



A comprehensive review of statistical variants and enhancements of SMOTE oversampling method

Sahar Idwan¹ · Wael Etaiwi² · Haneen Rafayia² · Izzeddin Matar³

Received: 14 December 2024 / Accepted: 11 July 2025

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2025

Abstract

The Synthetic Minority Oversampling Technique (SMOTE) has emerged as an effective and commonly used method for addressing class imbalance in machine learning modeling. This survey investigates the statistical variations of SMOTE, highlighting its enhancement types and techniques. Examining the statistical variants of SMOTE, specifically how synthetic instance production affects model performance, bias, overfitting, and generalization are the main ones. The research also emphasizes recent improvements, such as selection- and adjustment-based approaches. Overfitting, high-dimensional data, and generalization concerns are all covered. This survey provides an in-depth review of the usefulness and limitations of statistically based SMOTE and advises researchers and practitioners on selecting and customizing SMOTE techniques for specific datasets.

Keywords SMOTE · Statistical-based · Selection · Adjustment · Distribution · Overfitting · Generalization

1 Introduction

The learning process plays a predominant role in effectively handling the imbalanced dataset as it assists experts in designing, endorsing, and implementing reliable models throughout several class distributions. Different examples from the real world have an imbalanced dataset [1–3], which includes two kinds of classes: minority class and majority class. The number of instances in the minority class is much smaller than in majority class, which generated a complication in selecting the appropriate classifiers in machine

learning. Many solutions to the class imbalance problem were introduced earlier based on the data level [3] or the algorithmic level [4, 5]. The data level is based on the resampling technique, which includes two approaches: oversampling the minority group or undersampling the majority group. While oversampling focuses on increasing the number of instances in the minority class by generating synthetic instances or replicating the existing ones, undersampling targets to reduce the number of instances in the majority class by randomly removing the existing samples. It is important to note that the oversampling and undersampling techniques have faced challenges, such as overfitting or deleting several important instances. Numerous studies have shown that oversampling outperforms the undersampling technique [6–8]. Therefore, integrating the appropriate resampling methods will guarantee a better capture of the underlying patterns in an imbalanced dataset. The algorithmic level is based on forcing the classifier to consider the minority class during the training process by assigning the imbalanced weights for the majority and minority classes. The algorithmic level is out of the scope of this paper.

Synthetic Minority Oversampling Technique (SMOTE) [1] is the most well-known oversampling technique to address class imbalance in datasets by utilizing feature space relationships between instances of the minority class. The SMOTE method concentrated on the region near the minor-

✉ Sahar Idwan
sahar@hu.edu.jo

Wael Etaiwi
w.etaiwi@psut.edu.jo

Haneen Rafayia
HAN20228039@std.psut.edu.jo

Izzeddin Matar
imatar@uop.edu.jo

¹ Department of Computer Science and Applications, Faculty of Prince Al Hussein Bin Abdullah II for Information Technology, The Hashemite University, Zarqa, Jordan

² Princess Sumaya University for Technology, Amman, Jordan

³ Software Engineering Department, University of Petra, P.O. Box 961343, Amman, Jordan

ity class by generating the synthetic instances along the line between the minority instance and its nearest neighbors of the same class. Since the inception of this method, many studies and research have been published to improve and modify this method considering different points of view to enhance its functionalities in various contexts. Considering the statistical-based strategy to investigate the SMOTE variant methods draws attention to identify and assess their usefulness in handling class imbalance by considering the characteristics of the data such as distance from the decision boundary, density, noise, the qualities of the feature space, determining the number of nearest neighbors and the degree of oversampling. This paper studies different SMOTE variant methods from the statistical-based approach. Moreover, a deeper investigation of how these methods combine with the characteristics of the data will improve the resilience and efficient methodologies to manage the unbalanced dataset.

This paper is organized as follows: This section introduces the distinction and the research methodology. Section 2 represents a general overview of SMOTE. Section 3 presents the statistical SMOTE variants methods from different points of view. A detailed discussion is presented in Sect. 4. Finally, the conclusion and future work are shown in Sect. 5.

1.1 Distinction

Between 2018 and 2024, different researchers discussed the SMOTE variant methods from different perspectives and assessed their efficiency in various fields. However, a comprehensive review of statistical-based SMOTE enhancements, specifically targeting selection- and adjustment-based methods, has been lacking. Therefore, this paper provides a novel categorization of statistical SMOTE variants, systematically classifying them into selection-based and adjustment-based approaches. A comprehensive evaluation of the developed SMOTE versions was presented in terms of both the enhancements made to the weakness found in the original concept, such as the interpolation technique, data overlapping, noise, and insufficient data, and its possible use for more challenging prediction issues [9]. This study notably lists several SMOTE extensions from two perspectives: Data destiny modifications and their adaptations to different applications. The methods reviewed in this paper were investigated from different viewpoints, such as how to initially select minority instances, investigate the filtering procedure, and present the new techniques of interpolation.

Recently, the development of SMOTE variant algorithms has been classified by [10] into many groups according to the goal and area of solving problems in different domains. In this paper, clustering issues, choosing minority samples, adjusting the dimensionality, and handling noise instances were presented from different points of view and their impact on improving outcomes related to the imbalanced dataset.

A comparison of several SMOTE variation methods that improved the accuracy of minority class classification and assessed their performance in various fields was presented by [11]. Moreover, they conducted a comparison between the SMOTE method and other methods that handled the imbalanced dataset, such as random oversampling, random oversampling examples (ROSE), and minority class weighting by presenting their advantages and impacts in this field [11].

The previous studies presented in this section addressed the performance of SMOTE variations by reviewing techniques to enhance the classification efficiency of minority datasets and examining the measurement strategies employed to evaluate the effectiveness of the SMOTE method. Further investigations are essential by considering data characteristics.

Data characteristics encompass specific aspects or qualities of a dataset that are predominant in how well various SMOTE variants perform. It is crucial to comprehend these features, since they are used to identify the SMOTE variations that may be best suited to correct the disparity while maintaining data quality. This study investigates the SMOTE variant methods from the statistical-based viewpoint: selection-based and adjustment-based. Further details are presented in the later subsections. Another key distinction of this paper is its emphasis on post-2020 SMOTE variants, which introduce advanced statistical frameworks for more adaptive and data-driven oversampling strategies. This paper systematically examines these newer contributions, positioning them within the broader landscape of SMOTE research, and highlighting their impact on imbalanced learning. By addressing both the evolution of statistical-based SMOTE methods and the latest advancements, this survey provides a timely and in-depth perspective that not only updates previous taxonomies but also offers researchers practical guidance in selecting appropriate SMOTE variants based on the characteristics and application needs of the dataset.

1.2 Research methodology

The research methodology aims to highlight and compare the proposed statistical SMOTE studies in the state of the art. The methodology includes three stages: research requirements, selection criteria, and analysis.

The research requirements used to govern this study are determined to include formulating a basic understanding of the different statistical SMOTE variants and a comprehensive comparison between the proposed SMOTE variants in terms of their underlying mechanisms.

The search and selection criteria that were applied are listed. The research must be:

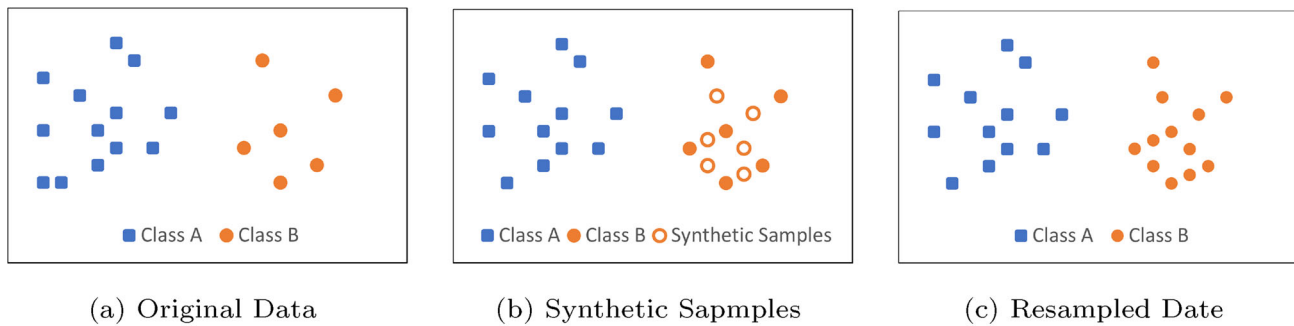


Fig. 1 SMOTE resampling example

- Indexed in one of the following digital databases: IEE-Explore,¹ ACM Digital Library,² ScienceDirect³ or Springer.⁴
- A conference paper or a journal article.
- In English.
- Has one of the following keywords in the paper title, keywords list, or abstract: SMOTE; Synthetic Minority Oversampling Technique; Statistical SMOTE; SMOTE enhancement; improved SMOTE.

