Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

DOI: 10.37943/21VWQH9068

Zholdas Buribayev
PhD and Acting Associate Professor, Department of Computer Science
zhburibaev@gmail.com, orcid.org/0000-0002-3486-227X
Al-Farabi Kazakh National University, Kazakhstan
Ainur Yerkos
PhD candidate, Department of Computer Science
yerkosova@gmail.com, orcid.org/0000-0001-5949-6942
Al-Farabi Kazakh National University, Kazakhstan
Zhibek Zhetpisbay
Bachelor’s student, Department of Computer Science
zhibekzhetpisbay@gmail.com, orcid.org/0009-0005-7807-1444
Al-Farabi Kazakh National University, Kazakhstan

NEW APPROACH TO ADDRESSING CLASS IMBALANCE IN MEDICAL DATASETS CONSIDERING SPECIFICS

Abstract: At the moment, the popularization of the integration of machine learning into the field of medicine for data processing and analysis is being traced, but at the same time difficulties such as class imbalance and noisy datasets arise. Due to the prevalence of the problem, there are already existing solutions, but in all of them there is an abstraction from the field of medicine, namely, gender, racial and other differences are not taken into account. It is this side of the problem that is solved in our resampling algorithm. A feature of our algorithm is the use of splitting the dataset by an important feature through the p-value of Spearman correlation, which helps to consider subgroups of observations without losing their unique characteristics and removing noise data using LOF and Z-score separately for minority and majority classes, respectively. Synthetic data is generated in a flexible way, adapting to the data set using algorithm parameters. Work is provided with both quantitative and nominative features. The algorithm was tested on datasets for heart attack, chronic kidney disease, and liver disease, and the Random Forest ensemble method was used to train the model. After applying this class balancing method, improvements were recorded on average in Accuracy by 36%, in AUC by 15-25%, in Precision by 39-42%, and in Recall by 21-37% compared with SMOTE, ADASYN algorithms and the data set before balancing. Applying the algorithm on medical data can improve the accuracy of the algorithm and reduce the loss of reliability compared to other resampling methods.

Keywords: imbalance; oversampling; medical data analysis; data analysis; classification; noise filtering; nominal data

Introduction

Currently, one of the priority areas of development of the Republic of Kazakhstan is the healthcare sector with a budget of 3.77 trillion tenge for 2024 [1]. Compared to 2023, expenditures on Applied scientific research in the field of health and sanitary and epidemiological welfare of the population have increased from 3.08 billion tenge to 4.24 billion tenge [1]. The integration of artificial intelligence into medicine is a global trend and the estimated market growth from 20.9 billion dollars in 2024 will amount to 148.4 billion dollars in 2029 [2]. However, the introduction of artificial intelligence into medical practice faces a number of challenges. One of the main ones is the imperfection of the data used to train models. Data often contains outliers, is unevenly distributed between classes, and does not take into account important factors such as gender and race. Another important point is ethical issues. Due to ethical constraints, access to quality data is limited. As a result, it has to be used open sources that are of even lower quality. An urgent problem now is the imbalance of classes, due to which the output values are shifted towards the majority class. The result is a low sensitivity of the trained models to the minority class and this can lead to serious consequences when applying them in practice. To solve this problem, there are many resampling approaches, among which the most well-known ones can be distinguished: oversampling and undersampling. But even when generating data, the reliability of the data is questioned by medical researchers, since the features that different subgroups of observations may have not been taken into account. Existing resampling algorithms do not take these factors into account and generate synthetic samples on a wide variety of data.

If the existing solutions are considered, the algorithm from the article is aimed at removing outliers [3], using the OCSVM method. Further balancing takes place using the well-known SMOTE and ADASYN algorithms, which distinguishes it from our method, where its own generation algorithm is considered. The article [4] considers the problem of eliminating class overlap by modifying the SMOTE algorithm, namely dividing the data into 2 groups (safe, noise) and further generating new observations based on safe ones within the radius to the nearest majority point. The method from the article [5] is also based on the calculation of radii, which are equal to the distance to the nearest points of the majority and minority and determine the degree of classification contribution degree F, which is the ratio of radii. The authors of the article [6] in turn, use a probability distribution determined from the weights found from the proximity coefficient through the radius of the distance to the majority class.

