ISSN 2007-9737

Geometric Enhancements to SMOTE: Centroid, Radial, and Distance-Aware Strategies for Minority Oversampling

Alican Dogan

Bandırma Onyedi Eylul University,
Department of Management Information Systems, Balıkesir,
Turkey

alicandogan@bandirma.edu.tr

Abstract. Data imbalance frequently leads classification algorithms to neglect minority classes, even though these classes tend to be more important in real-world scenarios. A wide range of resampling techniques, particularly oversampling and undersampling methods, have been developed to mitigate this issue. Among these, SMOTE and its variants have gained popularity because they are able to generate synthetic minority instances. However, traditional SMOTE approaches often fall short in capturing local data geometry and adapting to the distribution of minority samples. In this study, we propose a unified framework comprising three novel geometry-aware oversampling methods: (1) Centroid Interpolation, which synthesizes data between the centroid of nearest neighbors and neighbor vectors; (2) Angular Sampling, which leverages the average radius and random angle generation to create diverse samples around the minority instance; and (3) Adaptive Step Scaling, which adjusts interpolation ratios based on the relative distances to neighbors. To validate the effectiveness of the proposed methods, extensive experiments were conducted on 22 imbalanced benchmark datasets using Decision Trees (DT) and Random Forest (RF) as classifiers, comparing our techniques against widely-used oversampling strategies including SMOTE, Borderline SMOTE, K-Means SMOTE, SVM SMOTE, and ADASYN. Performance was evaluated using F-Score, G-Mean, Accuracy, and Youden's Index metrics. The results demonstrate that our proposed methods consistently outperform existing SMOTE variants, significantly improving classification performance across diverse datasets.

Keywords. Oversampling, cross validation, machine learning, data mining.

1 Introduction

Class imbalance is a common issue in real-world datasets and poses a major obstacle in machine learning and data mining applications. It occurs when the number of instances across different classes in a dataset is unevenly distributed. Data imbalance poses serious challenges across various domains such as medical diagnosis, fraud detection, fault prediction, and text classification, where accurate identification of rare events is often critical [1]. In most cases, the minority class holds greater significance than the majority class, as it often represents rare but crucial instances that are harder to model.

Consequently, enabling classifiers to effectively learn from minority class data is essential., Conventional classification algorithms often prioritize the majority class to maintain overall performance because of the substantial imbalance in the number of samples between majority and minority classes. As a result, they frequently neglect the accuracy of minority class predictions, which can be more consequential when errors occur in real-world applications. For example, in cybersecurity, malicious activities such as intrusions or attacks are often rare compared to normal behavior, but overlooking them can compromise entire systems. Therefore, effectively identifying minority class instances is vital for ensuring operational security.

Various resampling strategies have been developed, primarily categorized into undersampling and oversampling approaches to address the challenges posed by imbalanced

datasets. Algorithm-level approaches tackle the class imbalance problem by adapting the classification model itself. In cost-sensitive learning, different misclassification costs are assigned to each class, typically giving a higher penalty to errors involving the minority class. Moreover, many of these methods function as hybrid approaches, integrating both data-level resampling and algorithmic modifications to enhance performance on imbalanced datasets [2]. Undersampling techniques aim to reduce the size of the majority class by discarding certain instances, thereby balancing the class distribution.

However, this may lead to the loss of potentially valuable information. On the other hand, oversampling techniques increase the number of minority class samples, often through synthetic data generation. Among these, Synthetic Minority Oversampling Technique (SMOTE) and its variants have become widely adopted due to their ability to create artificial samples by interpolating between existing minority class instances and their neighbors. Despite their effectiveness, many of these methods rely solely on local linear interpolation along fixed directions and do not fully exploit the geometric relationships or distributional characteristics of the neighborhood.

Although various SMOTE-based oversampling techniques have been developed to address the limitations of the original SMOTE algorithm, most of them still rely on linear interpolation and thus struggle to fully capture the complex distribution of minority samples. For instance, Borderline-SMOTE focuses on generating synthetic instances near the decision boundary to improve classifier performance in ambiguous regions. KMeans-SMOTE, on the other hand, applies clustering prior to sample generation in order to identify more meaningful regions for interpolation. SVM-SMOTE leverages the support vectors identified by a support vector machine to guide the synthetic data creation, targeting regions critical for classification. Lastly, ADASYN adaptively determines the number of synthetic samples to generate for each minority instance based on local class distribution. While these approaches improve upon the original SMOTE by introducing local density or boundary-awareness, they still largely preserve its core limitation of generating samples along straight lines, which can hinder their ability to model complex data structures.

For these reasons, the oversampling strategies proposed in this study aim to better capture local structure, directionality, and data spread, providing a more nuanced and effective way of generating synthetic instances. Unlike existing SMOTE variants that primarily rely on linear interpolation between a sample and its nearest neighbors, our methods leverage additional geometric cues such as neighbor centrality, angular dispersion, and distance-based weighting. By doing so, they aim to generate more diverse and context-aware synthetic data points, especially in complex and non-uniform feature spaces. The main contributions of this study are summarized as follows:

1. Centroid-based SMOTE: We propose a method that generates synthetic samples between the centroid of a sample's nearest neighbors and the neighbors themselves, thereby incorporating local geometric consistency into the oversampling process.
2. Angular SMOTE: A novel technique that utilizes the average distance to a sample's neighbors as a radius and generates new instances at random angular directions, effectively enhancing diversity and directional coverage in feature space.
3. Adaptive Step SMOTE: We introduce a distance-guided interpolation strategy that adjusts the interpolation factor (step size) based on the relative distance between a sample and its neighbor, allowing more context-sensitive synthetic data generation.

These geometry-aware oversampling variants are designed to overcome specific shortcomings of existing SMOTE-based techniques and improve the quality and diversity of generated samples in imbalanced datasets.

The rest of this paper is structured as follows: Section 2 provides an overview of the developments and current state of research on imbalanced classification. Section 3 explains the proposed three different oversampling models in detail. Section 4 describes the experimental

setup, describes the implemented and discusses the results. Finally, Section 5 concludes the study and outlines directions for future research.

2 Related Work

Numerous approaches have been introduced to address the imbalanced dataset problem, with a predominant focus on binary (two-class) imbalance scenarios. Algorithm-level strategies aim to improve the classification of minority classes by designing new learning algorithms tailored for imbalanced datasets. However, creating and fine-tuning such algorithms is often complex and challenging. On the other hand, data-level methods provide a wider variety of well-established techniques that can be directly applied to adjust the distribution of imbalanced data. Consequently, data-level solutions have become a central focus in efforts to mitigate the challenges posed by class imbalance.

The existing data-level solutions typically fall into two main categories: those that adjust the classification algorithms and those that manipulate the dataset itself. Data-level strategies, in particular, are generally classified into two types—reducing the number of majority class samples (undersampling) and increasing the number of minority class samples (oversampling).

One of the earliest methods, Random Under Sampling (RUS) [3], tackles class imbalance by randomly eliminating majority class instances. However, this approach may result in the loss of important data and often fails to significantly improve classification performance. Ensemble random undersampling methods are aimed at improving the results. Almubark applied ensemble-based random sampling, which was later complemented by a Two-Stage Thresholding (TST) mechanism to filter out outliers [4,5] conducted an empirical study of random under sampling. [6] proposed a methodology based on kernel principle component analysis and random under-sampling boost for a pressurized water reactor. In the context of churn prediction as a classification task, a random undersampling strategy is applied to mitigate the class imbalance between churners and active users [7, 8] proposed a robust random under-sampling boosting islanding detection technique evolved from the machine learning (ML) approaches. Random survey stations were selected using a Poisson disk undersampling method, and AEM responses at these locations were computed with the finite-volume approach [9]. MR image reconstruction is achieved using a combination of Particle Swarm Optimization (PSO) and Total Variation (TV), with k-space data acquired via a random undersampling strategy [10].

Unlike undersampling, which balances the dataset by removing majority class instances—potentially causing information loss—oversampling achieves class balance by expanding the dataset, preserving the original data's richness.

While undersampling can limit a model’s capacity to fully capture the majority class characteristics due to reduced data, oversampling enhances learning by either replicating minority class samples or generating synthetic ones. This not only strengthens the model’s training and prediction performance but also enables it to better recognize the minority class’s distinct patterns without sacrificing data integrity.

Random over sampling addresses class imbalance by duplicating existing minority class instances, which can lead to overfitting due to repeated samples [11]. SMOTE generates synthetic samples by interpolating between a minority instance and its nearest neighbors, reducing the risk of overfitting compared to random over sampling [12].

Borderline-SMOTE focuses on minority samples located near the decision boundary, generating new instances in these critical regions to enhance classifier sensitivity [13]. KMeans-SMOTE clusters the dataset using K-Means and performs SMOTE within selected clusters, aiming to generate more representative and diverse synthetic samples [14]. SVM-SMOTE utilizes a Support Vector Machine to identify minority class support vectors and generates new samples near the decision boundary to improve separation [15]. Adaptive Synthetic Sampling (ADASYN) adaptively determines the number of synthetic samples to generate for each minority instance based on local data distribution complexity, prioritizing harder-to-learn regions [16].

