Advances in Artificial Intelligence and Machine Learning; Research 4 (3) 2703-2717 Received 15-07-2024; Accepted 21-09-2024; Published 28-09-2024

Active Down-Sampling Method for Knn When Dealing with Imbalance Dataset

Murad Mustafa Badarna mbadarna@gmail.com
Department of Information Systems,
The Max Stern Yezreel Valley Academic College.

Loai Cameel AbedAllah loaia@yvc.ac.il
Department of Information Systems,
The Max Stern Yezreel Valley Academic College.

Corresponding Author: Loai Cameel AbedAllah.

Abstract

This study introduces Active Down-sampling (ADS), a novel approach combining down-sampling with active learning to select informative samples from the majority class in imbalanced data scenarios, thereby enhancing machine learning model performance. Tested on three real-world datasets (BLOOD, Yeast, and Ecoli), ADS demonstrates superior classification accuracy over existing methods, efficiently balancing dataset representation while saving computational resources. It boosts accuracy across both minority and majority classes, optimizes resource use, and reduces misclassification costs. It emerges as a promising solution to the prevalent issue of data imbalance in machine learning, offering significant performance, resource, and cost advantages.

Keywords: Imbalanced data, Under sampling, Selective sampling

1. INTRODUCTION

A classification data set with skewed class proportions is called imbalanced. Classes that make up a large proportion of the data set are called majority classes. Those that make up a smaller proportion are minority classes. When the data is imbalanced, it is a problematic case for many supervised learning algorithms. Since, with so few positives relative to negatives, the training model will not learn enough from positive ones and spends most of its learning process on the negative class.

Imbalanced datasets are a common issue in many real-world classification problems, including predicting manufacturing equipment failures [1], detecting spam emails [2], identifying fraudulent credit card transactions [3], diagnosing medical diseases [4], detecting cyber-attacks [5], and many others, the classes are not represented equally. This inequality appears on two-class classification problems as well as on multi-class classification problems.

Citation: Murad Mustafa Badarna and Loai Cameel AbedAllah. Active Down-Sampling Method for Knn When Dealing with Imbalance Dataset. Advances in Artificial Intelligence and Machine Learning. 2024;4(3):157.

Many resampling methods were used for solving the imbalanced dataset problem. Overall, these methods fall into three groups [6, 7]:

(1) The first group is the over-sampling methods that create new minority class samples to eliminate the problem of the skewed distribution.

(2) The second group is the under-sampling method that discards the intrinsic samples in the majority class to deal with the problem of the skewed distribution.

(3) The third group is the Hybrid methods that combine the over-sampling method and the under-sampling method.

In this paper, we introduced a novel sampling algorithm belonging to the under-sampling group, referred to as the second group in our study. Existing under-sampling methods in the literature have a limitation: although they specify the samples to be selected, they do not consider the representation importance of each selected point in the data. Additionally, these methods overlook the significance of each selected point in solving the classification problem. Much research has been done on under-sampling methods for imbalanced dataset. The simplest yet most effective method is random under-sampling, which involved the random elimination of majority class examples [8–10].

It is very important to note that there is no single sampling method that is best suitable for all types of datasets and classifiers. Therefore, previous works suggested techniques which combined and serve specific type of learning algorithm such as [11–13], where they suggested dedicated approach for SVM classifier and [14], which proposed dedicated solution for logistic Regression model. The algorithm that suggested in [15], uses the K-means clustering algorithm for filtering the data from noise and weak points and its selection is not based on the points contribution.

Our study distinguishes itself from previous studies in the following way: Our method can deal with a multiclass dataset where each iteration considers only the two nearest neighbors. Our experiments demonstrate that this method consistently outperforms traditional resampling techniques across multiple metrics, making it a valuable tool for addressing class imbalance in a wide range of machine learning scenarios.

The rest of this article is organized as follows. Section 2 include discussion of related previous work. Section 3 describes how the proposed model addresses the imbalance problem using the active learning technique for the k-nearest neighbors' classifier. In section 4, we evaluate the performance of the proposed algorithm with other resampling techniques using real-world datasets. Finally, we conclude our research in section 5.

2. RELATED WORK

Many resampling methods were used for solving the imbalanced dataset problem for KNN classifier. The imbalanced feature-mixing K-center method (IFMK) [16], tailored for unbalanced datasets, might not directly apply to KNN effectively due to interpolation complexities, potentially failing to fully address class imbalance issues. The incremental clustering-based under-sampling method aims to enhance KNN's performance through feature-sum clusters, yet it might not capture all minority class variations adequately, risking the omission of critical instances [17]. The Neighborhood based Under-Sampling (N-US) algorithm tries to increase the visibility of minority data points without excessive reduction of majority points, but it may retain redundant majority samples, weakening the minority class's impact [18].

Methods that emphasize the difficulty and proportion of samples in active learning could theoretically provide a more balanced dataset handling for KNN, but they may not adapt well to the complex and dynamic nature of real-world datasets [19]. Hybrid approaches that combine preprocessing with ensemble learning, such as stochastic sample evaluation followed by Multilayer Perceptron training, might introduce unnecessary complexity to KNN, a model valued for its simplicity [20]. The ALIS framework, which alternates between major and minor class instance selection, may not sufficiently adjust its general strategy to meet KNN's specific requirements for data proximity [21]. Active balancing mechanisms using Gaussian naive Bayes and entropy for under-sampling proposed by Hongyi et. al., however, their reliance on Gaussian distribution assumptions may limit their applicability across diverse datasets [22].

Moreover, modifying active learning algorithms for KNN by changing acquisition functions to introduce a balance step could potentially enhance performance but at the cost of complicating the acquisition process and efficiency [23]. Iterative methods that generate samples and calculate validation scores for training refinement may significantly increase computational demands, delaying model deployment [24]. Enhancements to SMOTE focusing on minority samples and addressing noise and processing time issues aim to improve handling of imbalanced datasets for KNN, yet they risk increasing noise and computational complexity in the preprocessing stage [25].

Many other works can be found in literature that suggested techniques which combined and serve the SVM learning algorithm. Maheshwari et al. [26], suggested merging the random under-sampling method within the Boosting method of the DataBoost-IM model to improve the precision of the base classifiers. Sahni et al. [27], propose a framework to find a relation between datasets and sampling methods via a set of meta-features that characterizes the distribution of data. D'Addabbo and Maglietta [28], described an approach that selects data from the majority class to reduce imbalance data sets when using SVM classifier. Relevant examples from the majority class are selected and used in the successive classification step using SVM. A similar approach was proposed by Anand et al.[11], for the weighted SVM. It selects instances from the majority class which are more likely to be near a decision boundary. The proposed under sampling technique is then combined with the weighted-SVM.

Other works proposed ensemble clustering based techniques to deal with imbalance dataset. Sun et al. [12], suggested converting an imbalanced dataset into multiple balanced ones and then builds number of classifiers on these multiple data with a specific classification algorithm. Li et al. [13], described an ensemble SVM with segmentation for the classification of imbalanced datasets where a vector quantization algorithm is used to segment the majority class in order to improve the prediction accuracy of the minority class. Kumer et al. [15], proposed a method which removes noisy and weak instances from majority class to deal with the imbalance distributed data. Only the stronger instances are combined with the minority class dataset to be used for the training of the well-known K-means. Recently, a work by Wonjae et al. [14], proposed a down-sampling method using active learning to mitigate the effect of class imbalance. Here, the samples selection, based on optimal experimental design criteria, minimizes the generalization error in a penalized logistic Regression model.

