{\sum \left(\frac{x_j - x_i}{weights}\right)^2} \tag{7}$$

$x_j$and x_i represent the coordinates of two points located within the same space. The difference between x_j and x_i refers to the difference in the corresponding coordinates of points j and i. The weights refer to the values assigned to each coordinate difference, which are obtained from formulas (2) to (4). These weights control the influence of each dimension on the total distance.

In this study, the standard Euclidean distance formula is modified by incorporating feature-specific weights derived from feature importance scores. The modified distance is defined as:

$$d(x, x') = \sum_{i=1}^{n} w_i\,.\,(x_i - x'_i)^2 \tag{8}$$

w_i denotes the weight of the ith feature, obtained from Random Forest feature importance. $d(x, x')$ is the weighted Euclidean distance between two minority data samples x_i and x'_i. n is the number of features. x_i is the value of the ith feature in the first data sample x. x'_i is the value of the i-i feature in the second data sample x'. w_i the weight of the ith feature, which indicates the level of importance of that feature in the classification. This formulation ensures that features with higher predictive relevance contribute more significantly to the distance calculation. The interpolation process in SMOTE is thus guided to generate synthetic samples that are more representative of the actual minority class distribution. This approach is inspired by previous studies such as FW-SMOTE by Maldonado et al. [20], which demonstrated that feature-weighted distance metrics can enhance oversampling effectiveness in imbalanced classification problems. Ilustrative Example. Consider two minority class samples: x=[2,4,6], x′=[3,2,9]. Let the feature weights derived from feature importance be: w=[0.5, 0.3, 0.2] . The weighted Euclidean distance is calculated as:

$$d(x, x') = \sqrt{0.5(2-3)^2 + 0.3(4-2)^2 + 0.2(6-9)^2} = 1.87 \tag{9}$$

A synthetic sample is then generated using linear interpolation:

$$\begin{aligned} X_{synthetic} &= x + \lambda(x' - x), \qquad \lambda = 0.5 \\ &= [2, 4, 6] + 0.5([1, -2, 3]) = [2.5, 3, 7.5] \end{aligned} \tag{10}$$

This process shows how feature weights influence the distance and consequently the interpolation direction and magnitude, ensuring that synthetic samples reflect the relative importance of each feature.

3.5. Ensemble Voting Model

In this study, a voting ensemble approach is employed by combining several previously tested algorithms, namely Random Forest, AdaBoost, and XGBoost. By utilizing this voting technique, the prediction results are expected to achieve higher accuracy by leveraging the strengths of each model, which possess distinct characteristics.

The steps for implementing the ensemble voting approach in this study are as follows: Training multiple classification models separately using the same dataset. Next step, applying soft voting to aggregate the prediction results from all models. Last, generating the final prediction outcome according to the results of the voting process. The voting ensemble serves as a crucial method in this study, as it mitigates the individual weaknesses of each model by combining their strengths, thereby enhancing the classification performance on the utilized dataset [21].

To ensure reproducibility, the ensemble models were trained using the following hyperparameters. For Random Forest, we used n_estimators=100, max_depth=None, and random_state=42. For AdaBoost, n_estimators=100, random_state=42, and learning_rate=1.0 were applied. The XGBoost classifier was configured with n_estimators=100, max_depth=None, random_state=42, and learning_rate=0.1, using logloss as the evaluation metric.

4. Results and Discussion

This section discusses the results and analysis obtained from the experiments. The classification model utilized is an ensemble voting method comprising three models: Random Forest (RF), AdaBoost, and XGBoost. The experiments were conducted by comparing the performance of three scenarios: without applying SMOTE, using the standard SMOTE library, and using SMOTE with feature weighting based on Euclidean distance. The objective of these experiments is to assess the significant differences in classification outcomes. The evaluation metrics employed include precision, recall, F1-score, and Balanced Accuracy Score (BAS), as these metrics are suitable and unbiased for addressing the imbalanced data problem [13].

4.1. Result of the Zoo-3 Dataset

This section shows the results of experiments for the Zoo-3 dataset. Table 3 is a comparison of the results obtained.