The analysis stage aims to comprehensively compare the various studies that address the different variants of statistical SMOTE. A strict search and selection criteria were applied, more than 75 studies were identified, and 39 were selected for the final review. The studies analyzed are classified according to their underlying mechanisms.

2 Background

The enlargement of development in technology and science and the convenience of data caused the need to emphasize the learning of accurate algorithms. SMOTE method [1] has been widely used to manage the imbalanced class in machine learning datasets. The new synthetic instances that are similar to existing minority class instances were created to enhance the performance of the classifiers by generating balanced training data. SMOTE method determines the k nearest neighbor of the instances that belong to the minority class as shown in Fig. 1b. The Euclidean distance metric was employed to generate new instances while preserving the original feature space of the minority class. In SMOTE, the primary factors that influence optimal results and mitigate overfitting include the selection of the number of nearest

neighbors and the proportion of synthetic instances generated. However, since SMOTE generates synthetic samples by interpolating between existing minority class instances, the resulting data may not precisely adhere to the original distribution of the minority class, potentially affecting classification performance [12, 13]. Based on empirical analysis, SMOTE has better accuracy for large-sized data, so its use is not limited to small-sized data or minority class data [14]. It is worth mentioning that SMOTE faces different challenges when the distribution of the minority class is sparse compared with the majority class, which enlarges the overlapping between classes.

3 Statistical-based SMOTE techniques

Several variants of the SMOTE method have been developed to enhance its effectiveness in various scenarios. SMOTE variant methods can be classified into several categories based on their underlying mechanisms and approaches to address the class imbalance in datasets by considering the statistical-based approach to improve its usefulness or implementation to a certain dataset and problem. The timeline for the improvement and enhancement of SMOTE and its variants will be presented in this section. Categorizing SMOTE variant methods based on a statistical-based strategy provides researchers and practitioners with a better understanding of their underlying principles and helps select the most relevant variant.

Several versions of the SMOTE method have been created and developed to overcome the shortcomings of the original SMOTE method. These versions investigated several ways to produce synthetic instances, such as studying the data density and aiming at class boundaries. In this paper, the SMOTE variant methods based on two categories, selection-based and adjustment-based, are represented and discussed.

Selection-based SMOTE represents various SMOTE variant methods by employing the selection criteria, such as distance from the decision boundary, density, noise, and the qualities of the feature space to select the minority class

¹ <https://ieeexplore.ieee.org/>.

² <https://dl.acm.org/>.

³ <https://www.sciencedirect.com/>.

⁴ <https://www.springer.com/>.

samples to be oversampled. Meanwhile, the adjustment-based SMOTE represents the SMOTE variant methods that enhanced the original SMOTE method by determining the nearest neighbor's number and the oversampling degree to create more authentic instances that guarantee the effective representation of the original data distribution. These methods generate more realistic synthetic samples that better reflect the underlying data distribution and enhance the performance of machine learning models on imbalanced datasets.

3.1 Selection-based SMOTE

Selection-based SMOTE represents an improved version of the classic SMOTE method that aims to overcome some of the shortcomings of the SMOTE method by carefully choosing samples and procedures for producing new synthetic data. This section demonstrates in detail different methods that focus on optimizing factors such as distance from the decision boundary, density, noise, and the qualities of the feature space.

1. Distance from the decision boundary: The decision boundary is essential for comprehending the model's categorization parameters. The decision boundary views the level of significance at which an additional instance is classified as either a minority or a majority instance by the model. The likelihood of misclassified instances increases with proximity to the decision border. The distance from the decision boundary indicates the closeness of a sample to the classifier's decision boundary. A group of researchers enhanced the precision of the model on the minority class by recognizing data points that might be incorrectly classified to reduce the model's ambiguity and minimize the violation of separating the minority class from the majority, such as Borderline-SMOTE [15], (Stefanowski et al.) [16], MSYN [17], SMOTE-Cosine [18], SDSMOTE [19], SMOTE-IPF [20], and (Lee et al.) [21].
2. Density: The distribution of data instances of the minority class in the feature space is called a density. Determining how and where to create new instances based on the density term was explored by [22–26], SMOTE-Out, SMOTE-Cosine, and Selected-SMOTE [18, 19, 27, 28], DSMOTE, ESMOTE [29], and OM-SMOTE [30].
3. Noise: Data points that are inaccurate, deceptive, or do not accurately reflect the real data distribution are generated noise. The authors in [16, 26], SMOTE-Out [18, 20, 31], CP-SMOTE, IO-SMOTE [32], and ISMOTE [33] presented different solutions to minimize skew in the learning process.
4. Quality of feature Space: The quality of the feature space refers to the significance and characteristics of the fea-

tures used in the data. The high-quality feature space is essential to improve the effectiveness of the variant methods of SMOTE, which the authors present in [22, 34], SMOTE-Cosine in [18, 28].

Different studies in the literature tackled managing the study region to obtain the optimal representation of the new instances by considering different factors, such as the borderline between classes and the spreading of neighbors from the majority classes. Borderline-SMOTE presented by Han et al. in 2005 [15] selected the nearest neighbors of the minority class, which are close to the decision boundary, and generated synthetic instances for these borderline samples to improve the performance of the classifiers on the imbalanced datasets and preserve the underlying information of the original data. Moreover, Borderline-SMOTE targeted better generalization of the classifier by reducing the bias near the majority class, and it is preferred to be used when the class distribution is very high with complicated decision boundaries. Different experiments were performed on different datasets, such as Satimage, Haberman, Pima, and Circle, using different classification models to compare the Borderline-SMOTE method with other oversampling methods. Metrics such as precision, recall, F1 score, and AUC-ROC are used to assess performance. The findings demonstrated that the Borderline-SMOTE method considerably outperforms classification performance, particularly when minority instances are placed close to the decision boundary. It is important to mention that Borderline-SMOTE is very sensitive to noise instances near the decision boundary, which can cause the creation of inappropriate data.

Another study proposed by Stefanowski and Wilk in 2008 [16] focused on using a selective preprocessing technique to handle the context of the imbalanced dataset. Their strategy focused on detecting borderline and noisy examples within the minority class by selecting oversampling the "safe" instances, which are those situated furthest from the decision boundary, and avoiding oversampling the "noise" instances to enhance the classification process over the minority class. Two learning approaches were examined in the experiments in nine datasets to evaluate sensitivity, specificity, and precision. Their method overcomes the other preprocessing techniques, such as NCR and SMOTE, where the strongly correlated features can impact the effectiveness of oversampling in classification and accuracy.

Managing the whole space effectively to ensure the best estimate of the real class distribution was investigated by TANG and CHEN in 2008 [22]. They proposed a new oversampling algorithm with synthetic examples, Adjusting the Direction Of the Synthetic Minority class examples (ADOMS). The ADOMS method improved the quality of the feature space and improved the classification effectiveness by

creating synthetic instances over the first main component vector of the original data distribution. Twelve datasets were selected from the UCI repository to evaluate the performance of the ADOMS method against the other oversampling methods considering the average value of GM to reduce the degradation of classification quality. It is worth mentioning that the ADOMS method generated synthetic instances based on the density of the minority instances in the feature space to ensure that these examples are both meaningful and realistic. The main challenge in the ADOMS method was determining the impact of directional adjustment during the generation of the synthetic instances and its effect on the decision boundaries discovered by the classifier.

Hu et al. [23] in 2009 presented Modified Synthetic Minority Oversampling (MSMOTE) to manage the selection of the nearest neighbors and utilize the feature space in order to properly create synthetic instances. MSMOTE is a modified method for handling the minority class by focusing on the concept of information density. Synthetic instances were created in the area of the highest density of the minority class rather than in proximity to outliers or noisy data points. The minority class was classified into security, border, and latent noise, considering their neighboring instances. The security and noise groups are related to enhance or reduce the classifier's performance, respectively. On the other hand, instances difficult to classify by the classifier were labeled as border groups. After the classification process, different strategies were used to select new samples according to the minority sample class. Three datasets were used to assess the performance of the MSMOTE method that conclude that the MSMOTE outperformed the SMOTE method in terms of F values. Although the performance of MSMOTE is better than the other methods investigated in this study, handling important features requires further investigation.