Our study distinguishes itself from previous studies in the following way: Our method can deal with a multiclass dataset where each iteration considers only the two nearest neighbors. In conclusion, our study offers a significant advancement over existing methods by providing a more straightforward yet highly effective approach to dealing with imbalanced datasets in the context of KNN classifiers. By focusing on the two nearest neighbors during each iteration, our method reduces computational complexity while maintaining the integrity of the minority class. This approach not only simplifies the resampling process but also ensures that the classifier remains sensitive to critical minority instances, thus avoiding the pitfalls of overfitting and noise introduction that can plague other methods.

3. ACTIVE DOWN-SAMPLING ALGORITHM

Our algorithm is inspired by the concept of selective sampling for nearest neighbor classifiers (LSS) [29, 30], which introduces a selective sampling methodology for nearest-neighbor (NN) classification learning algorithms. At each iteration, LSS examines all samples from the majority class and selects the point that contributes the most to the expected accuracy, gradually building a balanced representation of the majority class until the desired balance is achieved.

In this research, we propose an active down-sampling technique to address the challenges posed by imbalanced datasets. The technique aims to identify the most informative points within the dataset, with the primary objective of selecting relevant points from the majority class to complement the minority class. The information value of each point is computed based on its impact on the classifier's accuracy. The method quantifies the information rate for each point and selects those that have the highest potential to improve the overall classification accuracy.

FIGURE 1 depicts the selective sampling algorithm steps. Our algorithm facilitates the creation of a balanced dataset, BD, from an imbalanced labeled dataset, D. The process begins by adding all points from the minority class, denoted as C_0, to the new subset BD. Subsequently, the algorithm aims to select n_0 points from the majority class, C_1, that best represent it. It leverages the information value of each point in C_1 to determine the optimal selection. The information value is computed by evaluating the contribution of a point to the classifier's accuracy.

The algorithm's core approach involves using a conditional probability-based analysis to evaluate the accuracy of BD by considering the nearest neighbors of each point in the original dataset D. Specifically, for a given point x in D, the k nearest neighbors from BD $(z_1, z_2, \ldots, z_k)$ are found, and the probability of x being correctly classified as class 0 or class 1 is calculated as:

$$p\left(z_1, z_2, \ldots, z_k\right) = \frac{\left|\left\{1 \leq i \leq k,\ z_i \in C_0\right\}\right|}{k} \tag{1}$$

In this research, the algorithm uses a specific approach where the number of nearest neighbors is limited to 2, i.e. the 2 nearest neighbors from BD (z_1 and z_2) are found and the probability of x being correctly classified as class 0 or class 1 is calculated as:

$$P\left(t\left(x\right) = 1 | z_1, z_2 \in BD\right) = \frac{1}{2} + \frac{t\left(z_1\right)\gamma\left(x, z_1\right) + t\left(z_2\right)\gamma\left(x, z_2\right)}{\frac{1}{2} + 2t\left(z_1\right)t\left(z_2\right)\gamma\left(z_1, z_2\right)}, \tag{2}$$

```
Input:
    $D$ - The original imbalanced labeled dataset
    $C_0$- Minority class in $D$
    $C_1$ - Majority class in $D$
    $n_0$ - Desired number of points from $C_1$ to include in $BD$
    $k$ - Number of nearest neighbors (k=2 in this case)
Output:
    $BD$ - Balanced dataset
  1. Initialize $BD$ with all points from $C_0$:
         $BD \leftarrow C_0$
  2. While the size of $BD$ does not reach the desired balance:
      a. Initialize $best_{accuracy} \leftarrow -\infty$
      b. Initialize $best_{point} \leftarrow None$
      c. For each point $y \in C_1$:
          i. Create a temporary dataset $TBD$: $TBD \leftarrow BD \cup \{y\}$
         ii. Compute $Acc(D|TBD)$:
              - For each point $x \in D$:
                  1. Find the k nearest neighbors of $x \in TBD$ $(z_1,\ z_2)$
                  2. Compute $P(t\,(x) = 1|z_1,\ z_2 \in TBD)$ using the formula (2)
                  3. Compute $P(t\,(x) = -1|z_1,\ z_2 \in TBD)$ using the formula (2)
              - Calculate the accuracy $Acc(D|TBD)$ using formula (8)
        iii. If $Acc\,(D|TBD) > best_{accuracy}$:
              - Update $best_{accuracy} \leftarrow Acc(D \mid TBD)$
              - Update $best_{point} \leftarrow y$
      d. Add $best_{point}$ to $BD$:
             $BD \leftarrow BD \cup \{best\_point\}$
      e. Remove $best_{point}$ from $C_1$: $C_1 \leftarrow C_1 \setminus \{best_{point}\}$
  3. Return $BD$ as the balanced dataset.
```

Figure 1: Active Down-Sampling Algorithm

for all $x \in C_1$, where $t\,(z_i) = 1$ if $z_i \in C_1$ and -1 if $z_i \in C_0$. Where,

$$P\,(t\,(x) = -1|z_1, z_2 \in BD) = 1 - P\,(t\,(x) = 1|z_1, z_2 \in BD)\,,$$

for all $x \in C_0$.

and,

$$\gamma\left(x,y\right)=\frac{1}{4}e^{-distance(x,y)}. \tag{3}$$

To explain how equation (2) works, let $z_1, z_2 \in C_1$. As a result, the probability of x belonging to C_1 will be more than $\frac{1}{2}$ as follows:

$$P\left(t\left(x\right)=1|t\left(z_1\right)=1,t\left(z_2\right)=1\in BD\right)=\frac{1}{2}+\frac{1\gamma\left(x,z_1\right)+1\gamma\left(x,z_2\right)}{\frac{1}{2}+211\gamma\left(z_1,z_2\right)}>\frac{1}{2}, \tag{4}$$

Assuming that the points z_1 and z_2 are close to the x, their distance from x converge to 0. As a result, we observe that: $\gamma \rightarrow \frac{1}{4}e^0 = \frac{1}{4}$ and equation (4) will be:

$$P\left(t\left(x\right)=1|t\left(z_1\right)=1,t\left(z_2\right)=1,\in BD\right)=\frac{1}{2}+\frac{\frac{1}{4}+\frac{1}{4}}{\frac{1}{2}+2\frac{1}{4}}=1, \tag{5}$$

On the other hand, if $z_1, z_2 \in C_0$ and close to x, we can observe that

$$P\left(t\left(x\right)=1|t\left(z_1\right)=-1,t\left(z_2\right)=-1\in BD\right)\rightarrow 0, \tag{6}$$

And if z_1, z_2 belong to different classes, then the point with the smallest distance will influence the classification towards its class.

As a result, the expected accuracy of the classifier, denoted as $Acc\left(D|BD\right)$, is calculated by summing the probabilities that each point $x \in D$, is correctly classified based on its two nearest neighbors within the balanced dataset BD. The formula is as follows:

$$\begin{aligned}Acc\left(D|BD\right)&=\sum\nolimits_{x\in D}P\left(t\left(x\right)|z_1,z_2\in BD\right)=\sum\nolimits_{x\in C_0}P\left(t\left(x\right)=-1|z_1,z_2\in BD\right)\\&\quad+\sum\nolimits_{x\in C_1}P\left(t\left(x\right)=1|z_1,z_2\in BD\right)\end{aligned} \tag{7}$$

Since the points $x \in D$ belonging to 2 classes we the equation (7) will be:

$$Acc\left(D|BD\right)=\sum\nolimits_{x\in C_0}P\left(t\left(x\right)=-1|z_1,z_2\in BD\right)+\sum\nolimits_{x\in C_1}P\left(t\left(x\right)=1|z_1,z_2\in BD\right) \tag{8}$$

The proposed approach uses two nearest neighbors to balance the trade-off between computational efficiency and classification accuracy. However, the method is general and can be used within any number of nearest neighbors. In this method, when we focus on only the two closest neighbors, we reduce the influence of distant and potentially irrelevant points, making it more effective in handling imbalanced datasets. This also simplifies the selection process by ensuring that the most informative points contribute to the classifier's decision-making process, leading to improved performance across all classes.