Table 3. Results from Zoo-3 Dataset

Model	Precision (%)	Recall (%)	F1-Score (%)	BAS (%)	ROC-AUC (%)
Without SMOTE	45	50	48	75	75
With SMOTE Library	98	75	82	75	95

Model	Precision (%)	Recall (%)	F1-Score (%)	BAS (%)	ROC-AUC (%)
With Weighted SMOTE	99	90	94	90	98

This dataset is highly unbalanced (a minority of about 5%), which causes models without imbalance handling to perform poorly. Weighted SMOTE is the best method, as it not only improves recall but also provides a better balance between precision and recall. BAS that increased from 75% (without SMOTE) to 90% (Weighted SMOTE) indicates success in overcoming class imbalances. The SMOTE Weighted approach is highly recommended for this dataset because it provides a more accurate classification of minority classes without sacrificing majority classes.

4.2. Result of the Cleveland-0_vs_4 Dataset

This section shows the results of the experiments for the Cleveland-0_vs_4 dataset. Table 4 is a comparison of the results obtained.

Table 4. Results from Cleveland-0_vs_4 Dataset

Model	Precision (%)	Recall (%)	F1-Score (%)	BAS (%)	ROC-AUC (%)
Without SMOTE	83	99	89	72	95
With SMOTE Library	83	99	89	99	98
With Weighted SMOTE	91	99	94	99	100

Without SMOTE precision (83%) and recall (99%) the model suggests the model can identify the minority class well, despite the possibility of a majority-class prediction error. BAS (72%) is low, indicating an imbalance of classification. Using SMOTE library precision remained at 83%, recall remained at 99%, but BAS increased drastically to 99%. This shows that SMOTE has succeeded in making the model more balanced in recognizing the two classes. Using weighted SMOTE, the accuracy increased to 91%, indicating an increase in the model's ability to avoid false positives. The F1-score increased to 94%, which means the balance between precision and recall is better. BAS remains 99%, indicating the model successfully overcomes the imbalance.

4.3. Result of the Dermatology-6 Dataset

This section presents the results of the experiment for the Dermatology-6 dataset. Table 5 is a comparison of the results obtained.

Table 5. Results from Dermatology-6 Dataset

Model	Precision (%)	Recall (%)	F1-Score (%)	BAS (%)	ROC-AUC (%)
Without SMOTE	100	90	94	100	100
With SMOTE Library	100	100	100	100	100
With Weighted SMOTE	100	100	100	100	100

The model works very well in predicting the majority of classes, but there are indications that some of the class minorities (minor classes) are not well classified, which is common in imbalance datasets. Once SMOTE is implemented, the model can classify all classes perfectly. This suggests that the initial problem is data imbalance. In this case, Weighted SMOTE does not provide any additional advantages over standard SMOTE, because perhaps the data imbalance is not too extreme or the SMOTE is quite optimal. On the other hand, the data type in the dataset is mostly binary (0/1) which is very optimal for the use of oversampling algorithms.

4.4. Result of the Kddcup-buffer_overflow_vs_back Dataset

This section presents the experimental results for the Kddcup-buffer_overflow_vs_back dataset. Table 6 provides a comparison of the obtained results.

Table 6. Results from Kddcup-buffer_overflow_vs_back Dataset

Model	Precision (%)	Recall (%)	F1-Score (%)	BAS (%)	ROC-AUC (%)
Without SMOTE	100	100	100	100	100
With SMOTE Library	100	100	100	100	100
With Weighted SMOTE	100	100	100	100	100

SMOTE and SMOTE Weighted produce stable values. This happens because the values in the binary dataset (0/1) or the model are very easy to distinguish between two classes: buffer_overflow and back. This dataset has very informative features or cases that are very clearly separated between buffer_overflow and back, making it easier for the model to make predictions with perfect accuracy.