In addition to determining the weights through the radius, there are other algorithms that calculate the optimality of points, which include the method [7] based on proximity factors to each class and generation occurring using M neighbors.

A different view of the definition of suitable elements is demonstrated by the algorithm [8] and uses KNN and from the selected observations new instances are generated by linear interpolation, which differs from our algorithm, which does not consider the minority and majority class together. KNN is also used in the article [9] to divide into “Danger” and “Safe” through the ratio of minority to majority objects and generation for each type occurs in a different way. Dividing only into 2 groups by location relative to the majority does not describe the full variability of the dataset, unlike our algorithm.

Another type of weighted sampling is applied in the article [10], the feature of which is the use of K-means clustering and the calculation of the sparsity of the cluster. Clustering is also used in the algorithm from the article [11] to divide the data space into areas with similar characteristics and, unlike the algorithm we developed, does not take into account outliers. The weights are calculated by estimating the density through the sparsity index and multiclass density. A feature of the method from the article [12] is the separation of observations from the minority group into clusters and the generation of new data using a Gaussian distribution with the parameters of the initial dataset.

The new BIRCH clustering method was used by the authors [13] after which the distribution density determines the weight and generation takes place using points as far apart as possible. The algorithm [14] uses the Mahalanobis distance to rank observations in clusters defined using K-Means and iteratively uses cross-synthesis with inheritance of 50% of the characteristics of the parent samples to generate new data. As in the methods [10], [11], [12], in contrast to our method, insufficient attention is paid to the effect of emissions.

The algorithm proposed in the article [15] considers ellipsoids instead of clusters, the size of which is determined by the multi-objective optimization scheme (MOPSO), and other EMDO parameters using the Mahalanobis distance. 2 methods are considered for generation, depending on the ratio of the minority in the ellipsoid. The Mahalanobis distance is also used in the article [16] when estimating the distance for constructing disjoint triangles using barycenters to generate new data. Oversampling is semi-controllable, which is expressed in the fact that new data can also be used in the generation process. The last method considered is the universal algorithm from the article [17], which differs from previous methods in the ability to work with nominative features. GSMOTE is used for data selection and generation in combination with a modified version of SMOTENC. Continuous and nominative signs are generated separately, but even with this method, it was not possible to take into account the nuances of medical data as in the methods [15], [16] in the form of factors requiring consideration of observation groups in isolation from each other.

Purpose and Objectives of Research

The purpose of this study is to create an algorithm for over-sampling suitable for processing medical data. To achieve this goal, the following tasks were set:

- analysis of current solutions for balancing classes;
- development of a new oversampling algorithm for medical data;
- application of the machine learning algorithm.

Methods and Materials

The process of this study can be divided into 6 stages: data collection, outlier removal, separation of the dataset by significant feature, clustering of subsamples, generation of new data and training with machine and deep learning algorithms. The whole process is demonstrated in Figure 1.