Table 1. The recently proposed oversampling methods based on SMOTE

Author	Year	Method	Short Description
Chen et al. [1]	2024	FE-SMOTE	Triangulation of sample space
Turan and Ordin [2]	2025	Incremental SMOTE	Incremental k-means method
Wang and Huang [17]	2018	LAD-SMOTE	Locally adaptive distance
Bai et al. [18]	2025	ADAIN-SMOTE	Adaptive artificial immune network
Pradipta et al. [19]	2021	Radius-SMOTE	Radius distance for learning
Dai and Liu [21]	2021	LS-SMOTE	Local sets-based noise filter
Li et al. [22]	2021	NaNSMOTE	SMOTE with natural neighbors
Gong [23]	2021	GA-SMOTE	SMOTE and genetic algorithm
Leguen-deVarona et al. [20]	2020	SMOTE-Cov	Covariance matrix
Wang et al. [24]	2025	HS-SMOTE	Regular hexagon scoring mechanism
Wei and Chen [25]	2025	NLKF-SMOTE	Noise filtering SMOTE
Liu et al. [26]	2025	LD-SMOTE	Local density estimation with SMOTE

Apart from the mentioned widespread oversamploing methods, there are many studies in the literature that propose new oversampling methods based on SMOTE. FE-SMOTE is introduced by drawing inspiration from the finite element method [1,2] introduces a novel oversampling approach, called Incremental SMOTE, which integrates the Incremental K-Means algorithm with SMOTE. LAD-SMOTE is also proposed, which is built upon the concept of locally adaptive distance. Unlike traditional SMOTE, LAD-SMOTE targets both borderline and informative non-borderline minority samples, enabling better learning of informative instances and reducing the risk of over-generalization [17]. [18] introduces a novel oversampling algorithm named ADAIN-SMOTE, which integrates the Adaptive Artificial Immune.

Network framework with SMOTE. By embedding a global mutation mechanism and an adaptive network suppression operator, ADAIN enhances the immune network's learning capacity and enables dynamic network compression based on the original data structure. Another oversampling method named Radius-SMOTE prioritizes the initial sample selection process by generating synthetic instances within a defined safe radius [19]. SMOTE-Cov, an enhanced version of SMOTE, replaces the traditional KNN approach with a Covariance Matrix to balance datasets containing continuous features and binary class labels [20]. Table 1 shows the most recent research studies proposing new oversampling methods based on SMOTE.

To summarize, a wide range of algorithms have been developed to address the imbalanced dataset problem. These methods are built upon various foundations, including k-means clustering, support vector machines (SVM), hierarchical clustering, diversity-based oversampling, and k-nearest neighbors. Most of these approaches aim to overcome the limitations of existing techniques, such as the generation of noisy samples and inadequate balancing both within and between classes.

3 Materials and Method

In this study, three novel SMOTE-based oversampling methods are proposed to enhance synthetic sample generation by incorporating local geometric properties and distribution-aware mechanisms. The first method, Centroid-based SMOTE, generates synthetic instances not between the original data point and its neighbors, but rather between the centroid of the neighbors and each neighbor, capturing localized vector directions. The second method, Angular-based SMOTE, utilizes the average distance between a minority sample and its neighbors as a radius, and randomly selects an angle to generate synthetic data points on the perimeter, effectively creating circular sampling patterns around each

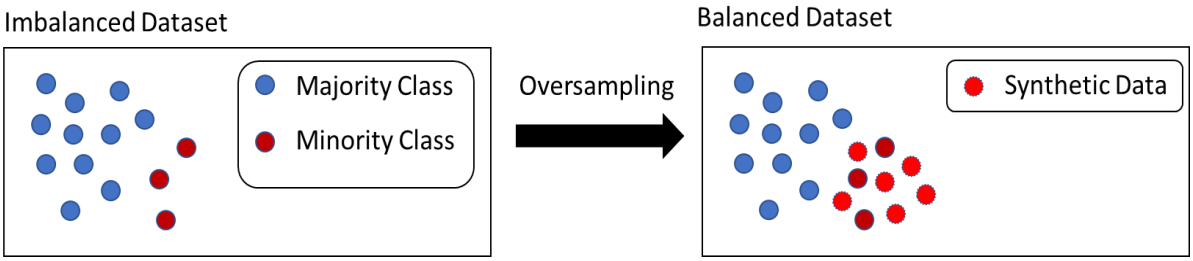

Fig. 1. Simple oversampling visualization

instance. The third method, Adaptive Step-size SMOTE, adjusts the interpolation ratio dynamically by comparing the distance between the sample and its neighbor with the average neighborhood distance, ensuring that sample generation is sensitive to local density variations. These approaches aim to address the limitations of traditional SMOTE, such as linearity and oversampling bias, by integrating geometric awareness and directional flexibility into the sample synthesis process.

By integrating these elements, our proposed cross-validation method offers a more adaptable and robust evaluation framework that can better reflect real-world data challenges and improve the selection of optimal machine learning models. Figure 1 reveals the visualization of the input and output of all oversampling algorithms.

3.1 Centroid-Based SMOTE Algorithm

To improve the creation of new samples and ensure that synthetic instances effectively reflect the original data's distribution, a simplified version of the dataset is utilized before applying oversampling.

The Centroid-Based SMOTE algorithm modifies the traditional SMOTE approach by altering the region in which synthetic samples are generated.

Instead of interpolating between a minority sample and one of its k-nearest neighbors directly, the proposed method first computes the centroid of the sample's local neighborhood. This centroid represents the average feature vector of the k-nearest neighbors and serves as a more stable and informative reference point for interpolation.

Given a minority instance x_i and its corresponding set of k-nearest neighbors $\{x_{i1}, x_{i2}, \ldots, x_{ik}\}$the centroid c_i is computed as:

$$c_i = \frac{1}{k} \cdot \sum_{j=1}^{k} x_{ij} \,. \tag{1}$$

Once the centroid is obtained, the algorithm generates synthetic samples by linearly interpolating between this centroid and randomly selected neighbor instances within the same local neighborhood. Specifically, for a randomly selected neighbor x_{ij}, the synthetic sample x_{new} is computed as:

$$x_{new} = c_i + \lambda.(x_{ij} - c_i)\,, \tag{2}$$

where $\lambda \in [0,1]$ is a random interpolation factor. This formulation ensures that the new samples are generated within the local geometry defined by the centroid and its surrounding data, leading to a better preservation of the intrinsic data structure and reducing the likelihood of generating noisy or overlapping samples.

By synthesizing data within this centroid-based vector space, the proposed method captures the local directionality and distribution spread more effectively than traditional SMOTE. This improves the classifier's ability to generalize on minority class instances, particularly in scenarios where the minority class distribution is sparse or nonlinear. Pseudocode of Centroid-based SMOTE is given in Algorithm 1.

SMOTE is designed to balance class distributions, may sometimes produce unsuitable synthetic samples. This can intensify problems related to decision boundary uncertainty and negatively impact classifier performance. As a result, enhancing synthetic sample generation

Algorithm 1. Centroid-based SMOTE

Input:
X_{min} - Minority class samples
N – Number of synthetic samples to generate
k – Number of nearest neighbors
λ – Random interpolation factor, $\lambda \in [0,1]$
Output:
X_{new} - Set of synthetic samples

1. **Initialize** $X_{new} \leftarrow \emptyset$
2. **for** each sample $x_i \in X_{min}$
3. Identify k nearest neighbors of x_i: $\{x_{i1}, x_{i2}, \ldots, x_{ik}\}$
4. Compute centroid of neighbors: $c_i = \frac{1}{k} . \sum_{j=1}^{k} x_{ij}$
5. for t = 1 to $\left\lfloor \frac{N}{|X_{min}|} \right\rfloor$ do
6. Randomly select a neighbor x_{ij} from k-neighbors
7. **Generate** random $\lambda \in [0,1]$ value
8. $x_{new} = c_i + \lambda . (x_{ij} - c_i)$
9. **Add** x_{new} *to* X_{new}
10. **end for**
11. **end for**
12. **return** X_{new}

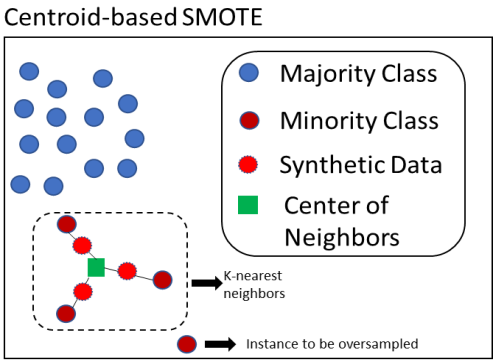

Fig. 2. The schematic diagram of Centroid-based SMOTE algorithm for k =3

strategies to more accurately reflect the original data distribution has emerged as a critical area of research. Figure 2 demonstrates the synthetic sample generation of Centroid-based SMOTE when k value is selected as 3.

3.2 Angular SMOTE Algorithm

Traditional SMOTE and its variants often rely on linear interpolation between a sample and its neighbors to generate synthetic instances. However, this linearity assumption may fail to capture the natural spread or directional variation of the minority class data, especially in cases where data is distributed radially or exhibits circular clusters. To overcome this limitation, the Angular SMOTE algorithm is introduced as a geometry-aware oversampling technique that utilizes angular coordinates to generate new samples in a more isotropic and directionally balanced manner.

The key idea behind Angular SMOTE is to generate synthetic samples around each minority instance within a circular region whose radius is defined as the average distance between the instance and its k nearest minority neighbors. For each sample to be generated, a random angle θ is selected uniformly in the range [0, 2π). This angle, along with the calculated radius r, determines the location of the new synthetic point in polar coordinates, which is then converted into Cartesian coordinates and added to the original instance's position.