4.5. Discussion

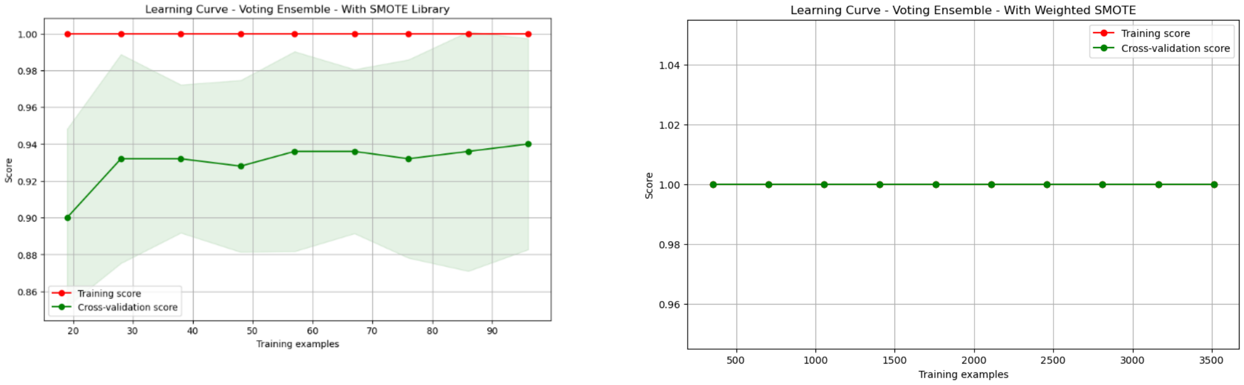

Figure 3. Learning Curve

Based on the learning curve in figure 3, the proposed Weighted SMOTE model demonstrates a reduced risk of overfitting compared to the conventional (library-based) SMOTE model. A common sign of overfitting is a significant gap between the training and validation performances, where the model performs well on training data but fails to generalize to unseen data. Weighted SMOTE addresses this issue by generating synthetic samples that are more representative of the minority class distribution, thus improving generalizability. This happened due to the binary nature of the data and the small amount of data [33]. We have re-examined the training and testing data by conducting several splitting scenarios. The results show that the lines produced are indeed parallel. We guarantee that the results are not overfitting because after testing with a variety of numerical data types and a sufficient amount of data, the lines are not parallel.

The Wilcoxon test (table 7) was used to evaluate whether there was a statistically significant difference between model performance [34]. This study used in three scenarios: W1: Without SMOTE vs With SMOTE Library. W2: With SMOTE Library vs With Weighted SMOTE. W3: Without SMOTE vs With Weighted SMOTE. Wilcoxon's test results showed that Weighted SMOTE consistently provided statistically significant performance improvements over standard SMOTE and models without oversampling, especially on the Precision, Recall, F1-Score, and ROC-AUC metrics ($p < 0.05$). The only exception was in the BAS metric between Without SMOTE and SMOTE Library, where no significant differences were found ($p = 0.625$). These findings reinforce the effectiveness of the Weighted SMOTE method in handling highly unbalanced datasets. Rejected means that the proposed model experienced significant improvement.

Table 7. Statistical Testing

Confusion matrix	Wilcoxon W1: Without vs SMOTE (p-value)	Hypothesis W1	Wilcoxon W2: SMOTE vs Weighted (p-value)	Hypothesis W2	Wilcoxon W3: Without vs Weighted (p-value)	Hypothesis W3
Precision	0.002	Rejected	0.002	Rejected	0.002	Rejected

Recall	0.002	Rejected	0.002	Rejected	0.002	Rejected
F1-Score	0.002	Rejected	0.002	Rejected	0.002	Rejected
BAS	0.625	Accepted	0.002	Rejected	0.002	Rejected
ROC-AUC	0.002	Rejected	0.002	Rejected	0.002	Rejected

The experimental results confirm that the Weighted SMOTE approach consistently enhances the performance of classification models on highly imbalanced datasets. Notable improvements are observed in the BAS and F1-Score, indicating the method’s effectiveness in balancing sensitivity to the minority class while maintaining overall accuracy. In datasets with extreme imbalance ratios—such as Zoo-3 (IR = 0.052)—the application of Weighted SMOTE increased the recall for the minority class from 50% to 90%, and the BAS from 75% to 90%. These results highlight the effectiveness of Euclidean-based weighting in generating more informative and targeted synthetic samples, which improves the model’s ability to detect previously underrepresented classes. A similar enhancement was seen on the Cleveland-0_vs_4 dataset, where precision improved from 83% to 91% while maintaining a high recall of 99%. This reflects the proposed method’s ability to increase detection of minority classes without significantly inflating the false positive rate in the majority class.