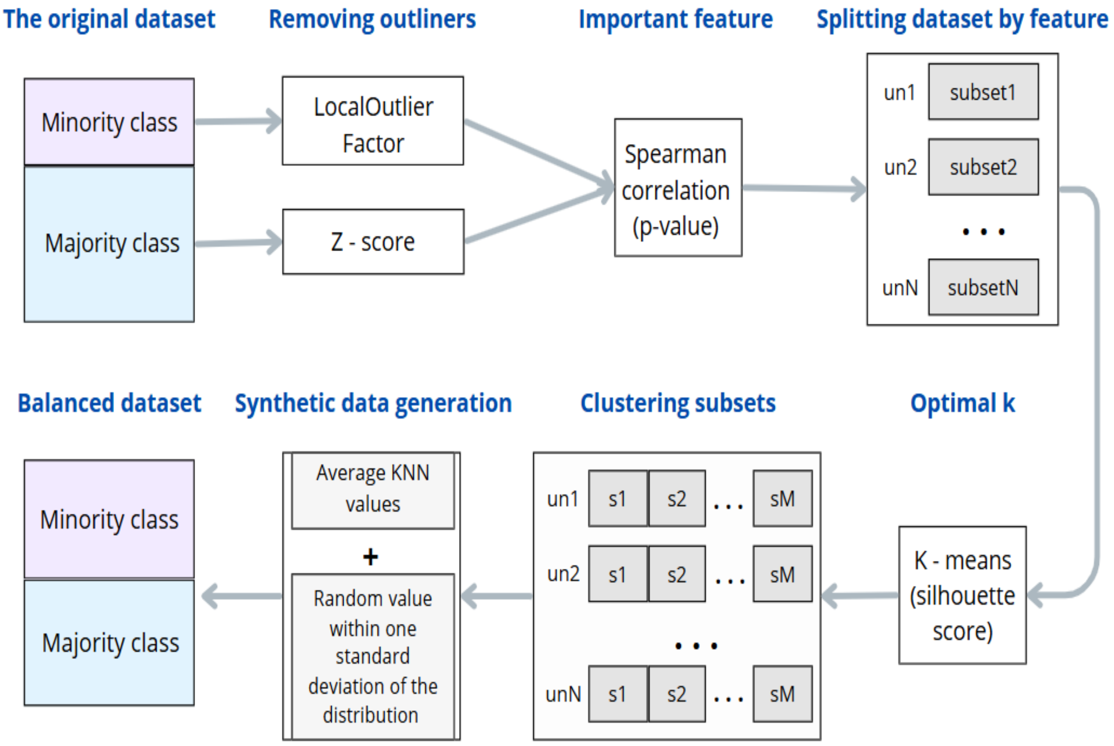


Figure 1. Diagram of the algorithm

A. Related works

1. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a resampling algorithm that generates synthetic instances based on two randomly selected observations. The values are taken randomly in the interval between the points, which allows you to create new samples near existing representatives of the minority class, while preserving their distribution and structure. However, this method can lead to the formation of duplicate or closely spaced points, which increases the risk of overfitting.

2. ADASYN

A special feature of the ADASYN (Adaptive Synthetic Sampling) method is the generation of minority class points in areas that are difficult to classify. Unlike SMOTE, ADASYN adaptively determines where more new samples are needed based on data density. Observations whose nearest neighbors are mostly instances of the majority class are more likely to be selected as the basis for new data. This allows the model to focus better on areas that are difficult to classify, but it can lead to more noisy data if the class boundaries are blurred.

Sources: sj.astanait.edu.kz, idaacs.net, embc.embs.org, jyi.org, mdpi.com, nature.com, sjsu.edu

B. Data collection

1. Collecting datasets

In this study, data sets on heart attack [18], chronic kidney disease [19] and liver disease [20] were used. These types of diseases were chosen because of their significant prevalence. In 2019, 19.91 million deaths due to cardiovascular diseases were recorded worldwide [21]. An international study in 2023 estimated that approximately 850 million people have chronic kidney disease, affecting all ages [22]. In 2024, liver diseases caused the death of about 2 million people, which accounted for 4% of the total mortality [23].

2. Data separation

The data was divided as follows: 90% was used for training and validation with 5 or 3 folds, depending on the size of the dataset, and the remaining 10% was intended for testing. The test sample includes data before generation that does not include synthetically created observations. Information with dataset ratios is shown in Table 1.

Table 1. Datasets used in the study

№	Name of dataframe	Count of objects	Count of features (with target)	Imbalance ratio	Target value counts
1	Heart Disease	4238	16	11:1	1524:135
2	Chronic Kidney Disease	1659	54	5:1	3594:644
3	Indian Liver Patient Records	583	11	5:2	414:165

C. Outlier removal

To clear the datasets of noise data, the minority class and the majority class were considered separately. The LOF method was applied on the minority class and the Z-score method on the majority class.

1. LOF (Local Outlier Factor)

LOF measures local deviations in the density of observations relative to neighboring samples and is the degree of isolation of objects. Not the entire set of observations is considered, but a certain k-nearest neighbors, the distances to which will be used for estimation, which is convenient for a minority class characterized by sparsity of data.

2. Z-score

In this method, the distribution density is estimated using the Z-score, which is a statistical estimate of the ratio of the observation to the sample average. The data whose values are less than the threshold values are selected. All elements from the majority class are considered and (-3, 3) are selected as thresholds.