The selection process for the most informative representative points from the majority class is as follows:

- For a given point y in C_1, the algorithm creates a temporary set (i.e., TBD) that includes the basic balanced data BD and the examined point y from C_1 (i.e., $TBD \leftarrow BD \cup \{y\}$).
- Based on D', the algorithm computes $Acc(D|TBD)$ and compares it to the accuracy of BD.

- The point y that yields the highest accuracy will be added to BD.
- This procedure repeats until the BD data is balanced.

Furthermore, our proposed method is designed to effectively handle multiclass datasets. During each iteration, the algorithm takes into consideration only the two nearest neighbors, allowing it to focus on comparing between two classes at a time. This approach ensures efficient computations and facilitates the processing of multiclass data. By incorporating the equations mentioned earlier, our algorithm can be seamlessly applied to multiclass datasets. Consequently, we can successfully run the algorithm on such complex datasets, benefiting from its ability to handle class imbalances and enhance classification performance across multiple classes simultaneously.

In summary, our proposed selective sampling algorithm offers an effective approach to address imbalanced multiclass datasets. By actively selecting informative points from the majority class and considering the trade-off between computation time and accuracy, the algorithm significantly enhances the balance of the dataset and improves the overall performance of classifiers.

4. EXPERIMENTS AND DISCUSSION

In this comparative study, we assessed the performance of four distinct algorithms – Down Sampling, SMOTE, Over Sampling, and ADS method – across three datasets: Blood, Yeast, and Ecoli from the UCI repository [31]. The primary goal of these experiments was to assess and compare the performance of various resampling algorithms—Down Sampling, SMOTE, Over Sampling, and the ADS method—applied in conjunction with the K-Nearest Neighbors (KNN) classifier. We aimed to evaluate the efficacy of these methods across three distinct datasets: Blood, Yeast, and Ecoli. These algorithms were evaluated using a K-Nearest Neighbors (KNN) classifier, with the number of neighbors varied as an experimental parameter ('k'). The 'k' parameter for the KNN classifier was adjusted across a range of values: 1, 3, 5, 7, 9, and 11 to analyze the sensitivity of the resampling algorithms to the choice of 'k'. Performance metrics such as Accuracy, F1 Score, Recall, Precision, ROC-AUC, and class-specific performance measures were analyzed.

To mitigate variability and ensure the robustness of our findings, each experimental condition was executed ten times for each value of 'k'. This iterative approach allowed us to account for the inherent randomness in the resampling algorithms and the KNN classifier's sensitivity to the initialization of the dataset. After ten iterations, we computed the average for each performance metric (Accuracy, F1 Score, Recall, Precision, and ROC-AUC) to obtain a reliable estimate of the classifiers' performance. The results presented are therefore the mean values of these ten independent runs, providing a solid statistical basis for the comparative analysis and enhancing the reliability of our conclusions.

The ”BLOOD” dataset contains 748 donor records, with 570 (76%) records from class 1 (donation) and the remaining from class 2 (no-donation). Similarly, the Yeast dataset contains 1484 records with 10 classes, and it suffers from class imbalance with varying numbers of samples in each class (as shown in TABLE 1). In our experiments, the positive samples belong to class ME2, while the negative samples belong to the rest, resulting in an imbalanced class distribution, with a ratio of 51 instances of the ME2 class to 1433 instances of the other classes. The Ecoli dataset contains 336 records with 8 classes (as shown in TABLE 2). Each record in the dataset represents a bacterial strain and contains various features such as the localization of the protein and the type of protein. In our experiments the positive samples belong to class imU and the negative samples belong to the rest. As a result, the class distribution for the Ecoli data is imbalanced, with a ratio of 35 instances of the imU class to 301 instances of the other classes.

Table 1: the distribution of the classes.

Class name	Description	# Samples
CYT	Cytosolc of Cytoskeletal	463
NUC	Nuclear	429
MIT	Mitochondrial	244
ME3	Membrane protein, no N-terminal signal	163
ME2	Membrane protein, uncleaved signal	51
ME1	Membrane protein, cleaved signal	44
EXC	Extracellular	37
VAC	Vacuolar	30
POX	Peroxisomal	20
ERL	Endoplasmic reticulum lumen	5

Table 2: the distribution of the classes.

Class name	Description	# Samples
cp	Cytoplasm	143
im	Inner membrane without signal sequence	77
pp	Periplasm	52
imU	Inner membrane, non-cleavable signal sequence	35
om	Outer membrane	20
omL	Outer membrane lipoprotein	5
imL	Inner membrane lipoprotein	2
imS	Inner membrane, cleavable signal sequence	2

TABLE 3 depicts the results of the experiments. Regarding the Blood Dataset, across varying ’k’, Over Sampling often yielded the highest Accuracy and F1 scores, reaching a peak Accuracy of 69.63% (k=7) and F1 Score of 71.61% (k=7). However, ADS method exhibited a significant improvement in ROC-AUC, particularly at k=9, where it reached 69.73%, suggesting a better trade-off between sensitivity and specificity despite lower accuracy in some instances.

On the Yeast Dataset, ADS method demonstrated superior performance consistently across all ’k’ values, with the highest Accuracy of 84.36% and an F1 Score of 84.04% at k=9, emphasizing its robustness in handling multiclass classification problems. Interestingly, ADS method also maintained a high level of ROC-AUC performance, peaking at 87.75% (k=9), indicating its ability to discriminate between multiple classes effectively.

On the Ecoli dataset, the ADS method outperformed other algorithms across nearly all metrics, particularly excelling in Precision and ROC-AUC. It achieved the highest accuracy of 93.73% (k=3)

Table 3: Results summary.