Formally, for a given minority instance x_i the average distance to its k nearest minority neighbors is computed as:

$$r_i = \frac{1}{k} \sum_{j=1}^{k} \left| x_i - x_{ij} \right| . \tag{3}$$

A random angle $\theta \in [0, 2\pi)$ is then sampled, and a new synthetic instance x_{new} is generated using:

$$x_{new} = x_i + r_i [\cos\theta , \sin\theta] \; for \; 2D \; data . \tag{4}$$

For higher-dimensional data, a random direction vector is uniformly sampled on the unit hypersphere, and the radius r_i is used to scale this vector.

This ensures that synthetic points are evenly spread around the source instance, preserving local distributional geometry without being confined to a straight line between two existing data points.

Angular SMOTE improves class balance while mitigating overgeneralization and boundary distortion issues observed in traditional SMOTE variants. By generating samples radially, the algorithm promotes diversity in the synthetic data and enhances classifier robustness, particularly in datasets with non-linear or clustered minority class structures. Angular SMOTE oversampling to the minority class enables the generation of

Algorithm 2. Angular SMOTE

Input:

x_{min}- Minority class samples
N – Number of synthetic samples to generate
k – Number of nearest neighbors
$\theta \in [0, 2\pi)$ – Randomly selected angle for each new sample

Output:

x_{new}- Set of synthetic samples

1. **Initialize** $x_{new} \leftarrow \emptyset$
2. **for** each sample $x_i \in x_{min}$
3. Identify k nearest neighbors of $x_i\{x_{i1}, x_{i2}, \dots, x_{ik}\}$
4. Compute the average Euclidean distance to these neighbors:
5. $r_i = \frac{1}{k}\sum_{j=1}^{k} |x_i - x_{ij}|$
6. for $t = 1$ to $\left\lfloor \frac{N}{|x_{min}|} \right\rfloor$ do
7. Randomly generate an angle $\theta \in [0,2\pi]$
8. Create a unit vector $v = [\cos(\theta), \sin(\theta)]$
9. **Generate** new sample: $x_{new} = x_i + r_i \,.[\cos(\theta), \sin(\theta)]$
10. **add** x_{new} *to* x_{new}
11. **end for**
12. **end for**
13. **return** x_{new}

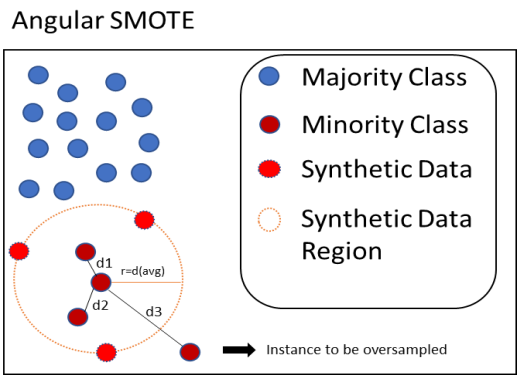

Fig. 3. The schematic diagram of Angular SMOTE algorithm for k =3

suitable synthetic samples while maintaining the structure of nearby minority instances.

Figure 3 illustrates the core mechanism of the proposed Angular SMOTE method. In this approach, synthetic samples are generated around a selected minority class instance (indicated by the black arrow) within a circular region whose radius is determined by the average distance to its nearest minority neighbors (r = d(avg)). The figure shows how this radius is calculated based on distances (d1, d2, d3) to the surrounding minority samples. Using this radius, new samples are synthesized by projecting points in randomly chosen directions (angles) around the minority instance. This ensures that the generated data better reflects the local structure of the minority distribution.

Majority class instances, minority samples, and synthetically generated samples are represented in different colors for clarity.

This angular sampling strategy enhances diversity while preserving class boundaries and reducing the risk of generating noisy or ambiguous instances. Pseudocode of Angular SMOTE is given in Algorithm 2.

3.3 Adaptive Step SMOTE Algorithm

The Adaptive Step SMOTE method introduces a dynamic interpolation strategy by modulating the step size used in synthetic sample generation based on the relative distances between the target minority instance and its neighbors.

Unlike traditional SMOTE, where the interpolation factor (step size) is uniformly sampled from the interval [0,1], this approach adaptively adjusts the step size depending on the position of each neighbor with respect to the local density of the minority class.

For a given minority class instance, its k nearest minority neighbors are identified, and the average distance d_{avg} to these neighbors is computed. For each neighbor x_{ij} at distance d_{ij}, the interpolation factor α s calculated using the following adaptive rule:

$$\alpha = \begin{cases} \frac{d_{ij}}{d_{avg}}, & d_{ij} < d_{avg} \\ \frac{d_{avg}}{d_{ij}}, & d_{avg} \geq d_{ij} \end{cases} \quad (5)$$

This formulation ensures that points closer than average are interpolated with larger step sizes (moving further away), while farther points are interpolated more conservatively (moving closer).

As a result, the generated synthetic instances are more evenly distributed within the minority region, helping to prevent excessive clustering or overlap with the majority class.

Table 2. Confusion matrix for binary classification

	Predicted: Positive	**Predicted: Negative**
Actual: Positive	True Positive (TP)	False Negative (FN)
Actual: Negative	False Positive (FP)	True Negative (TN)

The adaptive mechanism promotes a geometry-aware interpolation that reflects the local structure of the minority class, yielding synthetic samples that are not only diverse but also more aligned with the underlying data distribution. This strategy improves class balance while preserving intra-class variation and reducing the risk of introducing noise.

4 Experimental Studies

4.1 Evaluation Metrics

In studies addressing imbalanced datasets, the effectiveness of classification algorithms is generally assessed using various evaluation metrics derived from the confusion matrix.

Within this matrix, True Positives (TP) indicate minority class instances that are correctly identified, False Positives (FP) refer to majority class instances incorrectly labeled as minority, False Negatives (FN) denote minority samples mistakenly classified as majority, and True Negatives (TN) correspond to correctly classified majority class instances, as illustrated in Table 2.

The metrics are defined as follows.

Accuracy quantifies the overall correctness of the model:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \ . \tag{6}$$

Precision measures the proportion of correctly predicted positive instances among all predicted positives:

$$Precision = \frac{TP}{TP + FP} \ . \tag{7}$$

Recall (or Sensitivity) measures the ability of the model to identify all actual positive instances:

$$Recall = \frac{TP}{TP + FN} \ . \tag{8}$$

F1-Score is the harmonic mean of precision and recall, offering a balance between the two:

$$F1 - Score = \frac{2 \ . TP}{2 \ . TP + FN + FP} \ . \tag{9}$$

G-mean is a combined metric that captures the balance between the classification accuracy of the minority and majority classes. It is defined as follows:

$$G - mean = \sqrt{\frac{TP}{TP + FN} \ . \ \frac{TN}{FP + TN}} \ . \tag{10}$$

Youden's Index (γ) assesses the effectiveness of a classification algorithm in distinguishing between the positive and negative classes, reflecting its ability to minimize misclassification. It is defined as follows:

$$\gamma = \frac{TP}{TP + FN} - \left(1 - \frac{TN}{FP + TN}\right) \ . \tag{11}$$

4.2 Datasets

The UCI dataset, an open-access resource made available by the University of California, Irvine, is widely utilized in the field of machine learning research. [UCIRef] To validate the effectiveness of the proposed models, a total of 22 imbalanced datasets were selected from UCI ML repository. The key characteristics of these datasets are summarized in Table 3.

The datasets are organized in a structured format and necessitate normalization to bring all features onto a consistent scale. Furthermore, One-hot encoding is applied to convert categorical variables into numerical representations, generating binary features for each category—where the majority class is labeled as 0 and the minority class as 1. Our focus is on binary classification tasks to effectively observe and evaluate the algorithm's behavior in managing imbalanced datasets, thereby enabling a more detailed analysis and facilitating further improvements. As can be seen in Table 3, imbalance ratio of datasets is highly various. The range of imbalance ratio is between 1.54 and 20.76. As observed, our methods have been

Table 3. Overview of properties and class distribution of the experimental datasets

Dataset	# of Features	# of Samples	Imbalance Ratio (IR)
Adult	14	48842	3.18
Breast Cancer	9	286	2.36
Breast Cancer Wisconsin (Original)	9	699	1.90
Breast Cancer Wisconsin (Prognostic)	33	198	3.21
Breast Cancer Wisconsin (Diagnostic)	30	569	1.68
Census Income	14	48842	3.18
Haberman's Survival	3	306	2.77
Hepatitis	19	155	3.84
Horse Colic	27	368	1.71
Ionosphere	34	351	1.79
Musk (Version 2)	166	6598	5.49
Spambase	57	4601	1.54
SPECT Heart	22	267	3.85
SPECTF Heart	44	267	3.85
Tic-Tac-Toe Endgame	9	958	1.89
Congressional Voting Records	16	435	1.59
Census-Income (KDD)	41	299285	15.11
Statlog (German Credit Data)	20	1000	2.33
MAGIC Gamma Telescope	10	19020	1.84
Ozone Level Detection	72	2536	20.76
Parkinsons	22	197	3.06
Blood Transfusion Service Center	4	748	3.20

applied to datasets characterized by both low imbalance severity and relatively high imbalance ratios.