D. Separation of the dataset by significant feature

A feature of our algorithm is the separation of the dataset by an important feature, which helps to consider each group of objects in isolation and without losing their features. To determine a significant feature, the p-value from Spearman correlation was used. With nominative features, the dataset is divided by a unique value, in the absence of nominative features, the Quantile-based discretization function is used, which segments the feature into groups of the same size.

E. Clustering of subsamples

The possibility of selecting points that are far from each other is suppressed by the additional separation of subsamples using K-Means and Silhouette Coefficient. K-Means clustering is an algorithm based on dividing a dataset into clusters according to the degree of proximity to centroids. The number of clusters is determined using the Silhouette Coefficient, based on the ratio of the average between the cluster and the average within the cluster distance.

F. Generation of new data

The creation of synthetic data can be divided into 2 stages: setting weights, generating synthetic data.

1. Setting weights

Data clusters are given a weight to determine the number of synthetic elements that will be produced using cluster points. The weights are determined by the ratio of the number of observations in the sample in relation to the entire population of the minority class according to the formula (1):

$$W_c = \frac{N_c}{N_{min}} \tag{1}$$

where N_c is the number of observations in the cluster, N_{min} is the number of observations in the minority class.

2. Generating synthetic data

New observations are obtained using a random element and the nearest k data points selected by the Nearest Neighbors algorithm. The number k is an input parameter and is set by the user. The new value is determined by the average for the considered observation with the addition of a value from the Gaussian distribution scaled by feature as shown in (2):

$$X_{new} = \frac{\sum_{i=1}^{k} X_i}{k} + random(scale = \sigma_f) \tag{2}$$

where k is the number of observations to generate 1 new point, X_i is the values of existing observations, σ_f is the standard deviation of the feature under consideration. In the case of nominative features, the value with the maximum number of appearances among the considered group of neighbors is selected.

G. Classification algorithms

1. Machine learning algorithms

Random forest and Light Gradient-Boosting Machine (LightGBM) algorithms, which are ensemble methods convenient for classification, were used to train the model based on the generated data set. When applying them to the initial dataset, class balancing parameters were used to compare learning outcomes before and after synthetic sample generation.

2. Deep learning algorithm

In the work was used a Multi-Layer Perceptron two hidden layers that contain 64 and 32 neurons, ReLU and Sigmoid activation functions. The quality of the model was assessed using the following metrics: Accuracy, Precision, Recall, F1-score and ROC-AUC. Accuracy allows you to estimate the total proportion of correct predictions, Precision and Recall – the balance of predictions for each class, F1-score takes into account both completeness and accuracy, and ROC-AUC demonstrates the quality of the model at various classification thresholds.

The pseudocode of the developed oversampling algorithm is shown in Figure 2 below.

```
Algorithm 1 Oversampling Algorithm
Require: Dataset $D$, target feature target
Ensure: Balanced dataset with synthetic examples
```

Sources: sj.astanait.edu.kz, idaacs.net, embc.embs.org, jyi.org, mdpi.com, nature.com, sjsu.edu

```
Function for Outlier Removal (LOF and Z-score):
Apply LOF to $C_{min}$ (minority class)
Apply Z-score to $C_{maj}$ (majority class)
Remove outliers using specified threshold values
Function for Feature-based Splitting (Spearman Correlation):
Identify significant feature $F_{imp}$ with lowest $p$-value
if $F_{imp}$ is categorical then
    Split data by unique values of $F_{imp}$
else
    Use quantile-based discretization for data splitting
end if
Function for Clustering Subsets (K-means with silhouette score for cluster count):
for each data subset do
    Apply K-means with cluster count $K$ based on silhouette score
end for
Function for Synthetic Data Generation:
for each cluster k do
    for $i = 1$ to $N_{syn}$ (required number of synthetic samples) do
        Select k nearest neighbors $X_1, X_2, \dots, X_k$ of a random point from cluster $k$
        Generate new sample: $X_{new} = \frac{1}{k} \sum_{i=1}^{k} X_i$ + random(scale = $\sigma_f$)
    end for
end for
Training the Classifier:
Train a Random Forest classifier on the generated dataset
Evaluate model performance (Accuracy, Precision, Recall, AUC)
return Synthetic augmented dataset
```