A set of variations of the SMOTE method [17, 24, 34] was published in 2011, with the aim of managing the imbalanced class and improving the classifier's performance. The difficulties faced by the oversampling method in managing the huge margin were studied by Fan et al. [17] in 2011 to minimize this risk. They proposed the Margin-guided Synthetic Oversampling (MSYN), which provides more informative instances based on prioritizing the newly created instances close to the classes' boundary. The main step in the MSYN method is the selection process that determines the seed instances from the minority class near the decision boundary to control the degree of oversampling. The feature space characteristics were considered to ensure that the synthetic instances were created within the significant decision limit. Experiments have been conducted on synthetic datasets and ten actual datasets to evaluate the effectiveness of the MSYN method. According to comparisons with other oversampling methods, MSYN performs better than the other approaches in terms of both the AUC and the F -measure. The efficiency

and quality of the MSYN method are based on the selection of appropriate margin thresholds to determine seed examples.

On the other hand, improving the SMOTE method by utilizing dynamic selection has been viewed in the literature. Local Neighborhood Extension of SMOTE (LN-SMOTE) improved the generation of synthetic instances by utilizing a dynamic selection of the closest neighbors in the minority class to improve the quality of these samples was illustrated by Maciejewski and Stefanowski [34] in 2011. The traditional SMOTE method created synthetic examples globally across the dataset without respect for the dominant class. Meanwhile, LN-SMOTE generated synthetic examples by considering only the immediate vicinity of each minority class occurrence. LN-SMOTE addressed a minority class sample's K nearest neighbors and the wider local neighborhood framework, which will guide the creation of synthetic data points within an appropriate and useful region of the feature space. In terms of F -measure and G-means, the LN-SMOTE method achieved better compared to the other approaches presented in this study. The extent of the local neighborhood played an important role in creating the new synthetic instances, which influences the efficiency of LN-SMOTE. The selection of the size of the neighborhood in LN-SMOTE could result in inappropriate instances that affect the classification performance.

Zhang and Wang [24] in 2011 introduced an oversampling method named Distribution-SMOTE based on the distribution between different sets and SMOTE. It is specifically geared to the issue of data crossing in credit data. Distribution-SMOTE implemented the KNN to produce the synthetic instances that considered the intraclass distribution of the minority class by increasing the weight of the instances that may not be on the class borderline. The experiment's findings demonstrated that the Distribution-SMOTE strategy achieves better classification than alternative techniques. Although Distribution-SMOTE keeps the sample distribution in the minority class, the performance varies largely based on the dataset distribution.

García et al. [25] in 2012 proposed a variation of SMOTE that considers the density of the minority class based on the approach of the surrounding neighborhood. Rather than the k nearest neighbors being used in SMOTE, the nearest centroid neighborhood, and the graph neighborhood were used to expand the range of the new artificial instances based on spatial and proximity distribution parameters, which guarantees better capture of decision margins. Also, the surrounding neighborhood guarantees that the synthetic samples are impacted by the local pattern of the feature space rather than by using random interpolations. According to experimental results conducted in thirty-nine datasets with three distinct classifiers (1-NN, C4.5, and MLP), the Surrounding neighborhood method significantly outperforms the original SMOTE algorithm and other oversampling tech-

niques in terms of the AUC metric. This approach has several advantages over the traditional neighborhood, which only considers the minimum distance by spreading the influence region of the minority class. The distance metric used to identify the closest neighbors can be used to assess the performance of the surrounding neighborhood. Different distance metrics may reflect different aspects of sample similarity or be better suited for different data types.

Overgeneralization is considered a major disadvantage in the SMOTE method. Therefore, *Puntumapon and Waiyamai* [26] in 2012 suggested the TRIM-SMOTE method where the algorithm named TRIM was used as a preprocessing step for synthetic oversampling to prevent overgeneralization. TRIM preserves generality while searching for a subset of exact minority examples by implementing two steps: irrelevant data filtering and candidate splitting. The determination of high-density areas to ensure the production of relevant artificial samples was achieved in the irrelevant data filtering step. Then, the candidate splitting step detected and eliminated synthetic instances that were found in the area considered noisy or far away from the decision boundary. The exact set of seeds was fed into a synthetic minority oversampling algorithm. In the end, deleting inappropriate data instances and generating data instances that are significant in the feature space distribution in the minority class satisfies a good representation of the minority class. The outcome of TRIM-SMOTE was presented using eleven datasets with two binary classes, and the performance was estimated using two measurement methods: AUC and *F*-measure. For the Letter-A dataset, TRIM-SMOTE recorded the highest values (97.3%) and (95.2%) for the AUC and *F*-measure, respectively. It should be noted that regions that have been discovered may not be properly generalized to other datasets or real-world cases. The method may unintentionally skew synthetic samples to favor particular traits in the training set, reducing its usefulness in other situations.

Other significant advances in SMOTE variations appeared in 2014 through various studies [18, 19, 27, 35]. To manage scenarios not covered by SMOTE, *Koto* [18] in 2014 presented three SMOTE enhancements: SMOTE-Out, SMOTE-Cosine, and Selected-SMOTE methods. These variations considered three aspects. First of all, handling the spreading of instances that are extremely dense and near to other instances was represented by SMOTE-Out. The second variation considered discovering the nearest neighbor by using the angle or direction of two vectors to determine their similarity; SMOTE-Cosine represents this. Finally, the Selected-SMOTE method selected only informative cases from the minority class, which were used as basis examples. SMOTE-Out, SMOTE-Cosine, and Selected-SMOTE methods adopted the kNN, which has been used in the SMOTE method in different ways to control the generation of the new data points. Utilizing kNN and deleting

exceptions before creating the new synthetic instance was implemented using the SMOTE-Out method. Using feature space to reduce the extent of oversampling in the SMOTE-Out method compared to the SMOTE method. The density has been addressed in SMOTE-Out to manage the overlapping data and reduce noise before creating the synthetic instances. SMOTE-Cosine is targeted to handle the oversampling in the high-density area by considering two parameters of the vector: direction and distance to attain the better nearest neighbor for a minority sample. In addition, it improves the creation of appropriate instances in the feature space area by combining the Euclidean distance and the cosine similarity to find the nearest neighbor. The selection process in the Selected-SMTOE method incorporated significant attributes in the feature space and then implemented kNN to generate the new synthetic instances. The Selected-SMOTE was designed to produce synthetic samples by choosing data points that are crucial for classification and close to the border. Moreover, this selection mechanism reduced the noise by selecting the minority class instances that are most suited to create synthetic samples. Eighteen distinct binary datasets from UCI using B-ACC and *F*-Measure as evaluation methods to examine the performance of the suggested techniques. While the proposed SMOTE-Out represented the worst results for the IL and BCW datasets with values 80 and 97.61 of B-ACC and 58.08 and 96.32 for *F*-Measure, and the best results of the SLS and CE datasets with values 99.44 and 75.62 of B-ACC and 99.85 and 56.73 for *F*-Measure compared to the results achieved by traditional SMOTE. In contrast, the results obtained by SMOTE-Cosine indicate that eight datasets outperform SMOTE, with seven of them providing lower results, which leads to the conclusion that SMOTE-Cosine is insufficient to enhance the performance of the benchmark. The Selected-SMTOE method, which is similar to SMOTE-Out, demonstrated the best results of the CC and SLS datasets with values of 43.19 and 99.11 of B-ACC and 942.38 and 99.75 for *F*-Measure compared to the results achieved by traditional SMOTE. Although the experimental results showed the performance of these methods compared to SMOTE, the efficiency of these methods is related to a set of factors, such as the properties of the balanced dataset and the increasing computational complexity due to the selection of instances or the use of distinct feature spaces.

The determination of the number of new instances and the control of the selection process was considered a major disadvantage of the SMOTE method. *Li et al.* [19] in 2014 introduced SDSMOTE to overcome these challenges. The enhanced SMOTE approach utilized the support degree for a sample to determine the boundary of minority class instances. It generated the new synthetic instances based on the center of the minority class and the radius *k*. The density of the instances in the region was represented by the support degree, which concentrates on the area close to the decision boundary

to guarantee the creation of new instances that represent the low-density instances. In order to assess the performance of SDSMOTE, four datasets with varying ratios of positive to negative classes were used considering the F-value and AUC. Different results generated based on the different techniques were used to compute the support degree, which is considered the main challenge of this approach.