Dataset name	Algorithm name	k	Accuracy	F1_score	Recall	Precision	RocAuc	true positive	true negative	flase positive	false negative	Precision positive	Recall positive	F1 positive	Precision negative	Recall negative	F1 negative	F1 score
blood	Down Sampling	1	0.54	0.57	0.51	0.82	0.58	57.00	23.00	13.00	55.00	0.81	0.51	0.63	0.30	0.64	0.40	0.57
	SMOTE	1	0.66	0.67	0.72	0.80	0.59	82.00	16.00	20.00	31.00	0.80	0.73	0.76	0.34	0.44	0.39	0.67
	Over Sampling	1	0.68	0.68	0.81	0.78	0.56	91.00	11.00	25.00	21.00	0.78	0.81	0.80	0.34	0.31	0.32	0.68
	ADS	1	0.49	0.51	0.40	0.86	0.60	44.00	29.00	7.00	68.00	0.86	0.39	0.54	0.30	0.81	0.44	0.52
	Down Sampling	3	0.63	0.65	0.62	0.84	0.64	70.00	23.00	13.00	43.00	0.84	0.62	0.71	0.35	0.64	0.45	0.65
	SMOTE	3	0.68	0.69	0.74	0.82	0.64	84.00	17.00	18.00	29.00	0.82	0.74	0.78	0.37	0.49	0.42	0.70
	Over Sampling	3	0.69	0.70	0.78	0.81	0.62	88.00	15.00	20.00	25.00	0.82	0.78	0.80	0.38	0.43	0.40	0.70
	ADS	3	0.57	0.59	0.53	0.84	0.63	60.00	24.00	11.00	53.00	0.85	0.53	0.65	0.31	0.69	0.43	0.60
	Down Sampling	5	0.65	0.67	0.63	0.86	0.69	71.00	26.00	11.00	41.00	0.87	0.63	0.73	0.39	0.70	0.50	0.67
	SMOTE	5	0.69	0.70	0.74	0.83	0.68	83.00	20.00	17.00	29.00	0.83	0.74	0.78	0.41	0.54	0.47	0.70
	Over Sampling	5	0.68	0.69	0.73	0.82	0.66	82.00	19.00	18.00	30.00	0.82	0.73	0.77	0.39	0.51	0.44	0.69
	ADS	5	0.60	0.62	0.56	0.85	0.67	63.00	25.00	11.00	48.00	0.85	0.57	0.68	0.34	0.69	0.46	0.63
	Down Sampling	7	0.65	0.68	0.63	0.89	0.71	74.00	22.00	9.00	43.00	0.89	0.63	0.74	0.34	0.71	0.46	0.68
	SMOTE	7	0.69	0.71	0.73	0.86	0.69	86.00	17.00	14.00	31.00	0.86	0.74	0.79	0.35	0.55	0.43	0.72
	Over Sampling	7	0.70	0.72	0.74	0.86	0.68	86.00	17.00	14.00	31.00	0.86	0.74	0.79	0.35	0.55	0.43	0.72
	ADS	7	0.53	0.56	0.50	0.85	0.64	58.00	21.00	11.00	59.00	0.84	0.50	0.62	0.26	0.66	0.38	0.57
	Down Sampling	9	0.66	0.69	0.64	0.90	0.75	73.00	25.00	8.00	42.00	0.90	0.64	0.75	0.37	0.76	0.50	0.69
	SMOTE	9	0.68	0.71	0.70	0.86	0.70	81.00	21.00	12.00	34.00	0.87	0.70	0.78	0.38	0.64	0.48	0.71
	Over Sampling	9	0.68	0.70	0.70	0.86	0.69	81.00	20.00	13.00	34.00	0.86	0.70	0.78	0.37	0.61	0.46	0.71
	ADS	9	0.60	0.63	0.57	0.88	0.70	65.00	24.00	9.00	50.00	0.88	0.57	0.69	0.32	0.73	0.45	0.63
	Down Sampling	11	0.66	0.68	0.64	0.87	0.72	72.00	25.00	11.00	40.00	0.87	0.64	0.74	0.39	0.69	0.50	0.68
	SMOTE	11	0.69	0.71	0.71	0.85	0.70	80.00	23.00	13.00	33.00	0.86	0.71	0.78	0.41	0.64	0.50	0.71
	Over Sampling	11	0.67	0.69	0.69	0.85	0.68	78.00	22.00	14.00	34.00	0.85	0.70	0.77	0.39	0.61	0.48	0.70
	ADS	11	0.67	0.69	0.69	0.84	0.69	77.00	22.00	14.00	35.00	0.85	0.69	0.76	0.39	0.61	0.47	0.69
yeast	Down Sampling	1	0.75	0.75	0.77	0.83	0.73	72.00	33.00	14.00	21.00	0.84	0.77	0.80	0.61	0.70	0.65	0.75
	SMOTE	1	0.77	0.76	0.84	0.81	0.73	79.00	29.00	18.00	14.00	0.81	0.85	0.83	0.67	0.62	0.64	0.77
	Over Sampling	1	0.78	0.77	0.88	0.80	0.73	83.00	27.00	20.00	10.00	0.81	0.89	0.85	0.73	0.57	0.64	0.78
	ADS	1	0.78	0.77	0.87	0.81	0.73	81.00	28.00	19.00	12.00	0.81	0.87	0.84	0.70	0.60	0.64	0.77
	Down Sampling	3	0.76	0.77	0.79	0.85	0.80	74.00	34.00	13.00	19.00	0.85	0.80	0.82	0.64	0.72	0.68	0.77
	SMOTE	3	0.76	0.77	0.79	0.84	0.79	74.00	33.00	14.00	19.00	0.84	0.80	0.82	0.64	0.70	0.67	0.77
	Over Sampling	3	0.76	0.76	0.79	0.84	0.78	74.00	33.00	14.00	19.00	0.84	0.80	0.82	0.64	0.70	0.67	0.77
	ADS	3	0.79	0.79	0.88	0.82	0.80	82.00	30.00	18.00	11.00	0.82	0.88	0.85	0.73	0.63	0.67	0.79
	Down Sampling	5	0.79	0.79	0.83	0.85	0.84	77.00	34.00	13.00	16.00	0.86	0.83	0.84	0.68	0.72	0.70	0.79
	SMOTE	5	0.78	0.78	0.79	0.86	0.83	74.00	36.00	11.00	19.00	0.87	0.80	0.83	0.66	0.77	0.71	0.79
	Over Sampling	5	0.77	0.78	0.79	0.85	0.82	74.00	35.00	12.00	19.00	0.86	0.80	0.83	0.65	0.75	0.69	0.78
	ADS	5	0.82	0.81	0.91	0.83	0.83	85.00	30.00	17.00	8.00	0.83	0.91	0.87	0.79	0.64	0.71	0.82
	Down Sampling	7	0.77	0.78	0.81	0.85	0.83	76.00	33.00	13.00	18.00	0.85	0.81	0.83	0.65	0.72	0.68	0.78
	SMOTE	7	0.76	0.76	0.78	0.84	0.82	74.00	33.00	13.00	20.00	0.85	0.79	0.82	0.62	0.72	0.67	0.77
	Over Sampling	7	0.76	0.77	0.79	0.85	0.81	74.00	33.00	13.00	20.00	0.85	0.79	0.82	0.62	0.72	0.67	0.77
	ADS	7	0.80	0.80	0.88	0.83	0.83	83.00	30.00	17.00	11.00	0.83	0.88	0.86	0.73	0.64	0.68	0.80
	Down Sampling	9	0.82	0.82	0.83	0.89	0.88	80.00	35.00	10.00	15.00	0.89	0.84	0.87	0.70	0.78	0.74	0.82
	SMOTE	9	0.80	0.81	0.81	0.89	0.87	77.00	36.00	9.00	18.00	0.90	0.81	0.85	0.67	0.80	0.73	0.81
	Over Sampling	9	0.80	0.80	0.81	0.88	0.87	77.00	35.00	10.00	18.00	0.89	0.81	0.85	0.66	0.78	0.71	0.80
	ADS	9	0.84	0.84	0.92	0.86	0.88	88.00	31.00	14.00	8.00	0.86	0.92	0.89	0.80	0.69	0.74	0.84
	Down Sampling	11	0.79	0.79	0.82	0.85	0.85	76.00	36.00	12.00	16.00	0.86	0.83	0.84	0.69	0.75	0.72	0.80
	SMOTE	11	0.78	0.78	0.80	0.85	0.84	73.00	36.00	12.00	18.00	0.86	0.80	0.83	0.67	0.75	0.71	0.79
	Over Sampling	11	0.78	0.78	0.81	0.85	0.84	75.00	36.00	13.00	17.00	0.85	0.82	0.83	0.68	0.74	0.71	0.79
	ADS	11	0.82	0.82	0.91	0.83	0.85	84.00	32.00	17.00	8.00	0.83	0.91	0.87	0.80	0.65	0.72	0.82
ecoli	Down Sampling	1	0.84	0.87	0.83	0.99	0.92	50.00	6.00	0.00	10.00	1.00	0.83	0.91	0.38	1.00	0.55	0.88
	SMOTE	1	0.88	0.90	0.88	0.98	0.91	53.00	5.00	0.00	7.00	1.00	0.88	0.94	0.42	1.00	0.59	0.91
	Over Sampling	1	0.85	0.88	0.86	0.98	0.89	51.00	5.00	1.00	8.00	0.98	0.86	0.92	0.39	0.83	0.53	0.88
	ADS	1	0.92	0.92	0.95	0.96	0.91	57.00	4.00	2.00	3.00	0.97	0.95	0.96	0.57	0.67	0.62	0.93
	Down Sampling	3	0.84	0.87	0.83	0.99	0.95	49.00	7.00	0.00	10.00	1.00	0.83	0.91	0.41	1.00	0.58	0.87
	SMOTE	3	0.88	0.89	0.87	0.99	0.92	51.00	6.00	0.00	7.00	1.00	0.88	0.94	0.46	1.00	0.63	0.91
	Over Sampling	3	0.86	0.88	0.86	0.99	0.91	51.00	6.00	0.00	8.00	1.00	0.86	0.93	0.43	1.00	0.60	0.90
	ADS	3	0.94	0.94	0.97	0.96	0.95	57.00	5.00	2.00	1.00	0.97	0.98	0.97	0.83	0.71	0.77	0.95
	Down Sampling	5	0.82	0.85	0.80	1.00	0.94	47.00	7.00	0.00	12.00	1.00	0.80	0.89	0.37	1.00	0.54	0.85
	SMOTE	5	0.87	0.89	0.86	0.99	0.92	51.00	6.00	0.00	8.00	1.00	0.86	0.93	0.43	1.00	0.60	0.90
	Over Sampling	5	0.85	0.87	0.84	0.99	0.91	50.00	6.00	0.00	9.00	1.00	0.85	0.92	0.40	1.00	0.57	0.89
	ADS	5	0.92	0.92	0.96	0.96	0.94	57.00	4.00	2.00	2.00	0.97	0.97	0.97	0.67	0.67	0.67	0.94
	Down Sampling	7	0.83	0.86	0.82	0.99	0.94	49.00	5.00	0.00	11.00	1.00	0.82	0.90	0.31	1.00	0.48	0.87
	SMOTE	7	0.86	0.88	0.85	0.99	0.88	52.00	5.00	0.00	8.00	1.00	0.87	0.93	0.39	1.00	0.56	0.90
	Over Sampling	7	0.84	0.87	0.84	0.98	0.87	51.00	5.00	0.00	9.00	1.00	0.85	0.92	0.36	1.00	0.53	0.89
	ADS	7	0.92	0.93	0.95	0.96	0.93	58.00	3.00	2.00	2.00	0.97	0.97	0.97	0.60	0.60	0.60	0.94
	Down Sampling	9	0.81	0.84	0.78	1.00	0.94	46.00	7.00	0.00	12.00	1.00	0.79	0.89	0.37	1.00	0.54	0.85
	SMOTE	9	0.86	0.88	0.85	0.99	0.91	50.00	7.00	0.00	8.00	1.00	0.86	0.93	0.47	1.00	0.64	0.90
	Over Sampling	9	0.85	0.87	0.84	0.98	0.90	49.00	7.00	0.00	9.00	1.00	0.85	0.92	0.44	1.00	0.61	0.88
	ADS	9	0.91	0.91	0.95	0.95	0.94	55.00	5.00	2.00	3.00	0.97	0.95	0.96	0.63	0.71	0.67	0.93
	Down Sampling	11	0.79	0.83	0.77	0.99	0.91	46.00	6.00	0.00	13.00	1.00	0.78	0.88	0.32	1.00	0.48	0.84
	SMOTE	11	0.84	0.87	0.85	0.97	0.89	50.00	6.00	1.00	9.00	0.98	0.85	0.91	0.40	0.86	0.55	0.87
	Over Sampling	11	0.83	0.86	0.83	0.98	0.88	49.00	6.00	1.00	10.00	0.98	0.83	0.90	0.38	0.86	0.52	0.86
	ADS	11	0.92	0.92	0.95	0.96	0.92	56.00	5.00	2.00	3.00	0.97	0.95	0.96	0.63	0.71	0.67	0.93