Figure 2. Pseudocode of the new oversampling algorithm

Results

After applying the resampling algorithm, noticeable improvements were achieved for all 4 metrics. On average, Accuracy increased by 49%, AUC by 23%, Precision by 85%, Recall by 58% compared to the initial data set. Regarding SMOTE, the algorithm's results showed an improvement of 18% for Accuracy, 6% for AUC, 16% for Precision, and 17% for Recall. When compared with ADASYN, the results increased by 18% for Accuracy, 19% for Precision, and 21% for Recall. The results are shown in tables 2 and 3 below. An analysis of the classifiers showed that Random Forest is the best, as it consistently demonstrates high results both before and after oversampling, especially when using the new method. It shows a better balance between Precision and Recall, which is important for tasks with unbalanced data. LightGBM is also a good choice, especially when a balance between metrics is important, but its effectiveness depends on the type of data. Multilayer Perceptron (MLP) показал самые низкие результаты среди классификаторов, даже после передискретизации, уступая Random Forest и LightGBM по всем метрикам. Tables 2 -5 below demonstrate the results.

Table 2. Metrics before oversampling

Datasets	Method	Accuracy	AUC	Precision	Recall
Heart Disease	Random forest classifier	0.6583	0.7331	0.6939	0.5667

	LightGBM	0.5000	0.7183	-	-
	Multilayer Perceptron	0.5000	0.6312	0.5000	0.0256
Chronic Kidney Disease	Random forest classifier	0.6000	0.6000	0.5556	1.0000
	LightGBM	0.5500	0.7500	0.5263	1.0000
	Multilayer Perceptron	0.5000	0.7100	0.5000	1.0000
Indian Liver Patient Records	Random forest classifier	0.6000	0.7067	0.6000	0.6000
	LightGBM	0.5667	0.7022	0.6000	0.4000
	Multilayer Perceptron	0.4667	0.6400	-	-

Table 3. Metrics after oversampling with Random forest classifier

Datasets	Method	Accuracy	AUC	Precision	Recall
Heart Disease	SMOTE	0.7083	0.7786	0.7273	0.6667
	ADASYN	0.7000	0.7703	0.7222	0.6500
	New approach	**0.7917**	**0.8456**	**0.8302**	**0.7333**
Chronic Kidney Disease	SMOTE	0.6500	0.8100	0.5882	1.0000
	ADASYN	0.6500	0.9600	0.5882	1.0000
	New approach	**1.0000**	**1.0000**	**1.0000**	1.0000
Indian Liver Patient Records	SMOTE	0.7000	0.7467	0.7143	0.6667
	ADASYN	**0.7667**	**0.8400**	**0.7222**	**0.8667**
	New approach	0.7333	0.8222	0.7059	0.8000

Table 4. Metrics after oversampling with LightGBM

Datasets	Method	Accuracy	AUC	Precision	Recall
Heart Disease	SMOTE	0.6167	0.7911	0.8500	0.2833
	ADASYN	0.6250	0.8186	0.8000	0.3333
	New approach	**0.7500**	**0.8128**	**0.8571**	**0.6000**
	SMOTE	0.8000	0.9400	0.7500	0.9000

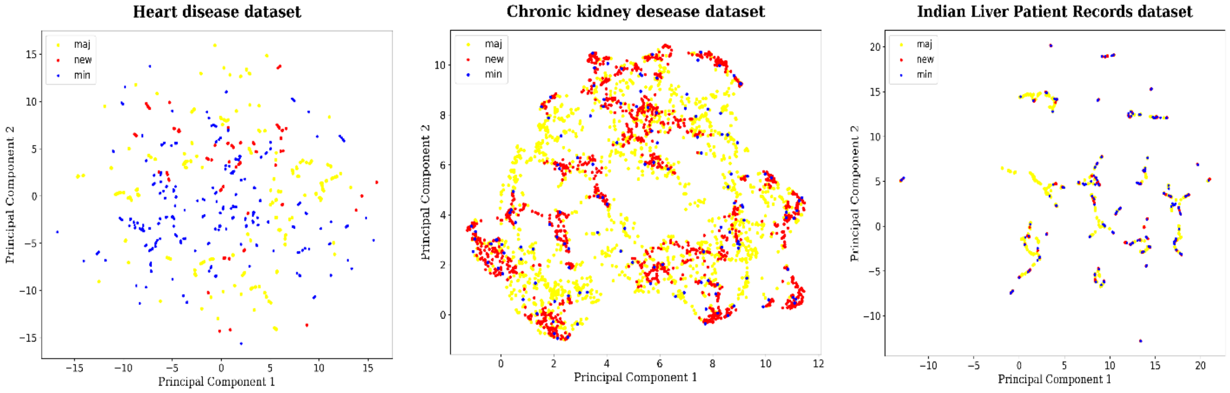
Sources: sj.astanait.edu.kz, idaacs.net, embc.embs.org, jyi.org, mdpi.com, nature.com, sjsu.edu