Sandhan and Choi [27] in 2014 enhanced the SMOTE method by utilizing the partially guided selection named G-SMOTE and the bootstrapping method to construct the classifier ensemble. The oversampling G-SMOTE extracted the linear and nonlinear patterns by considering different kinds of correlation in the minority class to avoid over-generalization. Then, the synthetic instances were created by partially utilizing these patterns to direct the random imputation in each feature space. The G-SMOTE oversampled the minority class considering its density pattern in the feature area. Later, the bootstrapping method was used by implementing oversampling in the minority class and undersampling in the majority class to manage the extremely uneven distributions of data. Four distinct imbalanced datasets (SCOP, Satimage, HypoThy, and EuThy) were used to test G-SMOTE. The classification performance was compared with the traditional SMOTE method based on the AUC values and depicted the following results for linear and nonlinear patterns as follows:

- The AUC of the SCOP dataset is 0.832 and 0.838
- The AUC of the Satimage dataset is 0.966 and 0.976
- The AUC of the HypoThy dataset is 0.921 and 0.973
- The AUC of the EuThy dataset is 0.952 and 0.952

It is worth mentioning that the suggested solution in this paper works very poorly with low-imbalance or real-time applications because of the extensive computations that excel its values.

The existence of noise and borderline instances in imbalanced datasets generates a huge challenge for learning classifiers. Therefore, Sáez et al. [20] in 2015 proposed SMOTE-IPF, which combines SMOTE with iterative partitioning filter (IPF). The SMOTE-IPF method started by implementing SMOTE to balance the class distribution, and then the IPF filter was used to delete the noisy instances that existed from the original dataset or instances generated by the SMOTE method. In addition, IPF was responsible for managing instances in the boundary region of the classes to increase their regularity. The work on imbalanced classification using noisy and borderline samples served as the basis for selecting the nine datasets examined in this study. In eleven of the thirty datasets examined, SMOTE-IPF outperforms the other resampling techniques, and in the remaining cases, it achieves results comparable to the top performance. It should be mentioned that the parameter setting for the

implementation of IPF employed in this study has been empirically determined to accurately represent the characteristics of the imbalanced dataset that contains noisy and borderline instances.

Despite its goal of addressing noisy and borderline samples, SMOTE-IPF was unable to manage the datasets when the noise level was high. Also, it is difficult to efficiently filter noisy samples without eliminating instances of the minority class that are helpful.

Lee et al. [21] in 2015 proposed a selective oversampling method that targets the creation of synthetic examples for minority class instances by locating minority class instances near the classifier's decision border and creating synthetic examples specifically for those cases. They aimed to enhance the classifier's performance on the minority class while lowering the danger of overfitting. Two kinds of datasets were chosen to examine the performance of our model: an artificial dataset and eight datasets from the UCI repository. Precision, recall, and G-mean were used to evaluate the performance of the classifier. According to the G-mean value, the proposed method scored the highest value of 0.9519 for the artificial dataset and 0.9997 for the Letter-recognition dataset compared to the other oversampling methods presented in this paper. The success of their oversampling strategy with rejection was strongly dependent on the efficiency of the essential classifier. If the classifier cannot correctly categorize minority class instances, the rejection process may fail to recognize cases near the decision boundary, resulting in poor oversampling.

Capturing the data underlying distribution effectively in high-dimensional areas is considered the most challenging for the original SMOTE method. To address this shortcoming, the Gaussian-based SMOTE algorithm uses Gaussian distributions, proposed by Lee et al. [28] in 2017, to mimic the underlying data distribution. Gaussian-based SMOTE generated synthetic instances by implementing the Gaussian distributions centered on minority class instances rather than linearly interpolating between them, as the traditional SMOTE method does. This enables more flexible and realistic production of synthetic samples to be created in higher-density regions, enhancing the fitness of the feature area and preventing the overgeneralizing problem. A weather forecasting dataset was chosen from the UCI repository, and sensitivity, accuracy, and specificity were selected as performance metrics to validate the suggested Gaussian-based SMOTE approach. The average accuracy of Gaussian-based SMOTE outperforms the other methods with a value of 87.45%, and sensitivity and specificity perform 3% to 5% better than the other methods. Although Gaussian-based SMOTE requires that the underlying data distribution be adequately described using the Gaussian distribution, in real-world situations, data allocation may not always be Gaussian, particularly in high-dimensional environments or

with complicated data structures. Gaussian distributions may not adequately represent the underlying distribution, resulting in biased or unrealistic synthetic sample production.

While the SMOTE method was designed to balance the dataset to enhance the implementation of the classifier in the imbalanced dataset, the SMOTE method considers every sample equitably, which may result in the unawareness of certain misclassified instances. To highlight these issues, Hengyu [29] in 2020 proposed two novel oversampling algorithms, DSMOTE and ESMOTE, based on the SMOTE method. The two versions of the SMOTE method primarily synthesize new samples close to minority samples by using different approaches to compute the misclassification in the sample and increase the weight of this sample. The DSMOTE method utilized the density of minority and majority samples to determine sample weight to create synthetic instances near minority instances. In contrast, the degree of uncertainty named entropy was utilized to determine the sample weight of a minority sample in the ESMOTE method. Samples with higher entropy are assigned greater weights, generating more samples in their neighborhood, and assessing their susceptibility to misclassification. Experiments carried out on eight different datasets reveal that the DSMOTE and ESMOTE algorithms outperform SMOTE in the F1 and G-mean evaluation measures. In addition, the ESMOTE algorithm outperforms the DSMOTE algorithm in the F1 assessment metrics.

Researchers in this field focused on determining the regions of the minority class with a number of instances before initiating the creation process to minimize the overlap between the data in the final result. Therefore, Pradipta *et al.* [31] in 2021 presented a new oversampling method named Radius-SMOTE, which improved the SMOTE method by using the concept of radius distance. Radius-SMOTE determined the k -nearest neighbors of a minority class instance within a given radius instead of creating synthetic samples at random intervals between minority class instances. Synthetic samples were created, restricted to the specified radius, along the line that links the chosen instance and its neighbors. This strategy minimizes the chance of introducing noise by ensuring that synthetic samples are created in areas of the feature space where the minority class is highly populated. For each selected minority class sample, the Euclidean distance is computed relative to all majority class instances. The nearest majority class data point is identified as the one with the shortest distance among all calculated distances for the selected minority sample, as represented by the following equations:

$$r_{ij} = \min \sum_{i=1}^n \sum_{i=1}^n \sqrt{(p_j - t_i)^2} \quad (1)$$

where r_{ij} is the minimum distance of minority sample j to the majority sample i . The performance of Radius-SMOTE was examined by employing random forest classification on selected datasets from the KEEL repository and compared with other variations of SMOTE considering measurement parameters for each class. Choosing the appropriate values for radius and the number of nearest neighbors k played a predominant role in determining the performance of Radius-SMOTE.

The SMOTE method has some drawbacks related to noisy samples and the possibility that the newly created synthetic sample will be less rational and drop the network's classification results. To overcome these difficulties, Bao and Yang [32] in 2023 introduced two new methods: the Center Point SMOTE (CP-SMOTE) method and the Inner and Outer SMOTE (IO-SMOTE) method that target to minimize the number of noise samples that could be used to create new samples. The CP-SMOTE approach created new instances by identifying numerous center points and then linearly integrating the minority examples with their associated center points. The IO-SMOTE approach separated the minority instances into two sets, the inner and outer sets, and then evaluated the uncertainty by using as many inner samples as possible in the process of creating new samples. Measuring the performance of the suggested methods was examined in four datasets, and the results showed that the IO-SMOTE and CP-SMOTE approaches outperform the no-sampling and SMOTE approaches based on different measurement metrics. While the CP-SMOTE approach was aimed at improving the distribution of artificial instances in the minority class, the IO-SMOTE approach was directed to create significant instances close to the boundary. It is worth mentioning that CP-SMOTE and IO-SMOTE may be affected by choosing the parameters and selecting the nearest neighbors.

To address class imbalance while minimizing class overlap, Yulin He *et al.* proposed Overlapping Minimization SMOTE (OM-SMOTE) in 2024 [30]. OM-SMOTE begins by mapping the original data points into a restructured feature space, effectively balancing the sample distribution while preserving classifier generalization. It then employs an advanced set of imputation strategies tailored for the minority class, generating synthetic instances that are strategically positioned to minimize their presence in overlapping regions between classes. This approach improves class separability and reduces the risk of misclassification. Experimental evaluations on thirty-two imbalanced datasets demonstrate that OM-SMOTE consistently outperforms 11 SMOTE variants when applied to naive Bayes, support vector machine, decision tree, and logistic regression classifiers, highlighting its effectiveness in improving classification performance in imbalanced learning scenarios.