Table 3: Continued..

Dataset name	Algorithm name	k	Accuracy	F1_score	Recall	Precision	RocAuc	true positive	true negative	flase positive	false negative	Precision positive	Recall positive	F1 positive	Precision negative	Recall negative	F1 negative	F1 score
blood	Down Sampling	1	0.54	0.57	0.51	0.82	0.58	57.00	23.00	13.00	55.00	0.81	0.51	0.63	0.30	0.64	0.40	0.57
	SMOTE	1	0.66	0.67	0.72	0.80	0.59	82.00	16.00	20.00	31.00	0.80	0.73	0.76	0.34	0.44	0.39	0.67
	Over Sampling	1	0.68	0.68	0.81	0.78	0.56	91.00	11.00	25.00	21.00	0.78	0.81	0.80	0.34	0.31	0.32	0.68
	ADS	1	0.49	0.51	0.40	0.86	0.60	44.00	29.00	7.00	68.00	0.86	0.39	0.54	0.30	0.81	0.44	0.52
	Down Sampling	3	0.63	0.65	0.62	0.84	0.64	70.00	23.00	13.00	43.00	0.84	0.62	0.71	0.35	0.64	0.45	0.65
	SMOTE	3	0.68	0.69	0.74	0.82	0.64	84.00	17.00	18.00	29.00	0.82	0.74	0.78	0.37	0.49	0.42	0.70
	Over Sampling	3	0.69	0.70	0.78	0.81	0.62	88.00	15.00	20.00	25.00	0.82	0.78	0.80	0.38	0.43	0.40	0.70
	ADS	3	0.57	0.59	0.53	0.84	0.63	60.00	24.00	11.00	53.00	0.85	0.53	0.65	0.31	0.69	0.43	0.60
	Down Sampling	5	0.65	0.67	0.63	0.86	0.69	71.00	26.00	11.00	41.00	0.87	0.63	0.73	0.39	0.70	0.50	0.67
	SMOTE	5	0.69	0.70	0.74	0.83	0.68	83.00	20.00	17.00	29.00	0.83	0.74	0.78	0.41	0.54	0.47	0.70
	Over Sampling	5	0.68	0.69	0.73	0.82	0.66	82.00	19.00	18.00	30.00	0.82	0.73	0.77	0.39	0.51	0.44	0.69
	ADS	5	0.60	0.62	0.56	0.85	0.67	63.00	25.00	11.00	48.00	0.85	0.57	0.68	0.34	0.69	0.46	0.63
	Down Sampling	7	0.65	0.68	0.63	0.89	0.71	74.00	22.00	9.00	43.00	0.89	0.63	0.74	0.34	0.71	0.46	0.68
	SMOTE	7	0.69	0.71	0.73	0.86	0.69	86.00	17.00	14.00	31.00	0.86	0.74	0.79	0.35	0.55	0.43	0.72
	Over Sampling	7	0.70	0.72	0.74	0.86	0.68	86.00	17.00	14.00	31.00	0.86	0.74	0.79	0.35	0.55	0.43	0.72
	ADS	7	0.53	0.56	0.50	0.85	0.64	58.00	21.00	11.00	59.00	0.84	0.50	0.62	0.26	0.66	0.38	0.57
	Down Sampling	9	0.66	0.69	0.64	0.90	0.75	73.00	25.00	8.00	42.00	0.90	0.64	0.75	0.37	0.76	0.50	0.69
	SMOTE	9	0.68	0.71	0.70	0.86	0.70	81.00	21.00	12.00	34.00	0.87	0.70	0.78	0.38	0.64	0.48	0.71
	Over Sampling	9	0.68	0.70	0.70	0.86	0.69	81.00	20.00	13.00	34.00	0.86	0.70	0.78	0.37	0.61	0.46	0.71
	ADS	9	0.60	0.63	0.57	0.88	0.70	65.00	24.00	9.00	50.00	0.88	0.57	0.69	0.32	0.73	0.45	0.63
	Down Sampling	11	0.66	0.68	0.64	0.87	0.72	72.00	25.00	11.00	40.00	0.87	0.64	0.74	0.39	0.69	0.50	0.68
	SMOTE	11	0.69	0.71	0.71	0.85	0.70	80.00	23.00	13.00	33.00	0.86	0.71	0.78	0.41	0.64	0.50	0.71
	Over Sampling	11	0.67	0.69	0.69	0.85	0.68	78.00	22.00	14.00	34.00	0.85	0.70	0.77	0.39	0.61	0.48	0.70
	ADS	11	0.67	0.69	0.69	0.84	0.69	77.00	22.00	14.00	35.00	0.85	0.69	0.76	0.39	0.61	0.47	0.69
yeast	Down Sampling	1	0.75	0.75	0.77	0.83	0.73	72.00	33.00	14.00	21.00	0.84	0.77	0.80	0.61	0.70	0.65	0.75
	SMOTE	1	0.77	0.76	0.84	0.81	0.73	79.00	29.00	18.00	14.00	0.81	0.85	0.83	0.67	0.62	0.64	0.77
	Over Sampling	1	0.78	0.77	0.88	0.80	0.73	83.00	27.00	20.00	10.00	0.81	0.89	0.85	0.73	0.57	0.64	0.78
	ADS	1	0.78	0.77	0.87	0.81	0.73	81.00	28.00	19.00	12.00	0.81	0.87	0.84	0.70	0.60	0.64	0.77
	Down Sampling	3	0.76	0.77	0.79	0.85	0.80	74.00	34.00	13.00	19.00	0.85	0.80	0.82	0.64	0.72	0.68	0.77
	SMOTE	3	0.76	0.77	0.79	0.84	0.79	74.00	33.00	14.00	19.00	0.84	0.80	0.82	0.64	0.70	0.67	0.77
	Over Sampling	3	0.76	0.76	0.79	0.84	0.78	74.00	33.00	14.00	19.00	0.84	0.80	0.82	0.64	0.70	0.67	0.77
	ADS	3	0.79	0.79	0.88	0.82	0.80	82.00	30.00	18.00	11.00	0.82	0.88	0.85	0.73	0.63	0.67	0.79
	Down Sampling	5	0.79	0.79	0.83	0.85	0.84	77.00	34.00	13.00	16.00	0.86	0.83	0.84	0.68	0.72	0.70	0.79
	SMOTE	5	0.78	0.78	0.79	0.86	0.83	74.00	36.00	11.00	19.00	0.87	0.80	0.83	0.66	0.77	0.71	0.79