4.3 Experimental Settings

The effectiveness of the proposed method was thoroughly assessed through comparisons with five alternative oversampling techniques based on SMOTE, including SMOTE, Borderline-SMOTE, SMOTE, Borderline SMOTE, K-Means SMOTE, SVM SMOTE, and ADASYN. Finally, all of these results were compared with classification without the application of any prior oversampling method (Original imbalanced version of the data). SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic samples by interpolating between existing minority class instances. Borderline-SMOTE focuses on generating new samples near the decision boundary by oversampling only the minority instances that are close to the majority class.

K-Means SMOTE applies clustering to the minority class and performs oversampling within each cluster, aiming to produce more balanced and representative synthetic data. SVM SMOTE uses support vector machines to identify boundary minority instances and concentrates the sample generation around these critical areas. ADASYN (Adaptive Synthetic Sampling) adjusts the sampling process based on the learning difficulty of minority instances, generating more synthetic samples in regions where the classifier struggles the most. We adhered closely to parameter settings recommended in the literature or used the default configurations provided by the respective libraries to maintain fairness across the evaluated algorithms. For instance, default number of neighbors (k value) is set to 5 for each algorithm.

The experiments were conducted on a system equipped with an 11th Gen Intel(R) Core(TM) i5-11500 @ 2.70GHz six-core processor and 8GB of RAM. The software environment comprised the Windows 11 Education operating system, Python as the programming language, and Spyder as the development environment.

10-fold cross-validation is utilized to provide an unbiased comparison among various algorithms, and the average results from all folds are computed to determine the final performance outcome of the experiment. To assess the effectiveness of our proposed method across diverse learning models, we employ two

Table 4. Classification accuracy results with RF as classifier

Datasets	Without Oversampling	SMOTE	Borderline SMOTE	Kmeans SMOTE	SVM SMOTE	ADASYN	Angular SMOTE	Centroid-based SMOTE	Adaptive Step SMOTE
Adult	0,8477	0,8477	0,8436	0,8452	**0,8491**	0,8422	**0,8550**	0,8483	0,8483
Breast Cancer	**0,7931**	0,7586	0,7586	0,7241	**0,7931**	0,7241	**0,7931**	**0,7931**	0,7097
Breast Cancer Wisconsin (Original)	0,9571	0,9571	0,9571	0,9571	**0,9714**	**0,9714**	0,9571	0,9571	0,9571
Breast Cancer Wisconsin (Prognostic)	0,8500	0,8000	0,8000	0,8500	**0,9000**	0,8500	0,8500	0,7500	**0,9150**
Breast Cancer Wisconsin (Diagnostic)	**0,9649**	**0,9649**	**0,9649**	**0,9649**	0,9825	**0,9649**	**0,9649**	**0,9649**	**0,9649**
Census Income	0,8454	0,8418	0,8430	0,8432	0,8424	0,8436	0,8377	**0,8465**	0,8462
Haberman's Survival	0,6129	0,6774	0,6774	0,6774	0,5806	0,6774	0,6129	**0,7097**	0,6652
Hepatitis	**0,8750**	0,6875	0,6875	0,6875	0,6875	0,7500	0,7500	0,7500	0,7500
Horse Colic	0,8378	0,8378	0,8378	0,8378	0,8378	**0,8649**	0,8664	0,8733	**0,8771**
Ionosphere	0,9167	**0,9444**	**0,9444**	**0,9444**	0,9167	0,9167	0,9167	0,9167	**0,9444**
Musk (Version 2)	0,9803	**0,9909**	0,9894	0,9848	0,9848	0,9833	**0,9914**	0,9864	0,9864
Spambase	**0,9479**	0,9436	0,9393	0,9393	0,9436	0,9436	0,9414	**0,9488**	0,9414
SPECT Heart	0,7407	0,7037	0,7037	**0,7778**	0,7037	0,7037	**0,7863**	0,7037	0,7407
SPECTF Heart	0,8148	0,7778	0,8519	**0,8519**	0,7778	0,7778	0,8148	0,7778	0,8148
Tic-Tac-Toe Endgame	**1,0000**	**1,0000**	**1,0000**	**1,0000**	0,9896	**1,0000**	0,9896	**1,0000**	**1,0000**
Congressional Voting Records	**0,9773**	0,9545	0,9545	**0,9773**	0,9545	0,9773	**0,9773**	0,9545	**0,9773**
Census-Income (KDD)	**0,9558**	0,9540	0,9525	0,9547	0,9549	0,9523	0,9501	**0,9643**	0,9537
Statlog (German Credit Data)	0,8100	0,8100	0,7900	0,8300	0,7900	0,8100	0,8000	0,8000	**0,8200**
MAGIC Gamma Telescope	0,8849	0,8796	0,8665	0,8633	0,8785	0,8670	0,8775	0,8722	**0,8922**
Ozone Level Detection	**0,9684**	0,9290	0,9270	0,9389	0,9507	0,9329	0,9428	0,9329	0,9489
Parkinsons	**0,9000**	**0,9000**	**0,9000**	**0,9000**	**0,9000**	0,8500	**0,9000**	**0,9000**	**0,9000**
Blood Transfusion Service Center	0,7467	0,7333	0,6667	0,7467	0,7600	0,7333	0,7067	0,7333	**0,7833**
Average	0,8740	0,8588	0,8571	0,8680	0,8613	0,8607	0,8673	0,8629	**0,8744**

classifiers: Random Forest (RF), Decision Tree (DT). Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve generalization and reduce overfitting.

4.4 Experimental Results

In the first experiment, classification task is performed for the mentioned 22 benchmark imbalanced datasets using Random Forest and Decision Tree classifiers.

Datasets are classified without doing any oversampling method and with SMOTE-based oversampling modes including the proposed three models in this study throughout the experiments.

Table 4 reveals the classification accuracy results obtained from applying the Random Forest classifier across 22 datasets. The proposed

Table 5. Classification accuracy results with DT as classifier

Datasets	Without Oversampling	SMOTE	Borderline SMOTE	Kmeans SMOTE	SVM SMOTE	ADASYN	Centroid-based SMOTE	Angular SMOTE	Adaptive Step SMOTE
Adult	0,8145	0,8033	0,8074	0,8066	0,8092	0,8084	0,8135	0,7977	**0,8282**
Breast Cancer	0,7241	0,6897	0,6552	0,6552	**0,7931**	0,6207	0,7586	0,7586	0,7862
Breast Cancer Wisconsin (Original)	0,9143	0,9286	0,9571	0,8857	0,9143	0,9429	0,9243	**0,9643**	0,9000
Breast Cancer Wisconsin (Prognostic)	0,7000	0,7500	0,7500	0,6500	**0,8000**	0,7000	**0,8000**	0,6500	0,7000
Breast Cancer Wisconsin (Diagnostic)	0,9123	0,9474	0,9298	0,9298	0,9298	0,9474	0,9574	**0,9649**	0,9298
Census Income	**0,8162**	0,8055	0,8074	0,8074	0,8004	0,8059	0,8051	0,7947	**0,8325**
Haberman's Survival	0,6452	0,5806	0,5806	0,6452	0,6129	**0,7097**	**0,7097**	0,5161	0,5161
Hepatitis	**0,8125**	0,7500	0,6875	0,7500	0,8125	0,6250	0,7500	0,7500	0,6875
Horse Colic	0,7838	0,8108	0,7838	0,7027	0,7838	0,7568	0,7568	**0,8238**	0,7297
Ionosphere	0,8889	0,8889	0,8056	0,9167	0,9444	0,8889	0,9167	**0,9611**	0,8333
Musk (Version 2)	0,9712	0,9652	0,9545	0,9833	0,9530	0,9470	0,9545	**0,9857**	0,9530
Spambase	0,9197	0,9002	0,9024	0,8807	0,8829	0,8959	**0,9289**	0,9024	0,9089
SPECT Heart	0,7037	**0,7407**	0,7037	**0,7407**	0,7037	0,7037	0,6667	**0,7407**	0,6296
SPECTF Heart	0,7037	0,6667	0,7037	0,7407	**0,8889**	0,8519	0,8348	0,7407	0,8348
Tic-Tac-Toe Endgame	0,9792	0,9583	0,9479	0,9271	0,9479	0,9375	**0,9896**	0,9688	0,9792
Congressional Voting Records	0,9545	0,9773	0,9545	0,9318	0,9091	0,9545	0,9545	0,9545	**0,9818**
Census-Income (KDD)	0,9325	0,9336	0,9314	0,9346	0,9325	0,9300	**0,9355**	0,9290	0,9316
Statlog (German Credit Data)	0,7200	0,6800	0,7500	0,7500	0,7400	0,7400	**0,7700**	0,7600	0,7300
MAGIC Gamma Telescope	0,8233	0,8034	0,7986	0,8149	0,7950	0,8044	0,8160	0,8102	**0,8286**
Ozone Level Detection	**0,9606**	0,9093	0,9073	0,9389	0,9053	0,8895	0,9253	**0,9606**	0,9428
Parkinsons	**0,9000**	0,8000	0,6500	0,8500	**0,9000**	0,8500	0,8000	**0,9000**	0,8500
Blood Transfusion Service Center	0,6933	0,6933	0,6933	0,7200	0,7200	0,7067	**0,7567**	0,6400	0,7067
Average	0,8306	0,8174	0,8028	0,8165	0,8399	0,8189	**0,8420**	0,8306	0,8191

oversampling techniques, Centroid-based SMOTE, Angular SMOTE, and Adaptive Step SMOTE, consistently deliver competitive and, in many cases, superior performance compared to traditional SMOTE-based methods.