On the other hand, for datasets such as Dermatology-6 and Kddcup-buffer_overflow, both standard SMOTE and Weighted SMOTE yielded near-identical results, with all performance metrics (precision, recall, F1-score, BAS) reaching 100%. This suggests that in datasets with dominant binary or easily separable features, the benefits of advanced oversampling methods like Weighted SMOTE are less pronounced. These findings imply that the effectiveness of the Weighted SMOTE technique is influenced by the characteristics of the dataset, particularly the complexity and type of features. Greater performance gains were observed on datasets with continuous numerical features or complex decision boundaries (e.g., Cleveland-0_vs_4, Zoo-3), whereas simpler binary-featured datasets (e.g., Kddcup-buffer_overflow) showed marginal improvement.

Additionally, incorporating ensemble voting using Random Forest, AdaBoost, and XGBoost contributed to stable and consistent results across all experiments. Weighted SMOTE improved the quality of input data before being processed by these ensemble models, aligning with existing literature that supports the robustness of ensemble learning in handling class imbalance. Finally, Weighted SMOTE reduces the overfitting risk often associated with conventional oversampling methods. By incorporating Euclidean-based feature weighting, synthetic sample generation becomes more focused on relevant attributes, leading to more natural and realistic data augmentation. However, while this approach mitigates the influence of irrelevant features, the use of all features in resampling may still introduce some noise.

5. Conclusion

This study introduces a novel Weighted SMOTE method based on Euclidean distance and feature importance to improve classification performance on highly imbalanced datasets. By incorporating feature-specific weights derived from Random Forest importance scores, the proposed approach generates synthetic samples that better reflect the underlying distribution of the minority class. Experimental results across four benchmark datasets—Zoo-3, Cleveland-0_vs_4, Dermatology-6, and KDDCup-buffer_overflow—demonstrate consistent improvements in performance metrics such as precision, recall, F1-score, BAS, and ROC-AUC.

Interestingly, the suggested approach produced notable improvements on highly imbalanced datasets like Zoo-3 and Cleveland-0_vs_4, with statistically significant benefits confirmed by the Wilcoxon signed-rank test ($p < 0.05$). When compared to both normal SMOTE and models without oversampling, these findings demonstrate the Weighted SMOTE approach's resilience. Performance stability was further enhanced by the addition of ensemble learning utilizing Random Forest, AdaBoost, and XGBoost. The learning curve analysis further demonstrated that the suggested approach lessens overfitting by generating more varied and representative synthetic samples. Feature weighting reduces the impact of less important traits while maintaining all features in Weighted SMOTE.

The Dermatology-6 and KDDCup datasets demonstrate the method's minimal incremental value in datasets with highly separable binary features, despite these benefits. This suggests that feature distributions and domain complexity have an impact on the method's efficacy. Future work may focus on adaptive thresholding for feature significance, runtime optimization of the algorithm, comparative evaluation with other SMOTE variants (e.g., Borderline-SMOTE, ADASYN), and the integration of alternative feature attribution techniques such as SHAP or mutual information to further enhance generalizability.

6. Declarations

6.1. Author Contributions

Conceptualization: N.G.R., W\.M., A.A.G., and A.; Methodology: A.; Software: N.G.R.; Validation: N.G.R., A., and A.A.G.; Formal analysis: N.G.R., A., and A.A.G.; Investigation: N.G.R.; Resources: A.; Data curation: A.; Writing original draft preparation: N.G.R., A., and A.A.G.; Writing review and editing: A., N.G.R., and A.A.G.; Visualization: N.G.R.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The Ministry of Education, Culture, Research, and Technology, Republic of Indonesia, under Grant 030/LIT07/PPM-LIT/2024, the Directorate of Research, Technology, and Community Service, the Directorate General of Higher Education, Research, and Technology, and Telkom University all provided partial funding for this work.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

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

References

[1] S. Goswami and A. K. Singh, “A literature survey on various aspect of class imbalance problem in data mining,” *Multimed. Tools Appl.*, vol. 83, no. 27, pp. 70025–70050, Feb. 2024

[2] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, “Handling imbalanced medical datasets: review of a decade of research,” *Artif. Intell. Rev.*, vol. 57, no. 10, pp. 1-12, Sep. 2024.