	Over Sampling	5	0.77	0.78	0.79	0.85	0.82	74.00	35.00	12.00	19.00	0.86	0.80	0.83	0.65	0.75	0.69	0.78
	ADS	5	0.82	0.81	0.91	0.83	0.83	85.00	30.00	17.00	8.00	0.83	0.91	0.87	0.79	0.64	0.71	0.82
	Down Sampling	7	0.77	0.78	0.81	0.85	0.83	76.00	33.00	13.00	18.00	0.85	0.81	0.83	0.65	0.72	0.68	0.78
	SMOTE	7	0.76	0.76	0.78	0.84	0.82	74.00	33.00	13.00	20.00	0.85	0.79	0.82	0.62	0.72	0.67	0.77
	Over Sampling	7	0.76	0.77	0.79	0.85	0.81	74.00	33.00	13.00	20.00	0.85	0.79	0.82	0.62	0.72	0.67	0.77
	ADS	7	0.80	0.80	0.88	0.83	0.83	83.00	30.00	17.00	11.00	0.83	0.88	0.86	0.73	0.64	0.68	0.80
	Down Sampling	9	0.82	0.82	0.83	0.89	0.88	80.00	35.00	10.00	15.00	0.89	0.84	0.87	0.70	0.78	0.74	0.82
	SMOTE	9	0.80	0.81	0.81	0.89	0.87	77.00	36.00	9.00	18.00	0.90	0.81	0.85	0.67	0.80	0.73	0.81
	Over Sampling	9	0.80	0.80	0.81	0.88	0.87	77.00	35.00	10.00	18.00	0.89	0.81	0.85	0.66	0.78	0.71	0.80
	ADS	9	0.84	0.84	0.92	0.86	0.88	88.00	31.00	14.00	8.00	0.86	0.92	0.89	0.80	0.69	0.74	0.84
	Down Sampling	11	0.79	0.79	0.82	0.85	0.85	76.00	36.00	12.00	16.00	0.86	0.83	0.84	0.69	0.75	0.72	0.80
	SMOTE	11	0.78	0.78	0.80	0.85	0.84	73.00	36.00	12.00	18.00	0.86	0.80	0.83	0.67	0.75	0.71	0.79
	Over Sampling	11	0.78	0.78	0.81	0.85	0.84	75.00	36.00	13.00	17.00	0.85	0.82	0.83	0.68	0.74	0.71	0.79
	ADS	11	0.82	0.82	0.91	0.83	0.85	84.00	32.00	17.00	8.00	0.83	0.91	0.87	0.80	0.65	0.72	0.82
ecoll	Down Sampling	1	0.84	0.87	0.83	0.99	0.92	50.00	6.00	0.00	10.00	1.00	0.83	0.91	0.38	1.00	0.55	0.88
	SMOTE	1	0.88	0.90	0.88	0.98	0.91	53.00	5.00	0.00	7.00	1.00	0.88	0.94	0.42	1.00	0.59	0.91
	Over Sampling	1	0.85	0.88	0.86	0.98	0.89	51.00	5.00	1.00	8.00	0.98	0.86	0.92	0.39	0.83	0.53	0.88
	ADS	1	0.92	0.92	0.95	0.96	0.91	57.00	4.00	2.00	3.00	0.97	0.95	0.96	0.57	0.67	0.62	0.93
	Down Sampling	3	0.84	0.87	0.83	0.99	0.95	49.00	7.00	0.00	10.00	1.00	0.83	0.91	0.41	1.00	0.58	0.87
	SMOTE	3	0.88	0.89	0.87	0.99	0.92	51.00	6.00	0.00	7.00	1.00	0.88	0.94	0.46	1.00	0.63	0.91
	Over Sampling	3	0.86	0.88	0.86	0.99	0.91	51.00	6.00	0.00	8.00	1.00	0.86	0.93	0.43	1.00	0.60	0.90
	ADS	3	0.94	0.94	0.97	0.96	0.95	57.00	5.00	2.00	1.00	0.97	0.98	0.97	0.83	0.71	0.77	0.95
	Down Sampling	5	0.82	0.85	0.80	1.00	0.94	47.00	7.00	0.00	12.00	1.00	0.80	0.89	0.37	1.00	0.54	0.85
	SMOTE	5	0.87	0.89	0.86	0.99	0.92	51.00	6.00	0.00	8.00	1.00	0.86	0.93	0.43	1.00	0.60	0.90
	Over Sampling	5	0.85	0.87	0.84	0.99	0.91	50.00	6.00	0.00	9.00	1.00	0.85	0.92	0.40	1.00	0.57	0.89
	ADS	5	0.92	0.92	0.96	0.96	0.94	57.00	4.00	2.00	2.00	0.97	0.97	0.97	0.67	0.67	0.67	0.94
	Down Sampling	7	0.83	0.86	0.82	0.99	0.94	49.00	5.00	0.00	11.00	1.00	0.82	0.90	0.31	1.00	0.48	0.87
	SMOTE	7	0.86	0.88	0.85	0.99	0.88	52.00	5.00	0.00	8.00	1.00	0.87	0.93	0.39	1.00	0.56	0.90
	Over Sampling	7	0.84	0.87	0.84	0.98	0.87	51.00	5.00	0.00	9.00	1.00	0.85	0.92	0.36	1.00	0.53	0.89
	ADS	7	0.92	0.93	0.95	0.96	0.93	58.00	3.00	2.00	2.00	0.97	0.97	0.97	0.60	0.60	0.60	0.94
	Down Sampling	9	0.81	0.84	0.78	1.00	0.94	46.00	7.00	0.00	12.00	1.00	0.79	0.89	0.37	1.00	0.54	0.85
	SMOTE	9	0.86	0.88	0.85	0.99	0.91	50.00	7.00	0.00	8.00	1.00	0.86	0.93	0.47	1.00	0.64	0.90
	Over Sampling	9	0.85	0.87	0.84	0.98	0.90	49.00	7.00	0.00	9.00	1.00	0.85	0.92	0.44	1.00	0.61	0.88
	ADS	9	0.91	0.91	0.95	0.95	0.94	55.00	5.00	2.00	3.00	0.97	0.95	0.96	0.63	0.71	0.67	0.93
	Down Sampling	11	0.79	0.83	0.77	0.99	0.91	46.00	6.00	0.00	13.00	1.00	0.78	0.88	0.32	1.00	0.48	0.84
	SMOTE	11	0.84	0.87	0.85	0.97	0.89	50.00	6.00	1.00	9.00	0.98	0.85	0.91	0.40	0.86	0.55	0.87
	Over Sampling	11	0.83	0.86	0.83	0.98	0.88	49.00	6.00	1.00	10.00	0.98	0.83	0.90	0.38	0.86	0.52	0.86
	ADS	11	0.92	0.92	0.95	0.96	0.92	56.00	5.00	2.00	3.00	0.97	0.95	0.96	0.63	0.71	0.67	0.93