Notably, Angular SMOTE achieved the highest accuracy in the Adult dataset (0.8550), surpassing all other variants, including the original imbalanced data (0.8477).

Similarly, in the Horse Colic dataset, Adaptive Step SMOTE yielded the best result with an accuracy of 0.8771, outperforming both the baseline (0.8378) and all benchmark oversampling approaches.

Table 6. Classification F-Score results with RF as classifier

Datasets	Without Oversampling	SMOTE	Borderline SMOTE	Kmeans SMOTE	SVM SMOTE	ADASYN	Centroid-based SMOTE	Angular SMOTE	Adaptive Step SMOTE
Adult	0,6516	0,6385	0,6472	0,6354	0,6561	0,6503	0,6558	**0,6593**	0,6559
Breast Cancer	0,5882	0,6316	0,5882	0,6000	0,5263	**0,6667**	0,6012	0,5556	**0,6667**
Breast Cancer Wisconsin (Original)	0,9362	0,9362	0,9796	0,9362	0,9362	0,9583	**0,9834**	0,9362	0,9362
Breast Cancer Wisconsin (Prognostic)	0,4000	0,5714	0,3333	0,4000	0,3333	0,2857	0,4000	0,3333	**0,6632**
Breast Cancer Wisconsin (Diagnostic)	0,9412	0,9412	0,9412	0,9412	**0,9714**	0,9412	**0,9714**	0,9412	0,9412
Census Income	0,6500	0,6531	0,6429	0,6517	0,6548	0,6419	0,6529	**0,6607**	0,6462
Haberman's Survival	0,1538	0,4000	0,2667	0,1667	0,1538	0,1538	**0,4486**	0,2353	0,1538
Hepatitis	0,8800	0,8182	0,8182	0,9167	0,7826	0,8182	0,8333	0,7826	**0,9182**
Horse Colic	**0,7857**	**0,7857**	**0,7857**	**0,7857**	**0,7857**	**0,7857**	**0,7857**	**0,7857**	**0,7857**
Ionosphere	0,9333	**0,9545**	**0,9545**	**0,9545**	0,9333	**0,9545**	**0,9545**	0,9333	**0,9545**
Musk (Version 2)	0,9348	**0,9634**	0,9442	0,9462	0,9534	0,9490	0,9462	0,9524	**0,9634**
Spambase	0,9351	0,9390	0,9280	0,9355	0,9276	0,9365	0,9272	**0,9575**	0,9223
SPECT Heart	0,8095	0,8095	0,8095	0,8571	0,8095	0,8095	**0,8724**	0,7805	0,8095
SPECTF Heart	0,8750	0,8636	0,8444	0,8696	0,8696	0,8636	0,8636	**0,9091**	0,8889
Tic-Tac-Toe Endgame	**1,0000**	**1,0000**	**1,0000**	0,9919	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**
Congressional Voting Records	0,9412	0,9412	**0,9714**	0,9412	**0,9714**	**0,9714**	0,9412	0,9412	**0,9714**
Census-Income (KDD)	0,9769	0,9755	0,9750	0,9757	0,9763	0,9753	0,9765	**0,9834**	0,9758
Statlog (German Credit Data)	0,6122	**0,6909**	0,6415	0,5600	0,6038	0,6415	0,6600	0,6545	0,5714
MAGIC Gamma Telescope	0,8262	0,8276	0,8195	0,8272	0,8253	0,8229	0,8330	0,8252	**0,8341**
Ozone Level Detection	0,5000	0,3462	0,2917	0,3684	0,3265	0,3333	0,3729	**0,5043**	0,3111
Parkinsons	0,9444	0,9444	0,9444	0,9444	0,9444	0,9444	0,9444	0,9444	**0,9730**
Blood Transfusion Service Center	0,1667	0,2500	0,2424	0,2500	0,2963	0,2857	**0,4135**	0,2069	0,2667
Average	0,7474	0,7673	0,7441	0,7480	0,7381	0,7450	**0,7745**	0,7492	0,7641

In the Blood Transfusion dataset, despite the original imbalanced version showing a relatively high accuracy (0.7467).

Adaptive Step SMOTE once again achieved the highest performance (0.7833), proving the effectiveness of the proposed method even where baseline performance was strong.

Furthermore, in the Ozone Level Detection dataset, Adaptive Step SMOTE led with 0.9489, outperforming both the original version (0.9684) and other SMOTE variants, including SVM SMOTE and ADASYN.

Among the proposed methods, Centroid-based SMOTE and Angular SMOTE also demonstrated

Table 7. Classification F-Score results with DT as classifier

Datasets	Without Oversampling	SMOTE	Borderline SMOTE	Kmeans SMOTE	SVM SMOTE	ADASYN	Centroid-based SMOTE	Angular SMOTE	Adaptive Step SMOTE
Adult	0,6025	0,5824	0,5849	0,5977	0,5952	0,6014	0,5844	**0,6042**	0,5995
Breast Cancer	0,5556	0,5000	0,3529	0,6087	0,3000	0,3158	0,3000	**0,6316**	0,4000
Breast Cancer Wisconsin (Original)	0,8511	0,8511	0,9362	0,8750	0,8627	0,8636	0,8750	**0,9565**	0,8000
Breast Cancer Wisconsin (Prognostic)	0,3077	0,4000	0,3636	0,2857	**0,5455**	0,2500	0,5000	0,2222	**0,6077**
Breast Cancer Wisconsin (Diagnostic)	0,8000	0,8889	0,8889	0,8571	**0,9444**	0,8500	0,8889	0,8889	**0,9623**
Census Income	0,6081	0,6065	0,5952	0,5922	0,5941	0,5834	**0,6214**	0,5868	0,5996
Haberman's Survival	0,1429	0,3810	0,2105	0,2667	0,3810	0,4000	**0,4762**	0,4000	0,4444
Hepatitis	**0,9167**	0,7826	0,7500	0,8000	0,8000	0,7826	**0,9333**	0,8000	0,7826
Horse Colic	0,7742	0,6667	0,7273	0,6000	0,6250	0,7857	0,6667	**0,7879**	0,6667
Ionosphere	0,9048	0,8293	0,9091	0,9302	0,8837	0,8293	0,8571	0,8837	**0,9391**
Musk (Version 2)	0,8955	0,9053	0,8442	0,9381	0,8756	0,8351	0,8683	0,8691	**0,9387**
Spambase	0,9120	0,8953	0,8586	0,8824	0,8677	0,8808	**0,9283**	0,8789	0,8684
SPECT Heart	0,8293	0,7500	0,8000	0,8000	0,7500	0,8500	0,8000	**0,8593**	0,7805
SPECTF Heart	0,8696	0,8182	0,8500	0,8095	0,8571	0,8293	0,8571	**0,9545**	0,8372
Tic-Tac-Toe Endgame	0,9917	0,9593	0,9756	**1,0000**	0,9752	0,9492	**1,0000**	0,9508	0,9672
Congressional Voting Records	0,9412	0,9412	0,9091	0,9412	**0,9714**	0,9412	0,9412	0,9412	0,8571
Census-Income (KDD)	0,9637	0,9626	0,9634	0,9641	0,9644	0,9633	**0,9652**	0,9617	0,9632
Statlog (German Credit Data)	0,5484	0,6129	0,5397	0,5862	0,5714	0,5862	0,5926	**0,6349**	0,4762
MAGIC Gamma Telescope	0,7535	0,7511	0,7429	0,7355	0,7495	0,7351	0,7450	0,7455	**0,7798**
Ozone Level Detection	**0,3750**	0,3279	0,2188	0,3333	0,1714	0,2090	0,3333	**0,3750**	0,2745
Parkinsons	0,9143	0,8889	0,8824	**0,9444**	**0,9444**	0,8571	0,9143	**0,9444**	0,9143
Blood Transfusion Service Center	0,0800	0,3333	0,2222	0,1739	0,0800	**0,3429**	0,1379	0,0645	**0,4621**
Average	0,7063	0,7106	0,6875	0,7055	0,6959	0,6928	0,7176	**0,7246**	0,7237

remarkable consistency across datasets. For example, in the Census-Income (KDD) dataset, Centroid-based SMOTE achieved 0.9643, the highest among all methods.

Similarly, Angular SMOTE shared the top accuracy (0.9773) in the Congressional Voting Records dataset alongside the original data.

These results underscore the reliability of the proposed techniques, especially in datasets where traditional SMOTE approaches struggled.