On the other hand, to ensure that the new samples are valid and contribute to the classification results, *Jiuxiang Song* and *Jizhong Liu* proposed Iteration-based SMOTE (ISMOTE) in 2024 [33]. Unlike SMOTE, the ISMOTE algorithm iteratively refines the training data over multiple cycles. In each iteration, it generates synthetic samples near strategically misclassified instances, which are then incorporated into the classifier. This iterative process progressively enhances the model's ability to distinguish between classes, leading to improved classification accuracy over successive iterations. Finally, a new sample is generated using the following formula 2:

$$x_{\text{new}} = x_i + r * (x_j - x_i) \quad (2)$$

where x_{new} is the newly generated sample, x_i is the safety data point that is already classified in the previous step, x_j is a randomly selected sample of the same kind, and r is a random number between 0 and 1. Several experiments were operated on twenty datasets obtained from KEEL repository and a comparison with other widely used oversampling methods in order to evaluate the efficacy of the suggested ISMOTE model using three metrics: AUC, F-Score, and G-mean, to ensure that our results were reliable.

SMOTE relies on linear interpolation techniques to generate synthetic data samples, which may be too simplistic and insufficient to capture complex feature spaces. To address this limitation, *Yu Liu* and *Qicheng Liu* proposed in 2025 an enhanced SMOTE algorithm utilizing generative adversarial networks (GANs) to improve the quality and representativeness of synthetic samples, named GAN-SMOTE [36]. GAN-SMOTE uses GANs to generate synthetic minority class samples. The process begins with the KNN algorithm, which identifies the neighboring points for each sample. The sample points are then randomly selected, and their corresponding neighbors serve as real data to train the GAN. Through an adversarial learning process between the generator and the classifier, synthetic data are produced that closely aligns with the local distribution of the minority class. Finally, the newly generated samples are integrated with the original dataset to create a balanced training set, which is then used for classification.

3.2 Adjustment-based SMOTE

Adjustment-based SMOTE represents an enhanced version of the traditional SMOTE method by determining the number of nearest neighbors and the degree of oversampling.

This section highlights in-depth different adjustment methods that target to optimize the original SMOTE method by proactively modifying various parameters of the oversampling process, such as determining the number of nearest neighbors (i.e., the k value) and the degree of oversampling

(i.e., increasing the size of the minority class by using all instances or some of them). While the value of k affects the diversity and distribution of the created samples, the degree of oversampling significantly impacts how well the model performs.

Determining the appropriate parameter k for a specific dataset is challenging. The traditional SMOTE method used a fixed value of $k = 5$, which represents the number of nearest neighbors based on various experiments. However, this approach may generate unsatisfactory outcomes for other datasets. The recent methods tackle this challenge and considering the k value using two ways:

1. Using a fixed value of k : A group of researchers represented SMOTE variant methods by considering the fixed value of k when creating synthetic instance for the minority class, such as *de la Calleja and Fuentes* [37], *He et al.* [38], *Gazzah and Essoukri Ben Amara* [39], *Kang and Won* [40], *Dong and Wang* [41] and *Torres et al.* [42].
2. Using a dynamic value of k : Other methods dynamically determined the value of k considering the local structure of the data, particularly in scenarios with varied densities, such as *Cao and Wang* [43], *Fernandez-Navarro et al.* [44], *Yun et al.* [45] and *Siriseriwan and Sinapiromsaran* [46].

Another factor considered in the proposed SMOTE variant methods is the degree of oversampling, which is considered from different perspectives, such as:

- Considering each instance in the minority class while generating synthetic samples, similar to the SMOTE method, such as *de la Calleja and Fuentes* [37], *Kang and Won* [40], *Dong and Wang* [41], *Cao and Wang* [43] and *Torres et al.* [42].
- Considering a selected portion of minority class instances in the oversampling procedure according to the classifier's sensitivity. The selection is based on different strategies and approaches, as presented in this section, i.e., *He et al.* [38], *Essoukri Ben Amara* [39], *Fernandez-Navarro et al.* [44], *Yun et al.* [45] and *Siriseriwan and Sinapiromsaran* [46].

In this section, a deeper investigation of these issues is explained. The oversampling of the minority class by employing a randomized weighted distance to select the number of the nearest neighbor was represented by *de la Calleja and Fuentes* [37] in 2007. A specific value has been assigned to k based on the data density to compute the k -NN for all minority instances using the weighted distance. The proposed method computes the average of the nearest neighbors of the minority sample to obtain the mean instance. The mean example was created in the experiments by choosing a fixed

K value ($k = 5$). Later, synthetic instances were added to the minority class by computing the distance between the mean instance and all the minority instances, representing the degree of oversampling. These intensive calculations generated an overhead on the system due to the increase in required resources and processing time, especially when managing high-dimensional data. The suggested method was evaluated in ten different datasets from the UCI MLR to determine its efficiency. Three data extraction metrics were used: precision, recall, and F -measure to assess the results of the suggested method for varying levels of oversampling.

A dynamic determination of how many synthetic instances must be created for each minority instance by utilizing the density distribution criterion was presented by *He et al.* [38] in 2008. The authors introduced the Adaptive Synthetic Sampling Approach (ADASYN) to enhance the learning process and minimize the existing bias in the imbalanced dataset. The proposed approach chooses a fixed value of k ($k = 5$) and emphasizes difficult-to-learn instances. ADASYN created more synthetic samples around minority class samples that are hard to classify by giving them high weights. Thus, none of the k 's closest neighbors will contribute equally. Using synthetic objects, ADASYN decreased the class imbalance and adaptively moved the decision boundary toward challenging samples. The total number of synthetic samples to be generated (g_i) is calculated using the following equation:

$$g_i = \hat{r}_i \times G \quad (3)$$

where $\hat{r}_i$ is the normalized value of ratio r_i that is calculated as:

$$r_i = \Delta_i / K \quad (4)$$

where Δ_i is the number of examples in the K nearest neighbors of x_i , where $x_i \in \text{minorityClass}$.

A variety of real-world machine learning datasets was used to evaluate the performance of ADASYN by considering the ROC measurements metric. ADASYN improved the performance of classifiers in imbalanced datasets, increased the accuracy for both majority and minority classes, and reduced the bias toward the majority class. While ADASYN ensured the robustness of the created synthetic instances, the possibility of transferring the overfitting to the majority class was increased.

The novel degree of oversampling methodology, which relates to how much synthetic data are added to the minority class, was presented by *Gazzah and Essoukri Ben Amara* [39] in 2008. Synthetic data instances were generated using polynomial fitting functions to balance the distribution in the minority class. The value of k is fixed ($K = 5$) in the experiments conducted in this research, which includes examples of sixty people's handwriting on an Arabic letter,

while the dynamic adjustment of the value of k is used in the real-life scenario. The effect of the degree of polynomial on the performance of the oversampling was explored by introducing four different approaches to determine the most appropriate degree that guarantees the accurate capture of the minority class distribution. Therefore, the oversampling process was not applied equally to all instances in the minority class. Rather, the proposed method utilized selective oversampling of instances according to the local features of the datasets. Although their work did not handle the feature space directly, the efficiency of the polynomial fitting for oversampling relies on the properties of the feature space.

Kang and Won [40] in 2010 proposed a weight decision algorithm (WDA) to address class imbalance during oversampling and improve the original SMOTE method. The weights were dynamically assigned to the instances in the minority class based on their distribution and the distance associated with the closest neighbor. Obviously, a fixed value of k is selected through the experiments. Selecting the appropriate weight for the minor instance is achieved by utilizing the edge detection algorithm on the spatial area depiction of training data, which determines the number of new samples for each instance. As a result, all minority class samples were considered during the oversampling process on the basis of their respective weights. The WDA utilizing SMOTE as the oversampling technique was evaluated using the SVM classifier. Nine datasets with varying degrees of class imbalance were tested and assessed the categorization performance by using G-mean and F -measure metrics. In this context, employing the derivative approach's edge detection algorithm of the derivative approach required many functions, which increased the computational complexity of this method. The experimental results showed that performance relies on the weighing method, where the structure of the dataset is affected.