Chronic Kidney Disease	ADASYN	0.9000	0.9800	0.9000	0.9000
	New approach	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Indian Liver Patient Records	SMOTE	0.7000	0.8000	0.6875	0.7333
	ADASYN	0.7667	**0.8756**	**0.8000**	0.8000
	New approach	**0.8000**	0.8222	0.7059	0.8000

Table 5. Metrics after oversampling with Multilayer Perceptron

Datasets	Method	Accuracy	AUC	Precision	Recall
Heart Disease	SMOTE	0.5667	0.6267	0.5714	0.5333
	ADASYN	0.5583	0.6217	0.5660	0.5000
	New approach	**0.6583**	**0.7442**	**0.6338**	**0.7500**
Chronic Kidney Disease	SMOTE	0.8000	**0.9000**	0.7500	0.9000
	ADASYN	0.8000	0.9700	0.7500	0.9000
	New approach	**0.8500**	0.8500	**0.8182**	0.9000
Indian Liver Patient Records	SMOTE	0.6667	0.7867	0.7273	0.5333
	ADASYN	0.4667	0.7911	0.3333	0.0667
	New approach	**0.8000**	**0.8133**	**0.7368**	**0.9333**

To visualize the data shown below in Figure 3, the UMAP algorithm was used, designed to reduce the dimensionality of the data.



a) Heart Disease
b) Chronic Kidney Disease
c) Indian Liver Patient Records

Figure 3 . Visualization of oversampling results

Discussion

The proposed oversampling algorithm demonstrates a significant improvement in classification metrics compared to traditional methods such as SMOTE and ADASYN when applied to medical data. This is due to the fact that these methods consider the entire data set, without taking into account the possibility of groups for which a deviation of synthetic data from the standard range for a feature can be considered serious and leads to a loss of model quality. The proposed approach, which accounts for the specific characteristics of medical data such as gender, race, and other differences, has significantly improved the quality of machine learning models. Dividing the data by a significant feature and subsequent clustering allow for a more accurate consideration of the features of each subgroup of observations, which is especially important for medical data where hidden heterogeneity is often present. The use of LOF and Z-score for outlier removal also improves data quality and enhances model performance.

Conclusion

This paper presents a new oversampling algorithm for class imbalance in medical data. The developed oversampling algorithm is an effective tool for balancing classes in medical data, taking into account their specific characteristics and allowing for more accurate results when building machine learning models. The results of the experimental evaluation showed that the proposed algorithm outperforms existing methods in terms of the main classification quality metrics. The developed algorithm can be used to solve a wide range of problems in the field of medical diagnosis and prognosis. In the future, it is planned to expand the algorithm's functionality by considering time series and multi-class classification.

Funding

This research was funded by the Grant No. AP23489229 “Development of a multimodal artificial intelligence-based application for kidney transplant data preprocessing and analysis” of Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

[1] *Bjudzhetnyj miks na 2024 god: +19%.* (n.d.). Retrieved November 5, 2024, from https://ulagat-m.kz/analyst/finansy-v-meditsine/byudzhetnyy-miks-na-2024-god-19.html

[2] *Artificial Intelligence (AI) in Healthcare Market Size, Share, Trends & Industry Growth Analysis Report 2032.* (n.d.). MarketsandMarkets. Retrieved November 5, 2024, from https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-healthcare-market-54679303.html

[3] Gnip, P., Vokorokos, L., & Drotár, P. (2021). Selective oversampling approach for strongly imbalanced data. *PeerJ Computer Science, 7*, e604. https://doi.org/10.7717/peerj-cs.604

[4] Pradipta, G. A., Wardoyo, R., Musdholifah, A., & Sanjaya, I. N. H. (2021). Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning From Imbalanced Data. *IEEE Access*, *9*, 74763–74777. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3080316