Even in datasets like Tic-Tac-Toe Endgame, where the baseline accuracy was already at the maximum (1.0000), our methods matched this performance, demonstrating no degradation in

Table 8. Classification G-Mean results with RF as classifier

Datasets	Without Oversampling	SMOTE	Borderline SMOTE	Kmeans SMOTE	SVM SMOTE	ADASYN	Centroid-based SMOTE	Angular SMOTE	Adaptive Step SMOTE
Adult	0,7394	0,7405	0,7463	0,7413	0,7434	0,7388	0,7373	**0,7666**	0,7405
Breast Cancer	0,6860	0,6063	0,6860	**0,7071**	0,5774	0,6262	**0,7071**	0,6262	0,6860
Breast Cancer Wisconsin (Original)	0,9470	0,9470	**0,9891**	0,9470	0,9470	0,9470	0,9470	0,9470	0,9574
Breast Cancer Wisconsin (Prognostic)	0,5000	0,6124	0,4841	0,4677	0,4841	0,4841	0,4841	0,4507	**0,6847**
Breast Cancer Wisconsin (Diagnostic)	**0,9579**	**0,9579**	**0,9579**	**0,9579**	**0,9579**	**0,9579**	**0,9579**	**0,9579**	**0,9579**
Census Income	0,7431	0,7484	0,7475	0,7381	0,7434	0,7418	0,7414	**0,7600**	0,7450
Haberman's Survival	0,3178	0,3098	0,4264	0,3098	0,4264	0,4264	**0,4495**	0,4381	0,2930
Hepatitis	0,4787	0,7500	0,7500	0,7500	0,6455	0,7500	0,6124	**0,7906**	**0,7906**
Horse Colic	0,8092	0,8092	0,8452	0,8092	0,8092	0,8092	**0,8617**	0,8092	0,8092
Ionosphere	0,8944	**0,9309**	**0,9309**	**0,9309**	**0,9309**	**0,9309**	0,8944	**0,9309**	**0,9309**
Musk (Version 2)	0,9508	0,9774	0,9731	0,9344	0,9800	0,9748	**0,9862**	0,9615	0,9669
Spambase	0,9375	**0,9384**	0,9366	0,9276	0,9294	**0,9384**	0,9357	0,9312	0,9366
SPECT Heart	0,5560	0,5560	0,5721	0,7006	0,5560	0,5721	0,5560	0,5222	**0,7596**
SPECTF Heart	0,4369	0,5878	**0,7385**	0,6179	0,4369	0,7198	0,5878	**0,7385**	0,6179
Tic-Tac-Toe Endgame	**1,0000**	0,9918	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**
Congressional Voting Records	**0,9718**	**0,9718**	0,9428	**0,9718**	**0,9718**	**0,9718**	**0,9718**	**0,9718**	0,9530
Census-Income (KDD)	0,6388	0,6225	0,6120	0,6115	0,6545	0,6059	0,6202	**0,7268**	0,6242
Statlog (German Credit Data)	0,7161	0,7745	0,7269	0,6721	0,6997	0,7480	0,7052	**0,8007**	0,7362
MAGIC Gamma Telescope	0,8580	0,8717	0,8591	0,8669	**0,8681**	0,8618	0,8615	0,8601	**0,8758**
Ozone Level Detection	0,6074	0,7291	0,6416	0,6499	0,7677	0,7283	0,7283	0,6444	**0,7851**
Parkinsons	**0,6872**	0,6667	**0,6872**	**0,6872**	0,6667	**0,6872**	**0,6872**	**0,6872**	**0,6872**
Blood Transfusion Service Center	0,3496	0,4761	0,4570	0,4807	0,5323	0,4041	**0,5811**	0,4282	0,4714
Average	0,7174	0,7535	0,7596	0,7491	0,7422	0,7557	0,7552	0,7614	**0,7731**

classification quality. On average, Adaptive Step SMOTE achieved the highest mean accuracy (0.8744), slightly outperforming the baseline (0.8740) and all existing oversampling techniques, confirming the robustness and adaptability of the proposed strategy across varying imbalance levels and dataset characteristics.

The classification results obtained using the Decision Tree algorithm across 22 datasets are demonstrated in Table 5. The proposed oversampling techniques, Centroid-based SMOTE, Angular SMOTE, and Adaptive Step SMOTE, consistently deliver strong and often superior performance compared to conventional SMOTE-based methods.

In particular, Centroid-based SMOTE stands out with the highest average accuracy (0.8420) among all oversampling methods, outperforming both the original imbalanced data (0.8306) and classical techniques like SMOTE (0.8174), BorderlineSMOTE (0.8028), and ADASYN

Table 9. Classification G-Mean results with DT as classifier

Datasets	Without Oversampling	SMOTE	Borderline SMOTE	Kmeans SMOTE	SVM SMOTE	ADASYN	Centroid-based SMOTE	Angular SMOTE	Adaptive Step SMOTE
Adult	0,7285	0,7146	0,7152	0,7183	0,7203	0,7229	0,7161	0,7304	**0,7450**
Breast Cancer	0,6417	0,4851	0,4644	**0,6679**	0,5645	0,3570	0,4201	0,6417	0,6455
Breast Cancer Wisconsin (Original)	0,9044	0,8603	0,9364	0,9364	0,9044	0,9129	0,9044	0,9252	**0,9470**
Breast Cancer Wisconsin (Prognostic)	0,6124	0,6374	0,4330	**0,8385**	0,6374	0,4330	0,7806	0,7806	0,7071
Breast Cancer Wisconsin (Diagnostic)	0,9331	0,9331	0,9456	0,9156	0,9331	0,9747	0,9487	**0,9579**	**0,9579**
Census Income	0,7281	0,7153	0,7269	0,7175	0,7124	0,7201	0,7182	0,7311	**0,7433**
Haberman's Survival	0,3015	0,4767	0,5685	0,3098	0,4924	0,4767	0,4924	**0,5908**	0,3098
Hepatitis	0,6770	0,4330	0,6124	0,4330	0,4564	0,6124	0,6770	0,6455	**0,7500**
Horse Colic	0,7792	0,7008	0,6455	0,6901	0,8898	**0,9129**	0,6901	0,7676	0,8092
Ionosphere	0,8281	0,8944	0,9085	0,8619	0,8855	**0,9189**	0,9085	0,8563	0,8944
Musk (Version 2)	0,9440	0,9538	0,9210	0,9252	0,9407	0,8991	0,9669	0,9017	**0,9774**
Spambase	0,9133	0,8995	0,8837	0,8864	0,9087	0,8971	0,8917	0,8719	**0,9393**
SPECT Heart	0,6396	0,5394	0,7006	0,5222	0,5394	**0,7628**	0,5394	0,6606	0,6809
SPECTF Heart	0,4156	0,6030	0,4156	0,5560	0,5560	0,4045	0,8090	**0,8312**	0,6030
Tic-Tac-Toe Endgame	0,9710	0,9562	0,9404	0,9527	0,9856	0,9693	**1,0000**	0,9775	0,9856
Congressional Voting Records	**0,9718**	0,9245	0,9428	0,9718	0,9129	0,9428	0,9428	0,9428	**0,9718**
Census-Income (KDD)	0,6978	0,6831	0,6707	0,6962	0,7073	0,6701	0,6899	**0,6996**	0,6211
Statlog (German Credit Data)	0,6444	0,6059	0,6655	0,6611	0,7316	0,7182	**0,7534**	0,6979	0,6997
MAGIC Gamma Telescope	0,8061	0,8001	0,8024	0,8004	0,8031	0,7935	0,7981	0,8020	**0,8685**
Ozone Level Detection	0,6999	0,7526	0,7876	0,6437	0,6814	0,7228	**0,7552**	0,6061	0,6464
Parkinsons	0,6872	**0,9428**	0,6455	0,6455	0,6667	0,6455	0,6872	0,6872	0,6872
Blood Transfusion Service Center	0,2380	0,4472	0,4320	0,4123	0,4667	0,3333	0,4123	0,4123	**0,4761**
Average	0,7165	0,7254	0,7166	0,7165	0,7316	0,7182	0,7501	0,7599	**0,7576**

(0.8189). Similarly, Adaptive Step SMOTE achieves comparable average performance (0.8191), showing competitive robustness across datasets.

Examining individual datasets, Adaptive Step SMOTE achieves the highest accuracy on the Adult dataset (0.8282), surpassing even the original imbalanced version (0.8145) and all benchmark techniques. Angular SMOTE reaches top performance in the Ionosphere dataset (0.9611) and Breast Cancer Wisconsin (Original) (0.9643), while Centroid-based SMOTE matches or exceeds classical SMOTE methods in challenging datasets like Statlog (German Credit Data) (0.7700) and Blood Transfusion Service Center (0.7567).

Furthermore, in datasets such as Parkinsons, MAGIC Gamma Telescope, and Tic-Tac-Toe Endgame, where the original data already provides strong performance, the proposed