Another variation of SMOTE named the Random-SMOTE method, where a random element was generated in minority classes to enhance the scarcity and to envisage unidentified examples in the sparse area, was proposed by *Dong and Wang* [41] in 2011. Random-SMOTE employed a technique different from SMOTE by randomly identifying the total number of synthetic instances in the triangular space to handle the amount of oversampling. The Random-SMOTE method selected a fixed value of K , i.e., ($k = 3$) during the experiments. Based on the selected k , Random-SMOTE created synthetic instances for each instance in the minority class. Random-SMOTE can be useful for non-numerical attributes to enhance sparsity in the minority classes. Experiments were conducted on ten UCI datasets to evaluate the effectiveness of Random-SMOTE compared to three other widely used sampling techniques. The evaluation was performed using the KNN algorithm as the classifier. The experimental results demonstrate that Random-SMOTE effectively handles cases

of absolute rarity and can be applied to datasets with non-numerical and heterogeneous attributes.

A new variation of SMOTE called Synthetic Minority Oversampling Based on Sample Density (SMOBD) was introduced by *Cao and Wang* [43] in 2011. The new method balances the class distribution by combining data density and cost sensitivity. While the data density generated synthetic instances by finding the sparsity inside the minority class, the cost-sensitive highlighted the proper classification by giving various costs to different classes. The new synthetic instances are created by choosing different values of the parameter k based on the density distribution, which guarantees that the amount of generated samples corresponds to the local density of each minority instance. Moreover, the oversampling process was applied to all instances in the minority class. Although this method guarantees the generation of new synthetic instances near the real dataset distribution, SMOBD can struggle with the higher computational complexity required for oversampling. Nine UCI datasets were used to evaluate SMOBD. The proposed method achieved the highest G-means metric in eight out of nine datasets and the highest AUC metric in seven out of nine datasets compared to other SMOTE variants.

A dynamic oversampling method that considers the sensitivity and performance of the output was introduced by *Fernandez-Navarro et al.* [44] in 2011. The proposed method was implemented in the preprocessing stage by using the memetic algorithm (MA) to enhance radial basis functions neural networks (RBFNNs). Two types of oversampling methods were designed to manage class imbalance: the Static SMOTE radial basis function (SSRBF) method and the dynamic SMOTE radial basis function (DSRBF) method. The major difference between these two algorithms depends on the number of times the oversampling process is applied. The SSRBF performed an oversampling step equally to all instances in the minority class to balance the class sizes. In DSRBF, the oversampling process was implemented on selected instances in the minority class during the preprocessing step and in the various generations produced by the MA based on the sensitivity of the classifier. Hence, the distribution pattern and properties of the data for each class define the oversampling rate and the value of k . The SSRBF method is consistent with the traditional implementation of the SMOTE method, in which k is fixed. In DSRBF, the value of k might be dynamic and change depending on the sensitivity of the classifier or the features of the local data. The SSRBF method works well when there is no significant class imbalance with lower computational adequacy. However, the DSRBF method was preferred for difficult datasets with extreme imbalances or complicated class distributions. The results confirm that the dynamic oversampling approach yields promising outcomes, achieving high mean sensitivity in nearly all thirteen datasets analyzed. However, the increase

in sensitivity comes at the cost of reduced accuracy in certain imbalanced datasets.

Later, several studies are proposed to improve the SMOTE method using various technologies, such as deterministic rules and dynamic creation of synthetic instances. A variant of SMOTE that produces synthetic objects for the minority class in a deterministic manner, called SMOTE-D, was presented by *Torres et al.* [42] in 2016. In SMOTE method, the synthetic instances were generated randomly, while in SMOTE-D, deterministic rules were utilized to determine how many new synthetic objects from each minority class item should be created and where these new objects should be located. It could be difficult to produce significant synthetic instances in areas with sparse data, which could cause problems with the quality of the oversampling using SMOTE-D. To evaluate the proposed method, 66 datasets from the KEEL repository were used. The performance of SMOTE-D was compared with SMOTE using decision trees, support vector machines (SVM), and KNN based on the F -measure. The results indicate that the proposed method achieved better performance in 67% of the datasets with decision trees, 61% with KNN, and 73% with SVM. On average, it demonstrated a statistical improvement of 3.14%, 2.73%, and 4.26% in F -measure for decision trees, KNN, and SVM, respectively.

The traditional SMOTE and its derivatives performed poorly in classification when a dataset has complicated patterns, small sub-clusters, or noise. Therefore, independently creating synthetic minority samples with the automatically predefined k values was proposed by *Yun et al.* [45] in 2016. The proposed method introduced automatic neighborhood size determination (AND) based on selecting the appropriate neighborhood size that increases the effectiveness of SMOTE. The selection process was handled by a method that automatically set the size of the neighborhood based on the minimum description length (MDL) basis to obtain the balance between the model complexity and its rightness of fit. As a result, less dense samples are assigned a larger k value, and more densely packed instances are assigned a smaller k value, which leads to oversampling only on the minority instances with lower data density. Automatic determination techniques could not perform effectively when dealing with datasets that deviate from the training set or contain previously unseen data. This may affect the algorithm's robustness and dependability regarding cross-domain scenarios or fresh datasets. Twelve datasets from the KEEL dataset repository were utilized, considering variations in the number of samples, attributes, and imbalance ratios. The classification and regression tree (CART) classifier was employed. The experimental results demonstrate that the proposed method outperformed existing approaches at a high noise level (15%) in terms of AUC-ROC and AUC-PR.

To dynamically change the number of neighbors required for oversampling in various minority regions, a new adap-

tive method called the Adaptive Neighbor Synthetic Minority Oversampling Technique (ANS) was introduced by *Siris-eriwan and Sinapiromsaran* [46] in 2017. The minority outcast was introduced to the represented instance without neighbors. Although most oversampling strategies ignored minority outcasts, ANS combined SMOTE, outlier handling, and an adaptive approach to provide a solution to performance improvement. Based on the quantity of nearby positive examples, ANS automatically determines the value of the parameter k for a positive instance. In addition, ANS managed minority outcasts to maintain their value without creating artificial instances. The local features of the data can affect how much oversampling occurs. This indicates that the method is flexible and adapts the number of synthetic samples produced to the requirements of various data regions rather than being restricted to a certain oversampling ratio. To evaluate the effectiveness of ANS compared to other oversampling techniques, experiments were conducted on nine datasets from the UCI repository as well as five datasets from the PROMISE repository. The study used five classifiers. The results indicate that ANS achieved the highest number of cases in which it yielded the best F -measure and AUC values. To confirm its statistical superiority over other techniques, Wilcoxon signed-rank tests were applied. It is worth mentioning that ANS faced difficulties related to cases where inappropriate instances bounded the minority examples from the majority class.

4 Discussion

The number of statistical-based SMOTE variations has increased in recent years, as shown in Tables 1 and 2, emphasizing selection-based and adjustment-based approaches; this is due to increased data availability and the diversity of data characteristics, which necessitates different approaches to handle each data distribution.

In the selection-based SMOTE variants, different data characteristics were considered in the proposed methods, such as distance from the decision boundary, density, noise, and the qualities of the feature space. Each approach proposed in Table 1 aims to deal with some data with one or more of the aforementioned characteristics; additionally, they aim to address certain problems and eliminate some drawbacks. In this context, improving generalization is the most appealing challenge, motivating researchers, such as Distribution-SMOTE [24], Surrounding neighborhood-based SMOTE [25], TRIM-SMOTE [26], G-SMOTE [27], Gaussian-based SMOTE [28], and DSMOTE [29], to assist in dealing with and reducing the risk of overgeneralization, which refers to creating synthetic samples that overgeneralize the minority class, potentially introducing noise and reducing model performance. Furthermore, the proposed

statistical selection-based SMOTE versions aim to address two of the most challenging problems: producing noise-free samples and minimizing overfitting, such as MSYN [17], SMOTE-IPF [20], MSMOTE [23], and Radius-SMOTE [31] and (*Lee et al.*) [21], ADOMS [22], G-SMOTE [27], and ISOMTE [33], respectively.

Adjustment-based SMOTE variations improve SMOTE by adjusting specific parameters, such as the number of nearest neighbors (fixed or dynamic k value) and the degree of oversampling (considering some or all minority instances). As shown in Table 2, the proposed adjustment-based SMOTE variants aim to improve SMOTE performance and efficiency by enhancing its processes rather than dealing with data characteristics. For example, a specified fixed value of k is frequently utilized due to its simplicity and low complexity. However, random weighted distance methods are prevalent among researchers for the same reason. Most of the proposed adjustment-based SMOTE variations are successful in preserving the minority class distribution, such as (Gazzah and Essoukri Ben Amara) [39], SMOTE-D [42], SMOBD [43], and ANS-SMOTE [46], and others are successful in dealing with different and complex feature space characteristics such as (de la Calleja and Fuentes) [37], ADASYN [38], (*Kang and Won*) [40], Random-SMOTE [41], and (*Yun et al.*) [45]. On the other hand, improving internal SMOTE processes may have certain problems, such as relying on technique performance to improve SMOTE, creating noise instances, and giving less generic solutions that function well in specific scenarios rather than all cases (i.e., sparse data).