and an F1 Score of 93.64% (k=3). The ADS method's precision was notably perfect (1.0) in Down Sampling and SMOTE at specific 'k' values. Additionally, its superior balance in Recall suggests an enhanced ability to minimize false negatives. However, the performance of the ADS method was notably less impressive on the blood dataset, potentially due to the loss of critical minority class information following down sampling.

For the SMOTE method it performed better than Down Sampling and Over Sampling, especially in the yeast and ecoli datasets, indicating its effectiveness in synthesizing new minority class instances. However, it still fell short of ADS method, potentially due to the latter's superior handling of class imbalances. SMOTE is particularly advantageous when the minority class is significantly underrepresented, but it may introduce synthetic samples that do not add substantial new information, leading to overfitting, particularly at lower 'k' values.

Over Sampling method revealed competitive performance, particularly in lower 'k' values in the blood dataset, suggesting that simple replication of minority class instances may be initially beneficial. Nevertheless, as the complexity of the model increased with higher 'k' values, its performance plateaued or declined. Over Sampling's simplicity is both its strength and its weakness: while it can quickly address class imbalance, it may not provide the nuanced improvements seen with more sophisticated methods like ADS, especially in more complex or high-dimensional datasets. Furthermore, the primary disadvantage is the potential for overfitting due to the replication of minority class instances. As the model complexity increases, or as 'k' increases, the effectiveness of Over Sampling tends to plateau or decline.

Down Sampling showed reasonable performance due to its simplicity and lower computational cost, making it a viable option when computational resources are limited. However, its major drawback is the potential loss of important information from the majority class, which can lead to poorer performance, as observed in the Blood dataset. Down Sampling can be particularly harmful in scenarios where the majority class contains critical information that is necessary for accurate classification. Furthermore, the method may lead to a significant loss of information from the majority class, as seen in the Blood dataset, which can result in suboptimal performance. This method is less effective when the dataset contains critical minority information that needs to be preserved.

The consistently high performance of the ADS method suggests that it likely employs a more nuanced approach to balancing class distributions, possibly incorporating mechanisms that effectively address both minority and majority classes. This is further supported by its high Precision and Recall, indicating that the ADS method successfully avoids the common trade-off between these metrics seen in many resampling techniques. Notably, the ADS method continued to maintain or even improve its performance as the value of 'k' increased, which contrasts with traditional methods that often experience diminishing returns with higher 'k' values. This implies that ADS method may benefit from considering more neighbors, possibly due to a more refined approach to weighting or selecting neighbors based on the dataset's characteristics.

Despite its effectiveness, the ADS method may introduce computational complexity, which could be a drawback when applied to large-scale datasets. Its performance was also less impressive on the Blood dataset, suggesting that it may struggle with datasets where critical minority information is prone to loss during resampling.

In conclusion, the experimental results indicate that the ADS method offers a compelling alternative to conventional resampling methods in KNN classification tasks across various datasets. By consistently achieving high Accuracy, F1 Score, and ROC-AUC, ADS method demonstrates its ability to deliver balanced performance, effectively addressing class imbalance challenges. Future research should explore the scalability of the ADS method, its adaptability to different classifier paradigms, and its application to a broader range of datasets to further validate its effectiveness in diverse machine learning scenarios.

5. CONCLUSIONS

In this study, we introduced the Active Down-sampling (ADS) method, a novel approach designed to tackle the problem of imbalanced data in machine learning. ADS utilizes a down-sampling technique coupled with an active learning approach to intelligently select informative points from the majority class for training. Through extensive experimentation on three publicly available datasets (”BLOOD,” ”Yeast,” and ”Ecoli”), we demonstrated the superior performance of our ADS algorithm compared to state-of-the-art methods in terms of classification accuracy.

The positive impact (pos) of ADS was evident in various aspects. Firstly, ADS effectively addressed the class imbalance issue, leading to more accurate predictions for both minority and majority classes. By selecting informative majority class samples, ADS enhanced the dataset representation, resulting in improved recognition of minority class patterns and overall classifier performance.

Moreover, ADS positively impacted resource optimization. The active learning approach in ADS allowed for the selection of a smaller yet representative subset of majority class samples, reducing the need for extensive training data and optimizing computational costs without compromising classifier accuracy. This cost-effective attribute makes ADS a compelling choice for real-world applications with limited resources.