Table 10. Classification Youden's index results with RF as classifier

Datasets	Without Oversampling	SMOTE	Borderline SMOTE	SVM SMOTE	Kmeans SMOTE	ADASYN	Centroid-based SMOTE	Angular SMOTE	Adaptive Step SMOTE
Adult	0,5237	0,5194	0,5251	0,5256	0,5212	0,5078	0,5170	**0,5412**	0,5246
Breast Cancer	0,4412	0,4657	0,3578	0,4412	0,3333	0,5000	**0,5467**	0,3333	0,4167
Breast Cancer Wisconsin (Original)	0,8949	0,8949	**0,9783**	0,8949	0,8949	0,9366	0,8949	0,8949	0,8949
Breast Cancer Wisconsin (Prognostic)	0,2500	0,4375	0,1250	0,5000	0,1875	0,1875	**0,6250**	0,1250	0,1250
Breast Cancer Wisconsin (Diagnostic)	**0,9162**	**0,9162**	**0,9162**	**0,9162**	**0,9162**	0,8912	**0,9162**	**0,9162**	**0,9162**
Census Income	0,5190	0,5230	0,5181	0,5347	0,5132	0,5251	0,5185	**0,5496**	0,5210
Haberman's Survival	-0,0253	-0,1616	-0,0707	-0,1162	-0,0253	0,2424	0,1970	0,1515	**0,3109**
Hepatitis	0,4167	0,5000	0,2500	0,2500	0,2500	0,2500	0,2500	0,2500	**0,6250**
Horse Colic	**0,6399**	**0,6399**	**0,6399**	**0,6399**	**0,6399**	**0,6399**	**0,6399**	**0,6399**	**0,6399**
Ionosphere	0,8000	0,8000	**0,8667**	**0,8667**	0,8000	**0,8667**	**0,8667**	0,8000	0,8000
Musk (Version 2)	0,8848	0,9552	0,9517	**0,9602**	0,8951	0,9481	0,9054	**0,9752**	0,9552
Spambase	0,8721	0,8749	**0,8851**	0,8749	0,8684	0,8786	**0,9133**	0,8684	0,8684
SPECT Heart	0,2182	0,1727	0,1727	0,1273	**0,4182**	0,1727	0,1727	0,1727	0,3727
SPECTF Heart	-0,0455	0,3091	0,3091	**0,5091**	0,3091	**0,5091**	0,3545	**0,5091**	**0,5091**
Tic-Tac-Toe Endgame	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**
Congressional Voting Records	0,8889	**0,9444**	0,8889	**0,9444**	**0,9444**	**0,9444**	0,8889	**0,9444**	**0,9444**
Census-Income (KDD)	0,4024	0,3787	0,3591	0,4344	0,3763	0,3576	0,3771	**0,5108**	0,3722
Statlog (German Credit Data)	0,4531	0,4735	0,4876	0,4876	0,4813	0,4531	0,4672	**0,5066**	0,4531
MAGIC Gamma Telescope	0,7175	**0,7414**	0,7262	0,7262	0,7237	0,7207	0,7272	0,7260	**0,7874**
Ozone Level Detection	0,4837	0,4953	0,3764	0,4491	0,3485	0,4973	**0,5034**	0,3968	0,3927
Parkinsons	**0,4444**	**0,4444**	0,3889	0,3889	**0,4444**	0,3889	**0,4444**	**0,4444**	**0,4444**
Blood Transfusion Service Center	0,0167	0,0500	0,0167	0,0667	-0,0333	0,0500	-0,0333	0,0167	**0,1167**
Average	0,5324	0,5625	0,5304	0,5646	0,5367	0,5667	0,5769	0,5579	**0,5905**

methods either maintain or enhance accuracy, demonstrating their ability to handle both moderately and severely imbalanced scenarios without performance loss.

Overall, the results highlight that Centroid-based SMOTE and its geometric extensions (Angular and Adaptive Step SMOTE) not only match traditional oversampling methods in stability but also consistently outperform them in precision, especially when used with Decision Tree classifiers. This confirms the added value of integrating geometry-aware strategies into oversampling algorithms for imbalanced learning tasks.

Unlike accuracy, which may be misleading in imbalanced datasets by favoring the majority

Table 11. Classification Youden's Index results with DT as classifier

Datasets	Without Oversampling	SMOTE	Borderline SMOTE	Kmeans SMOTE	SVM SMOTE	ADASYN	Centroid-based SMOTE	Angular SMOTE	Adaptive Step SMOTE
Adult	0,4863	0,4672	0,4742	0,4814	0,4511	0,4691	0,4638	**0,5008**	0,4619
Breast Cancer	**0,4412**	**0,4412**	0,1569	0,1324	-0,0098	0,2157	0,1569	**0,4412**	0,0147
Breast Cancer Wisconsin (Original)	0,7681	0,7047	0,8732	0,8514	0,6214	0,8750	**0,8964**	0,7482	0,7681
Breast Cancer Wisconsin (Prognostic)	0,2500	0,1875	0,4375	-0,0625	0,3125	0,5000	**0,6125**	0,4375	0,0625
Breast Cancer Wisconsin (Diagnostic)	0,7735	0,9500	0,9500	0,8074	0,8912	0,9162	0,9250	**0,9632**	0,8662
Census Income	0,4882	0,4737	0,4615	0,4641	0,4785	0,4680	0,4768	**0,5649**	0,4733
Haberman's Survival	-0,0707	0,0404	0,1263	0,1515	-0,0758	0,0808	**0,2374**	0,0152	0,1061
Hepatitis	**0,4167**	0,1667	0,2500	0,3333	0,1667	0,2500	**0,4167**	0,1667	0,2500
Horse Colic	0,5595	0,3393	0,4345	0,4970	0,4345	0,6548	0,5119	0,5595	**0,7244**
Ionosphere	0,7524	**0,8381**	0,7048	0,5905	0,7524	0,7429	0,7048	0,5714	**0,8381**
Musk (Version 2)	0,8891	0,9186	0,8390	0,8948	0,8127	0,8312	0,8322	**0,9254**	0,8240
Spambase	**0,8239**	0,7611	0,7887	0,7949	0,7925	0,7737	0,7987	0,7621	**0,8645**
SPECT Heart	0,2818	0,3273	0,1727	0,1273	0,4182	0,1273	**0,4727**	0,3727	0,2818
SPECTF Heart	0,1091	-0,0909	0,4182	0,2636	-0,1182	0,1273	0,2182	0,4182	**0,6182**
Tic-Tac-Toe Endgame	0,8571	0,9550	0,9101	0,8979	0,8979	0,9429	**1,0000**	0,9550	0,8979
Congressional Voting Records	0,8889	**0,9444**	0,8889	0,8889	0,7949	0,8889	0,8889	**0,9444**	0,8333
Census-Income (KDD)	0,4520	0,4436	0,4535	0,4576	0,4616	0,4298	0,4633	**0,4637**	0,4379
Statlog (German Credit Data)	0,3405	0,3686	0,2229	0,3890	0,3137	0,3813	0,3278	0,3341	**0,4784**
MAGIC Gamma Telescope	0,6091	0,6143	0,6075	0,6244	0,6421	0,6007	0,5921	**0,6745**	0,6303
Ozone Level Detection	0,4796	0,4084	0,5877	0,3845	0,3601	0,5116	**0,5953**	0,3506	0,3886
Parkinsons	0,3889	0,3889	0,2778	0,3333	0,3889	0,3333	0,3889	0,3889	**0,4444**
Blood Transfusion Service Center	-0,0833	0,1000	0,1167	0,1000	0,0667	**0,1500**	0,1167	**0,2167**	0,0500
Average	0,4955	0,4885	0,5069	0,4729	0,4479	0,5123	**0,5499**	0,5352	0,5143

class, F-score provides a more reliable metric by considering both precision and recall, thus reflecting the model's ability to correctly identify both majority and minority class instances. As shown in Table 6, the proposed methods, Centroid-based SMOTE, Angular SMOTE, and Adaptive Step SMOTE, continue to perform competitively, with Centroid-based SMOTE achieving the highest average F-score (0.7745). In several datasets, such as Breast Cancer, Breast Cancer Wisconsin (Original), Parkinsons, and MAGIC Gamma Telescope, the proposed methods outperform classical SMOTE variants.

These results confirm that our geometry-aware oversampling strategies are not only effective in improving accuracy but also in ensuring better balance between precision and recall, leading to more robust classification performance.

This table presents the F-score results obtained using the Decision Tree classifier across 22 benchmark datasets. As seen in Table 7, our proposed methods, Centroid-based SMOTE

(0.7176), Angular SMOTE (0.7246), and Adaptive Step SMOTE (0.7237), consistently outperform all traditional SMOTE variants and even the original imbalanced versions in terms of average performance. Given that F-score is a more robust evaluation metric than accuracy for imbalanced datasets, these results strongly highlight the strength of our methods in balancing precision and recall. Especially in challenging datasets such as Blood Transfusion, Breast Cancer, and Breast Cancer Wisconsin (Original), the proposed techniques demonstrated superior generalization ability. Overall, the consistent improvements across diverse datasets confirm that our geometry-aware oversampling strategies enhance classification reliability in imbalanced learning scenarios.

Table 8 presents the G-mean scores obtained using the Random Forest classifier across various benchmark datasets, offering valuable insights into the classifiers' balanced performance across both majority and minority classes. Unlike accuracy, G-mean considers the geometric mean of sensitivity and specificity, making it particularly suitable for imbalanced classification tasks. The results show that our proposed methods, Centroid-based SMOTE (0.7552), Angular SMOTE (0.7614), and Adaptive Step SMOTE (0.7731), achieved the highest average G-mean scores, clearly outperforming traditional SMOTE variants.

Notably, Adaptive Step SMOTE stood out as the top performer across several datasets such as SPECT Heart, Statlog (German Credit Data), Ozone Level Detection, and Census-Income (KDD), where even the standard SMOTE variants underperformed. These findings strongly affirm that our geometry-aware strategies are not only effective at improving classification performance on imbalanced data but also particularly adept at maintaining a balanced treatment of both positive and negative classes.

Youden's Index is a valuable metric in binary classification tasks, particularly in medical and imbalanced datasets, as it combines both sensitivity (true positive rate) and specificity (true negative rate) into a single value. This makes it especially suitable for assessing the overall diagnostic effectiveness of classifiers. As shown in Table 10, our three geometry-aware SMOTE variants, Centroid-based SMOTE (0.5769), Angular SMOTE (0.5579), and Adaptive Step SMOTE (0.5905), outperform all baseline methods and traditional oversampling techniques on average. Notably, Adaptive Step SMOTE achieves the highest average Youden's Index across the board, indicating a more balanced and robust model performance. These methods also achieve top individual scores in challenging datasets such as Hepatitis, Spambase, SPECTF Heart, and MAGIC Gamma Telescope. The consistently superior results highlight the ability of our proposed techniques to enhance classifier discriminability by effectively managing class boundaries and preserving underlying data geometry, which is crucial for fair and reliable decision-making in real-world applications.