The effectiveness of SMOTE and its variants is highly influenced by key hyperparameters, such as the number of nearest neighbors (k), oversampling ratios, and radius thresholds. Several practices could be employed to fine-tune these parameters and select the most appropriate parameter values, such as grid search and cross-validation. These techniques are utilized to identify suitable values that balance synthetic data generation without introducing excessive noise or overfitting. While there is no one-size-fits-all approach, prior studies suggest that setting k between 3 and 7 often yields stable results, while oversampling ratios should be adjusted based on the degree of class imbalance. Methods such as adaptive SMOTE and density-based approaches (i.e., DBSMOTE) further refine parameter selection by considering the local distribution of minority samples. Incorporating automated hyperparameter tuning techniques, such as Bayesian optimization or evolutionary algorithms, can further enhance parameter selection in future work.

5 Conclusion

In this article, several statistical-based SMOTE variants are studied, analyzed, and classified into two groups: selection-

Table 1 Selection-based SMOTE variant methods

Method/(Author(s))	Year	Methodology	Data characteristics	Advantages	Disadvantages
Borderline-SMOTE [15]	2005	Oversampling the closest instances to the borderline in the minority class	Distance	Useful when the decision boundary is important	Extra overhead in determining the percentage of oversampling and the number of nearest neighbor and
(Stefanowski et al.) [16]	2008	Combining filtered samples from the majority classes with local oversampling of samples from the minority class	Distance and Noise	Identifying difficult regions in the minority class and modifies only these samples	The efficiency relies on the way of choosing system parameters
ADOMS [22]	2008	Using directional adjustment	Density and Features	Reduce overfitting	The performance relies on the directional adjustment
MSMOTE [23]	2009	Considering the distribution in the neighbor selection process	Density	Eliminate noise samples	Not considering the features importance
MSYN [17]	2011	Prioritizing the newly created instances that are close to the boundary	Distance	Generate no noise samples	The efficiency relies on selecting the appropriate margin thresholds
LN-SMOTE [34]	2011	Dynamic selection of the closest neighbors in the minority class	Features	Considering more local neighborhood of the seed minority example	Generating inappropriate instances
Distribution-SMOTE [24]	2011	Sampling within-class distribution of the minority groups	Density	Keeping the sample's distribution in minority class	The performance varies largely with different dataset distribution
Surrounding neighborhood-based SMOTE [25]	2012	Based on the spatial distribution and the proximity to the selected sample using surrounding neighborhood	Density	Spread the influence region of the minority class	The efficiency relies on the selected distance metrics and the data types
TRIM-SMOTE [26]	2012	Using TRIM algorithm as a preprocessing	Density and Noise	Preventing overgeneralization	The performance declines as the sampling percentage increases
SMOTE-Out [18]	2014	Selecting the extremely dense and near to one another	Density and Noise	Dealing with dense the distribution	Increasing computational complexity
SMOTE-Cosine [18]	2014	Using the angle or the direction of two vectors	Density and Features	Considering distance and the direction of samples	Increasing computational complexity
Selected-SMOTE [18]	2014	Selecting the informative samples	Distance and Density	Enriching the variation of minority samples	Increasing computational complexity

Table 1 continued

Method/(Author(s))	Year	Methodology	Data characteristics	Advantages	Disadvantages
SDSMOTE [19]	2014	Considering support degree to determine the boundary of the minority class	Distance and Density	Increasing the ability to improve the positive class's classification	Different results generated based on the different techniques were used for computing the support degree
G-SMOTE [27]	2014	Utilizing the second-order and higher-order correlation + Bootstrapping	Density	Preserving minority class features and avoiding overfitting and overgeneralization	The performance relies on the selected pattern classifiers
SMOTE-IPF [20]	2015	Combining the SMOTE with the IPF	Distance and Noise	Regulating noise and improving the regularity of the class boundaries	Could not manage datasets when the noise level was high
(Lee <i>et al.</i>) [21]	2015	Locating minority class instances near the classifier's decision border	Distance	Minimizing overfitting	The performance depends on the efficiency of the essential classifier
Gaussian-based SMOTE [28]	2017	Using Gaussian distributions	Density and Features	Mimic the underlying data distribution	May not adequately represent the underlying distribution
DSMOTE [29]	2020	Computing the misclassification in the sample based on the density of neighboring minority and majority samples	Density	Improving generalization	Increasing computational complexity
ESMOTE [29]	2020	Computing the misclassification in the sample based on information entropy	Density	Improving generalization	Increasing computational complexity
Radius-SMOTE [31]	2021	Using radius distance	Noise	Lowering the chance of introducing noise	Choosing the appropriate values for the radius played a predominant role in determining the performance
CP-SMOTE [32]	2023	Based on identified center points	Noise	Reducing error in the training data	—

Table 1 continued

Method / (Author(s))	Year	Methodology	Data characteristics	Advantages	Disadvantages
IO-SMOTE [32]	2023	splitting minority examples into inner and outer instances	Noise	Reducing error in the training data	–
OM-SMOTE [30]	2024	employs an advanced set of imputation strategies	Density and Features	Minimizing class overlap	Parameter sensitivity and potential loss of minority class diversity
ISMOTE [33]	2024	Iterative training and random selection from the safe list	Noise	Lowering the chance of introducing noise	Increasing computational complexity
GAN-SMOTE [36]	2025	Using generative adversarial networks	Density and features	Dealing with complex data feature spaces	Increasing computational complexity

Table 2 Adjustment-based SMOTE variant methods

Method /(Author(s))	Year	Methodology	K value	Degree of oversampling	Advantages	Disadvantages
(de la Calleja and Fuentes) [37]	2007	Employing a randomized weighted distance	Fixed	Fully	Operate in the feature space rather than the data space	increasing the resource requirements and processing time
ADASYN [38]	2008	Utilizing the density distribution	Fixed	Partially	More emphasis on difficult-to-learn instances and reducing the bias toward the majority class	Transferring the overfitting to the majority class
(Gazzah and Essoukri Ben Amara) [39]	2008	Utilizing the polynomial fitting functions	Fixed	Partially	Guarantees the accurate capturing of the minority class distribution	The efficiency relies on the used polynomial function
(Kang and Won) [40]	2010	Using a Weight Decision Algorithm	Fixed	Fully	Prioritize oversampling based on sample features	The performance relies on the weighing method, and increasing the computational complexity
Random-SMOTE [41]	2011	Random Sample Selection process	Fixed	Fully	Improving sparseness' cases	Selecting the appropriate random seed
SMOBD [43]	2011	Combining the data density and the cost sensitivity	Dynamic	Fully	Generate new synthetic instances that are near to the real dataset's distribution	High computational complexity
SSRBF [44]	2011	Combining sampling methods into RBFNNs optimized by an MA which implemented at preprocessing step only	Static	Fully	simplicity and less expensive	High level of misclassification and work poorly for the diverse dataset
DSRBF [44]	2011	Combining sampling methods into RBFNNs optimized by an MA implemented at the preprocessing stage and various stages of MA	Dynamic	Partially	improving the performance of the classifier	Difficult to implement with high computational complexity
SMOTE-D [42]	2016	Based on the feature space density	Fixed	Fully	Considering the minority class distribution	Difficult to apply for sparse data
(Yun et al.) [45]	2016	Based on the minimum description length (MDL) basis	Dynamic	Partially	Useful when a dataset has complicated patterns, small sub-clusters, or noise	Could not perform effectively for all datasets
ANS-SMOTE [46]	2017	combining SMOTE, outlier handling, and adaptive approach	Dynamic	Partially	Considering the minority outcasts	Encountering issues when inappropriate instances circumscribed minority examples from the majority class

based SMOTE and adjustment-based SMOTE. By systematically analyzing these variants, their key advantages, limitations, and applicability across different data scenarios are highlighted. Selection-based SMOTE methods improve sample selection processes, leveraging distance metrics, density estimations, and noise filtering to refine the oversampling process. Adjustment-based SMOTE methods focus on optimizing nearest-neighbor selection (k) and oversampling ratios, ensuring more adaptive data balancing strategies. The selection-based SMOTE variants are further classified based on the processed data characteristics, i.e., distance from the decision boundary, density, noise, and the qualities of the feature space, while the adjustment-based SMOTE variants are further classified based on the type of enhancement proposed, i.e., the number of required nearest neighbors and the degree of oversampling.