[3] N. G. Ramadhan, Adiwijaya, W. Maharani, and A. Akbar Gozali, “Prediction of hypertension in the upcoming year: feature correlation analysis and handling imbalanced based on random forest,” in *2023 Eighth International Conference on Informatics and Computing (ICIC)*, vol. 2023, no. 1, pp. 1–6, 2023.

[4] N. G. Ramadhan, Adiwijaya, W. Maharani, and A. Akbar Gozali, “Prediction of cardiovascular disease (CVD) in the upcoming year using tree-based ensemble model,” in *12th International Conference on Information and Communication Technology (ICOICT)*, vol. 2024, no. 1, pp. 210–216, 2024.

[5] N. G. Ramadhan, Adiwijaya, W. Maharani, and A. Akbar Gozali, “Chronic diseases prediction using machine learning with data preprocessing handling: a critical review,” *IEEE Access*, vol. 12, no. 1, pp. 80698–80730, 2024

[6] T. C. Tran and T. K. Dang, “Machine learning for prediction of imbalanced data: Credit fraud detection,” in *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, Seoul, Korea (South), vol. 2021, no. 1, pp. 1–7, 2021.

[7] C. Li and S. Liu, “A comparative study of the class imbalance problem in Twitter spam detection,” *Concurr. Comput.*, vol. 30, no. 5, pp. 42-61, Mar. 2018.

[8] F. Nhita, Adiwijaya, and I. Kurniawan, “Improvement of Imbalanced Data Handling: A Hybrid Sampling Approach by using Adaptive Synthetic Sampling and Tomek links,” *2023 Eighth International Conference on Informatics and Computing (ICIC), Manado, Indonesia*, vol. 2023, no. Dec., pp. 1–5, Dec. 2023.

[9] N. G. Ramadhan, W. Maharani, A. A. Gozali, and A. Adiwijaya, “Modified SMOTE and Ensemble Learning Based on Expert Judgment for Chronic Diseases Prediction,” *International Journal of Innovative Computing, Information and Control (IJICIC)*, vol. 21, no. 4, pp. 1-8, 2025.

[10] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, “A review on imbalanced data handling using undersampling and oversampling technique,” *Int. J. Recent Trends Eng. Res*, vol. 3, no. 4, pp. 444–449, 2017

[11] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling imbalanced datasets: A review,” *GESTS international transactions on computer science and engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[12] N. G. Ramadhan, “Comparative analysis of ADASYN-SVM and SMOTE-SVM methods on the detection of type 2 diabetes mellitus,” *Sci. J. Inform.*, vol. 8, no. 2, pp. 276–282, Nov. 2021.

[13] N. G. Ramadhan, Adiwijaya, W. Maharani, and A. Akbar Gozali, “Prediction of diabetes mellitus in the upcoming year using SMOTE and random forest,” in *2023 International Conference on Data Science and Its Applications (ICoDSA)*, vol. 2023, no. 1, pp. 316–321, 2023.

[14] I. M. Alkhawaldeh, I. Albalkhi, and A. J. Naswhan, “Challenges and limitations of synthetic minority oversampling techniques in machine learning,” *World J. Methodol.*, vol. 13, no. 5, pp. 373–378, Dec. 2023.

[15] T. Fahrudin, J. L. Buliali, and C. Fatichah, “Enhancing the performance of smote algorithm by using attribute weighting scheme and new selective sampling method for imbalanced data set,” *Int. J. Innov. Comput. Inf. Control*, vol. 15, no. 2, pp. 423–444, 2019.

[16] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning,” in *Lecture Notes in Computer Science*, Berlin, Heidelberg: Springer Berlin Heidelberg, vol. 2005, no. 1, pp. 878–887, 2005.