Furthermore, the cost impact of ADS proved significant in practical applications. By mitigating class imbalance, ADS minimized the costs associated with false negatives, where critical positive instances are incorrectly classified as negatives. This aspect is particularly crucial in domains such as medical diagnosis and fraud detection, where misclassification can have severe consequences.

The experiments also showcased the adaptability and versatility of ADS. By successfully handling imbalanced data across diverse datasets, ADS demonstrated its potential as a robust solution for various real-world scenarios. Its compatibility with different classifiers, such as Naive Bayes, Support Vector Machine, and Decision Trees, further highlights its adaptability and efficacy in enhancing classifier performance.

As a future research direction, we propose to expand the evaluation of ADS on a wider range of datasets to reinforce its effectiveness and generalize its applicability. Additionally, integrating ADS with other classifiers will enable us to explore optimal combinations that leverage its strengths in diverse contexts. Moreover, investigating the impact of selecting samples from both the minority and majority classes could potentially lead to further improvements in the ADS method.

In conclusion, the ADS method provides a promising and effective solution to the challenges posed by imbalanced data in machine learning. Its outstanding performance, positive impact on resource optimization, and cost-saving benefits position it as a valuable tool for real-world applications. As we continue to refine and validate ADS through further research and experimentation, we look forward to unlocking its full potential and advancing the field of machine learning in dealing with imbalanced data scenarios.

References

[1] Lee W, Seo K. Early Failure Detection of Paper Manufacturing Machinery Using Nearest Neighbor-Based Feature Extraction. Eng Rep. 2020.

[2] Liu S, Wang Y, Zhang J, Chen C, Xiang Y. Addressing the Class Imbalance Problem in Twitter Spam Detection Using Ensemble Learning. Comput Secur. 2017;69:35-49.

[3] Dhankhad S, Mohammed E, Far B. Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study. IEEE International Conference on Information Reuse and Integration (IRI). IEEE PUBLICATIONS. 2018:122-125.

[4] Mena LJ, Gonzalez JA. Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic. Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference (Flairs Conference). 2006:574-579.

[5] Okutan A, Yang SJ, McConky K. Forecasting Cyber Attacks With Imbalanced Data Sets and Different Time Granularities. 2018. Arxiv preprint: https://arxiv.org/pdf/1803.09560

[6] Susan S, Kumar A. The Balancing Trick: Optimized Sampling of Imbalanced Datasets—a Brief Survey of the Recent State of the Art. Eng Rep. 2021;3: e12298.

[7] Tyagi S, Mittal S. Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning. In: Proceedings of the ICRIC 2019. Springer; 2020:209-221.

[8] Tahir MA, Kittler J, Mikolajczyk K, Yan F. A Multiple Expert Approach to the Class Imbalance Problem Using Inverse Random Under Sampling. In Multiple Classifier Systems: 8th International Workshop, MCS 2009. Proceedings. Springer Berlin Heidelberg. 2009:82-91.

[9] Ariannezhad A, Karimpour A, Qin X, Wu YJ, Salmani Y. Handling Imbalanced Data for Real-Time Crash Prediction: Application of Boosting and Sampling Techniques. J Transp Eng A Syst. 2021;147:04020165.

[10] Fiorentini N, Losa M. Handling Imbalanced Data in Road Crash Severity Prediction by Machine Learning Algorithms. Infrastructures. 2020;5:61.

[11] Anand A, Pugalenthi G, Fogel GB, Suganthan PN. An Approach for Classification of Highly Imbalanced Data Using Weighting and Undersampling. Amino Acids. 2010;39:1385-1391.

[12] Sun Z, Song Q, Zhu X, Sun H, Xu B, et al. A Novel Ensemble Method for Classifying Imbalanced Data. Pattern Recognit. 2015;48:1623-1637.

[13] Li X, Shao Q, Wang J. Classification of Tongue Coating Using Gabor and Tamura Features on Unbalanced Data Set. IEEE International Conference on Bioinformatics and Biomedicine. IEEE PUBLICATIONS. 2013:108-109.

[14] Lee W, Seo K. Downsampling for Binary Classification With a Highly Imbalanced Dataset Using Active Learning. Big Data Res. 2022;28:100314.

[15] Kumar NS, Rao KN, Govardhan A, Reddy KS, Mahmood AM. Undersampled K-Means Approach for Handling Imbalanced Distributed Data. Prog Artif Intell. 2014;3:29-38.

[16] Chen Y, Ye H. Improving Active Learning on Imbalanced Datasets by Features Mixing. International Conference on Cloud Computing, Performance Computing, and Deep Learning. 2023;12712.

[17] Karthikeyan S, Kathirvalavakumar T. Modified Leader Algorithm for Under-Sampling the Imbalanced Dataset for Classification. In Intelligent Sustainable Systems: Proceedings of ICISS 2021. Springer. 2021 Aug 27 (pp. 107-118).

[18] Goyal S. Handling Class-Imbalance With Knn (Neighbourhood) Under-Sampling for Software Defect Prediction. Artif Intell Rev. 2022;55:2023-2064.

[19] Cai X. Active Learning for Imbalanced Data: The Difficulty and Proportions of Class Matter. Wirel Commun Mob Comput. 2022;2022:1-9.

[20] Diallo M, Xiong S, Emiru ED, Fesseha A, Abdulsalami AO, et al. A Hybrid Multilayer Perceptron Under-Sampling With Bagging Dealing with a Real-Life Imbalanced Rice Dataset. Information. 2021;12 :291.

[21] Wang X, Liu B, Cao S, Jing L, Yu J. Important Sampling Based Active Learning for Imbalance Classification. Sci China Info Sci. 2020;63:1-4

[22] Zhang H, Zhang H, Pirbhulal S, Wu W, Albuquerque VH. Active Balancing Mechanism for Imbalanced Medical Data in Deep Learning–Based Classification Models. ACM Trans Multimedia Comput Commun Appl. 2020;16:1-5.

[23] Aggarwal U, Popescu A, Hudelot C. Active Learning for Imbalanced Datasets. IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA. 2020;2020:1417-1426.

[24] Cai J, Agrawal S, Idicula S, Varadarajan V, Yakovlev A, et al. Inventors; Oracle International Corp, assignee. Adaptive sampling for imbalance mitigation and dataset size reduction in machine learning. United States patent US11562178B2. 2023.

[25] Khaldoon A, Sujan R, Ali AG, Dharma P A. Enhancing Imbalanced Dataset by Utilizing (K-NN Based SMOTE_3D Algorithm). Ann Robot Automation. 2020;4:1-6.

[26] Maheshwari S, Jain RC, Jadon RS. Rusdataboost-Im: Improving Classification Performance in Imbalanced Data. In: Intelligent Data Engineering and Analytics. Springer. 2021;2:623-635.

[27] Sahni D, Pappu SJ, Bhatt N. Aided Selection of Sampling Methods for Imbalanced Data Classification. Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD). Association for Computing Machinery. 2021:198-202.

[28] D'Addabbo A, Maglietta R. Parallel Selective Sampling Method for Imbalanced and Large Data Classification. Pattern Recognit Lett. 2015;62:61-67.

[29] https://cris.haifa.ac.il/en/publications/lookahead-selective-sampling-using-a-distance-metric-based-on-ens

[30] Lindenbaum M, Markovitch S, Rusakov D. Selective Sampling for Nearest Neighbor Classifiers. Mach Learn. 2004;54:125-152.

[31] https://archive.ics.uci.edu/