There are promising future work directions to work on in this area such as (1) conducting an empirical evaluation to compare the different variants' performance on real datasets, (2) more new SMOTE variants can be developed to improve the quality of individuals by combining multiple variants into one method or proposing an ensemble of SMOTE variants that capture the benefits of multiple variants into one new method, and (3) analyzing the complexity of different statistical-based SMOTE variants and validating them on a large dataset.

References

- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
- Zheng, Z., Wu, X., Srihari, R.: Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explor. Newsl.* **6**, 80–89 (2004)
- Kubát, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. In: *International Conference on Machine Learning* (1997)
- Domingos, P.: Metacost: a general method for making classifiers cost-sensitive. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD99*, ACM (1999)
- Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., Brunk, C.: Reducing misclassification costs: knowledge-intensive approaches to learning from noisy data. In: *Proceedings of the Eleventh International Conference on Machine Learning, ML-94*. New Brunswick: New Jersey, pp. 100–109 (1994)
- Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **6**, 20–29 (2004)
- He, G., Han, H., Wang, W.: An over-sampling expert system for learning from imbalanced data sets. In: *2005 International Conference on Neural Networks and Brain*, vol. 1, pp. 537–541 (2005)
- Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: *Proceedings of the 24th International Conference on Machine Learning, ICML '07 & ILP '07*, ACM (2007)
- Fernández, A., García, S., Herrera, F., Chawla, N.: Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **61**, 863–905 (2018)
- Pradipta, G.A., Wardoyo, R., Musdholifah, A., Sanjaya, I.N.H., Ismail, M.: Smote for handling imbalanced data problem: a review. In: *2021 Sixth International Conference on Informatics and Computing (ICIC)*, pp. 1–8 (2021)
- Nemade, B., Bharadi, V., Alegavi, S.S., Marakarkandy, B.: A comprehensive review: smote-based oversampling methods for imbalanced classification techniques, evaluation, and result comparisons. *Int. J. Intell. Syst. Appl. Eng.* **11**, 790–803 (2023)
- Elreedy, D., Atiya, A.F., Kamalov, F.: A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. *Mach. Learn.* **113**, 4903–4923 (2023)
- Kamalov, F., Choutri, S.E., Atiya, A.F.: Analytical formulation of synthetic minority oversampling technique (smote) for imbalanced learning. *Gulf J. Math.* **19**, 400–415 (2025)
- Elreedy, D., Atiya, A.F.: A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Inf. Sci.* **505**, 32–64 (2019)
- Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) *Advances in Intelligent Computing*, pp. 878–887. Springer, Berlin (2005)
- Stefanowski, J., Wilk, S.: *Selective Pre-processing of Imbalanced Data for Improving Classification Performance*, pp. 283–292. Springer, Berlin (2008)
- Fan, X., Tang, K., Weise, T.: *Margin-Based Over-Sampling Method for Learning from Imbalanced Datasets*, pp. 309–320. Springer, Berlin (2011)
- Koto, F.: Smote-out, smote-cosine, and selected-smote: an enhancement strategy to handle imbalance in data level. In: *2014 International Conference on Advanced Computer Science and Information System, IEEE* (2014)
- Li, K., Zhang, W., Lu, Q., Fang, X.: An improved smote imbalanced data classification method based on support degree. In: *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*. IEEE (2014)
- Sáez, J.A., Luengo, J., Stefanowski, J., Herrera, F.: Smote-ipf: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Inf. Sci.* **291**, 184–203 (2015)
- Lee, J., Kim, N.-R., Lee, J.-H.: An over-sampling technique with rejection for imbalanced class learning. In: *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, IMCOM '15*, ACM (2015)
- Tang, S., Chen, S.-P.: The generation mechanism of synthetic minority class examples. In: *2008 International Conference on Technology and Applications in Biomedicine*. IEEE (2008)
- Hu, S., Liang, Y., Ma, L., He, Y.: Msmote: improving classification performance when training data is imbalanced. In: *2009 Second International Workshop on Computer Science and Engineering*. IEEE (2009)
- Zhang, L., Wang, W.: A re-sampling method for class imbalance learning with credit data. In: *2011 International Conference of Information Technology, IEEE, Computer Engineering and Management Sciences* (2011)
- García, V., Sánchez, J.S., Martín-Félez, R., Mollineda, R.A.: Surrounding neighborhood-based smote for learning from imbalanced data sets. *Prog. Artif. Intell.* **1**, 347–362 (2012)
- Puntumapon, K., Waiyamai, K.: A Pruning-Based Approach for Searching Precise and Generalized Region for Synthetic Minority Over-Sampling. pp. 371–382. Springer, Berlin (2012)
- Sandhan, T., Choi, J.Y.: Handling imbalanced datasets by partially guided hybrid sampling for pattern recognition. In: *2014 22nd International Conference on Pattern Recognition*. IEEE (2014)

28. Lee, H., Kim, J., Kim, S.: Gaussian-based smote algorithm for solving skewed class distributions. *Int. J. Fuzzy Logic Intell. Syst.* **17**, 229–234 (2017)
29. Hengyu, Z.: Improved smote algorithm for imbalanced dataset. In: 2020 Chinese Automation Congress (CAC). IEEE (2020)
30. He, Y., Lu, X., Fournier-Viger, P., Huang, J.Z.: A novel overlapping minimization smote algorithm for imbalanced classification. *Front. Inf. Technol. Electron. Eng.* **25**, 1266–1281 (2024)
31. Pradipta, G.A., Wardoyo, R., Musdholifah, A., Sanjaya, I.N.H.: Radius-smote: a new oversampling technique of minority samples based on radius distance for learning from imbalanced data. *IEEE Access* **9**, 74763–74777 (2021)
32. Bao, Y., Yang, S.: Two novel smote methods for solving imbalanced classification problems. *IEEE Access* **11**, 5816–5823 (2023)
33. Song, J., Liu, J.: Ismote: a more accurate alternative for smote. *Neural Process. Lett.* **56**, 240 (2024)
34. Maciejewski, T., Stefanowski, J.: Local neighbourhood extension of smote for mining imbalanced data. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). IEEE (2011)
35. Xu, Y.H., Li, H., Le, L.P., Tian, X.Y.: Neighborhood triangular synthetic minority over-sampling technique for imbalanced prediction on small samples of Chinese tourism and hospitality firms. In: 2014 Seventh International Joint Conference on Computational Sciences and Optimization. IEEE (2014)
36. Liu, Y., Liu, Q.: Smote oversampling algorithm based on generative adversarial network. *Clust. Comput.* **28**, 271 (2025)
37. de la Calleja, J., Fuentes, O.: A distance-based over-sampling method for learning from imbalanced data sets. In: The Florida AI Research Society (2007)
38. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE (2008)
39. Gazzah, S., Amara, N.E.B.: New oversampling approaches based on polynomial fitting for imbalanced data sets. In: 2008 The Eighth IAPR International Workshop on Document Analysis Systems. IEEE (2008)
40. Kang, Y.-I., Won, S.: Weight decision algorithm for oversampling technique on class-imbalanced learning. In: ICCAS 2010. IEEE (2010)
41. Dong, Y., Wang, X.: A New Over-Sampling Approach: Random-SMOTE for Learning from Imbalanced Data Sets, pp. 343–352. Springer, Berlin (2011)
42. Torres, F.R., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: SMOTE-D a Deterministic Version of SMOTE, pp. 177–188. Springer, Berlin (2016)
43. Cao, Q., Wang, S.: Applying over-sampling technique based on data density and cost-sensitive svm to imbalanced learning. In: 2011 International Conference on Information Management, Innovation Management and Industrial Engineering. IEEE (2011)
44. Fernández-Navarro, F., Hervás-Martínez, C., Antonio Gutiérrez, P.: A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recognit.* **44**, 1821–1833 (2011)
45. Yun, J., Ha, J., Lee, J.-S.: Automatic determination of neighborhood size in smote. In: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, IMCOM '16, ACM (2016)
46. Siriseriwan, W., Sinapiromsaran, K.: Adaptive neighbor synthetic minority oversampling technique under 1nn outcast handling. *Songklanakarin J. Sci. Technol.* **39**, 5 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.