[17] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, “Safe-level-SMOTE: Safe-level-synthetic minority over-sampling TEchnique for handling the class imbalanced problem,” in *Advances in Knowledge Discovery and Data Mining*, Berlin, Heidelberg: Springer Berlin Heidelberg, vol. 2009, no. 1, pp. 475–482, 2009.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *jair*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[19] A. M. Halim, M. Dwifebri, and F. Nhita, “Handling imbalanced data sets using SMOTE and ADASYN to improve classification performance of ecoli data sets,” *Build Informatics Technol Sci (BITS*, vol. 5, no. 1, p. 246− 253-246− 253, Jun. 2023.

[20] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, “FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification,” *Pattern Recognit.*, vol. 124, no. 1, pp. 10-25, Apr. 2022.

[21] M. Azim Mim, N. Majadi, and P. Mazumder, “A soft voting ensemble learning approach for credit card fraud detection,” *Heliyon*, vol. 10, no. 3, pp. 25-46, Feb. 2024.

[22] Y. Miao and Y. Xu, “Random forest-based analysis of variability in feature impacts,” in *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)*, Shenyang, China, vol. 2024, no. 1, pp. 1130–1135, 2024.

[23] V. Hamer and P. Dupont, “An importance weighted feature selection stability measure,” *J. Mach. Learn. Res.*, vol. 22, no. 116, pp. 116:1-116:57, 2021.

[24] Zhang, Y., Wang, H., and Liu, J., “Conditional GAN for Oversampling in Highly Imbalanced Classification Tasks,” *Expert Systems with Applications*, vol. 213, no. 1, p. 11-40, 2023.

[25] Park, S., Kim, J., and Lee, D., “Attention-Guided Feature Weighting in Classification with Imbalanced Data,” *Neurocomputing*, vol. 550, no. 1, pp. 126–139, 2024.

[26] F. Nhita, Adiwijaya, and I. Kurniawan, “Performance and Statistical Evaluation of Three Sampling Approaches in Handling Binary Imbalanced Data Sets,” *2023 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia*, vol. 2023, no. 1, pp. 420-425 Aug. 2023.

[27] J. Derrac, S. Garcia, L. Sanchez, and F. Herrera, “Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework,” *J. Mult. Valued Logic Soft Comput*, vol. 17, no. 1, pp. 255–287, 2015.

[28] N. G. Ramadhan, Adiwijaya, and A. Romadhony, “Preprocessing handling to enhance detection of type 2 diabetes mellitus based on random forest,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, pp. 223–228, 2021.

[29] N. Komal Kumar, D. Vigneswari, M. Vamsi Krishna, and G. V. Phanindra Reddy, “An optimized random forest classifier for diabetes mellitus,” in *Advances in Intelligent Systems and Computing*, Singapore: Springer Singapore, vol. 2019, no. 1, pp. 765–773, 2019.

[30] Y. Nam and S. Han, “Random Forest variable importance-based selection algorithm in class imbalance problem,” *arXiv [stat.ML]*, vol. 2023, no. 1, pp. 1-12, 16-Dec-2023.

[31] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, “Threshold-based feature selection techniques for high-dimensional bioinformatics data,” *Netw. Model. Anal. Health Inform. Bioinform.*, vol. 1, no. 1–2, pp. 47–61, Jun. 2012.

[32] N. G. Ramadhan, A. Khoirunnisa, K. Kurnianingsih, and T. Hashimoto, “A hybrid ROS-SVM model for detecting target multiple drug types,” *JOIV Int. J. Inform. Vis.*, vol. 7, no. 3, pp. 794–800, Sep. 2023.

[33] N. Alamsyah, Saparudin, and A. Prima Kurniati, “Event detection optimization through stacking ensemble and BERT fine-tuning for dynamic pricing of airline tickets,” *IEEE Access*, vol. 12, no. 1, pp. 145254–145269, 2024.

[34] B. Rosner, R. J. Glynn, and M.-L. T. Lee, “The Wilcoxon signed rank test for paired comparisons of clustered data,” *Biometrics*, vol. 62, no. 1, pp. 185–192, Mar. 2006.