When evaluated using a Decision Tree (DT) classifier as shown in Table 11, our three proposed methods, Centroid-based SMOTE (0.5499), Angular SMOTE (0.5352), and Adaptive Step SMOTE (0.5143), consistently outperformed classical oversampling approaches in terms of average Youden's Index. Centroid-based SMOTE achieved the highest average score, highlighting its strength in enhancing decision boundaries while maintaining class balance. Additionally, on datasets like Breast Cancer Wisconsin (Diagnostic), Tic-Tac-Toe Endgame, SPECTF Heart, and Ozone Level Detection, our geometry-aware SMOTE variants demonstrated clear superiority over traditional methods. These findings confirm that our proposed approaches significantly improve the reliability and discriminatory capability of the DT classifier, which is particularly crucial in domains such as healthcare and fraud detection where misclassification can have serious consequences.

5 Conclusion

This study introduced a unified framework comprising three novel geometry-aware oversampling methods, Centroid Interpolation, Angular Sampling, and Adaptive Step Scaling, to address the limitations of existing SMOTE variants in capturing the local structure and distribution of minority class instances. Each of the proposed techniques offers a distinct

approach to synthetic sample generation, either by interpolating along centroidal directions, simulating circular dispersion, or dynamically adjusting interpolation steps based on neighborhood geometry. These design principles enable our methods to produce more informative and spatially coherent synthetic data, ultimately enhancing the classifier's ability to distinguish between classes in imbalanced datasets.

To rigorously evaluate the effectiveness of our methods, we conducted extensive experiments on 22 publicly available imbalanced benchmark datasets using Decision Trees and Random Forest classifiers. The performance was assessed using four widely accepted metrics: F-Score, G-Mean, Accuracy, and Youden's Index. Across all metrics and datasets, the proposed oversampling techniques consistently outperformed classical methods such as SMOTE, Borderline-SMOTE, KMeans-SMOTE, SVMSMOTE, and ADASYN. In particular, Centroid-based SMOTE and Angular SMOTE achieved the highest average G-Mean and Youden's Index scores across both classifiers, while Adaptive Step SMOTE also delivered robust and stable results. These outcomes demonstrate that incorporating geometric awareness into the oversampling process can significantly improve classification performance, especially in challenging imbalanced learning scenarios.

In conclusion, our findings validate the feasibility, effectiveness, and generalizability of the proposed methods across diverse datasets and classifiers. The improvements observed in predictive performance suggest that geometry-aware oversampling holds strong promise for advancing the state-of-the-art in imbalanced classification. Future work will explore the extension of these strategies to multi-class settings and real-time applications, as well as their integration with adaptive parameter tuning mechanisms to further enhance usability and scalability.

References

1. **Chen, Y., Pedrycz, W., Wang, J. (2024).** A New Oversampling Method Based on Triangulation of Sample Space. IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 54, No. 2, pp. 1093–1106. DOI: 10.1109/TSMC.2023.3319694.
2. **Turan, D.S., Ordin, B. (2025).** The Incremental SMOTE: A New Approach Based on the Incremental K-Means Algorithm for Solving Imbalanced Data Set Problem. Information Sciences, Vol. 711, pp. 122103. DOI: 10.1016/j.ins.2025.122103.
3. **Ling, C.X., Li, C. (1998).** Data Mining for Direct Marketing: Problems and Solutions. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), pp. 73–79.
4. **Almubark, I. (2024).** Advanced Credit Card Fraud Detection: An Ensemble Learning Using Random under Sampling and Two-Stage Thresholding. IEEE Access, Vol. 12, pp. 192079–192089.
5. **Liu, S.M., Chen, J.H., Liu, Z. (2023).** An Empirical Study of Dynamic Selection and Random Under-Sampling for the Class Imbalance Problem. Expert Systems with Applications, Vol. 221, pp. 119703. DOI: 10.1016/j.neucom.2025.132361.
6. **Wang, Y., Sun, P. (2023).** Kernel Principle Component Analysis and Random Under Sampling Boost Based Fault Diagnosis Method and Its Application to a Pressurized Water Reactor. Nuclear Engineering and Design, Vol. 406, pp. 112–258.
7. **Wang, G.Y. (2022).** Churn Prediction for High-Value Players in Freemium Mobile Games: Using Random Under-Sampling. Statistika, Vol. 102, No. 4, pp. 443–453. DOI: 10.54694/stat.2022.18.
8. **Bhatt, N., Chandel, A.K. (2021).** A Random Under-Sampling Based Passive Approach for Fast and Accurate Detection of Islanding in Electrical Distribution System. IETE Technical Review, Vol. 39, No. 5, pp. 1152–1166.
9. **Shrividya, G., Bharathi, S.H. (2019).** A Characterization of CS-MRI Reconstruction Using PSO for Random Under Sampling Pattern. International Journal of Recent Technology and Engineering, Vol. 8, No. 11, pp. 519–524.

10. **Wang, H., Liu, Y., Yin, C. (2021).** Fast 3D Time-Domain Airborne EM Forward Modeling Using Random Under-Sampling. Journal of Applied Geophysics, Vol. 191, pp. 104–357. DOI: 10.1016/j.jappgeo.2021.104357.
11. **Baykal, T., Terzi, Ö., Yıldırım, G. (2025).** Investigating the Performance of Random Oversampling and Genetic Algorithm Integration in Meteorological Drought Forecasting with Machine Learning. Discover Applied Sciences, Vol. 7, No. 6, pp. 566.
12. **Chawla, N.V., Bowyer, K.W., Hall, L.O. (2002).** SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, Vol. 16, pp. 321–357. DOI: 10.1613/jair.953.
13. **Han, H., Wang, W.Y., Mao, B.H. (2005).** Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Advances in Intelligent Computing, Vol. 3644, pp. 878–887. DOI: 10.1007/11538059_91.
14. **Douzas, G., Bacao, F., Last, F. (2018).** Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE. Information Sciences, Vol. 465, pp. 1–20. DOI: 10.1016/j.ins.2018.06.056.
15. **Nguyen, H.M., Cooper, E.W., Kamei, K. (2009).** Borderline Over-Sampling for Imbalanced Data Classification. International Journal of Knowledge Engineering and Soft Data Paradigms, Vol. 3, No. 1, pp. 4–21. DOI: 10.1504/IJKESDP.2011.039875.
16. **He, H., Bai, Y., Garcia, E.A. (2008).** ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. IEEE International Joint Conference on Neural Networks, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.
17. **Wang, H., Huang, H. (2018).** LAD-SMOTE: A New Oversampling Method Based on Locally Adaptive Distance. International Conference on Intelligent Control and Information Processing, pp. 305–311.
18. **Bai, L., Sun, M., Jiang, X. (2025).** An Oversampling Method Based on Adaptive Artificial Immune Network and SMOTE. Genetic Programming and Evolvable Machines, Vol. 26, pp. 20. DOI: 10.1007/s10710-025-09516-7.
19. **Pradipta, G.A., Wardoyo, R., Musdholifah, A. (2021).** Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data. IEEE Access, Vol. 9, pp. 74763–74777. DOI: 10.1109/ACCESS.2021.3080316.
20. **Leguen-deVarona, I., Madera, J., Martínez-López, Y. (2020).** SMOTE-Cov: A New Oversampling Method Based on the Covariance Matrix. Data Analysis and Optimization for Engineering and Computing Problems. DOI: 10.13140/RG.2.2.19978.21441.
21. **Dai, F., Liu, Y. (2021).** A Novel Oversampling Method Based on SMOTE and Local Sets for Imbalanced Classification. Journal of Physics: Conference Series, Vol. 2026, pp. 012007.
22. **Li, J., Zhu, Q., Wu, Q. (2021).** A Novel Oversampling Technique for Class-Imbalanced Learning Based on SMOTE and Natural Neighbors. Information Sciences, Vol. 565, pp. 438–455. DOI: 10.1016/j.ins.2021.03.041.
23. **Gong, J. (2021).** A Novel Oversampling Technique for Imbalanced Learning Based on SMOTE and Genetic Algorithm, Neural Information Processing. ICONIP 2021, Lecture Notes in Computer Science, Vol. 13110, pp. 201–212. DOI: 10.1016/j.ins.2021.03.041.
24. **Wang, S., Bao, Y., Yang, S. (2025).** HS-SMOTE: Oversampling Method for Multiple Dynamic Interpolations Based on Regular Hexagon Scoring Mechanism. Expert Systems with Applications, Vol. 265, pp. 125–855.
25. **Wei, Z., Chen, Y. (2025).** NLKF-SMOTE: A Novel Noise-Filtering SMOTE Without Nearest Neighbor Parameter K for Oversampling. Proceedings of the International Conference on Computer Science and Artificial Intelligence (CSAI), pp. 153–159. DOI: 10.1145/3709026.3709035.
26. **Lyu, J., Yang, J., Su, Z. (2025).** LD-SMOTE: A Novel Local Density Estimation-Based

Oversampling Method for Imbalanced Datasets. Symmetry, Vol. 17, No. 2, pp. 160. DOI: 10.3390/sym17020160.

Article received on 07/10/2025; accepted on 09/06/2026.
**Corresponding author is Alican Dogan.*