[5] Jiang, Z., Pan, T., Zhang, C., & Yang, J. (2021). A new oversampling method based on the classification contribution degree. *Symmetry*, *13*(2), 1–13. Scopus. https://doi.org/10.3390/sym13020194

[6] Tao, L., Li, H., Wang, F., Liu, M., Tang, Z., & Wang, Q. (2024). An Adaptive Safe-Region Diversity Oversampling Algorithm for Imbalanced Classification. *IEEE Access*, *12*, 63713–63724. Scopus. https://doi.org/10.1109/ACCESS.2024.3396155

[7] Kunakorntum, I., Hinthong, W., & Phunchongharn, P. (2020). A Synthetic Minority Based on Probabilistic Distribution (SyMProD) Oversampling for Imbalanced Datasets. *IEEE Access*, *8*, 114692–114704. Scopus. https://doi.org/10.1109/ACCESS.2020.3003346

[8] Yi, X., Xu, Y., Hu, Q., Krishnamoorthy, S., Li, W., & Tang, Z. (2022). ASN-SMOTE: A synthetic minority oversampling method with adaptive qualified synthesizer selection. *Complex and Intelligent Systems*, *8*(3), 2247–2272. Scopus. https://doi.org/10.1007/s40747-021-00638-w

[9] Chen, Y., Zou, J., Liu, L., & Hu, C. (2024). Improved Oversampling Algorithm for Imbalanced Data Based on K-Nearest Neighbor and Interpolation Process Optimization. *Symmetry*, *16*(3), Article 3. https://doi.org/10.3390/sym16030273

[10] Wang, Z., & Wang, H. (2021). Global Data Distribution Weighted Synthetic Oversampling Technique for Imbalanced Learning. *IEEE Access*, *9*, 44770–44783. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3067060

[11] Deng, Y., & Li, M. (2023). An Adaptive and Robust Method for Oriented Oversampling With Spatial Information for Imbalanced Noisy Datasets. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2023.3329560

[12] Hassan, M. M., Eesa, A. S., Mohammed, A. J., & Arabo, W. K. (2021). Oversampling method based on gaussian distribution and K-means clustering. *Computers, Materials and Continua*, *69*(1), 451–469. Scopus. https://doi.org/10.32604/cmc.2021.018280

[13] Wang, C.-R., & Shao, X.-H. (2020). An Improving Majority Weighted Minority Oversampling Technique for Imbalanced Classification Problem. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2020.3047923

[14] Zhang, Y., Zuo, T., Fang, L., Li, J., & Xing, Z. (2020). An Improved MAHAKIL Oversampling Method for Imbalanced Dataset Classification. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2020.3047741

[15] Yao, L., & Lin, T.-B. (2021). Evolutionary Mahalanobis Distance-Based Oversampling for Multi-Class Imbalanced Data Classification. *Sensors*, *21*(19), Article 19. https://doi.org/10.3390/s21196616

[16] Liu, D., Qiao, S., Han, N., Wu, T., Mao, R., Zhang, Y., Yuan, C.-A., & Xiao, Y. (2020). SOTB: Semi-Supervised Oversampling Approach Based on Trigonal Barycenter Theory. *IEEE Access*, *8*, 50180–50189. Scopus. https://doi.org/10.1109/ACCESS.2020.2980157

[17] Fonseca, J., & Bacao, F. (2023). Geometric SMOTE for imbalanced datasets with nominal and continuous features. *Expert Systems with Applications*, *234*. Scopus. https://doi.org/10.1016/j.eswa.2023.121053

[18] *Heart Disease Dataset*. (n.d.). Retrieved November 1, 2024, from https://www.kaggle.com/datasets/mirzahasnine/heart-disease-dataset

[19] *Chronic Kidney Disease Dataset*. (n.d.). Retrieved November 1, 2024, from https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis

[20] *Indian Liver Patient Records*. (n.d.). Retrieved November 1, 2024, from https://www.kaggle.com/datasets/uciml/indian-liver-patient-records

[21] *More than half of U.S. adults don’t know heart disease is leading cause of death, despite 100-year reign*. (n.d.). American Heart Association. Retrieved November 1, 2024, from https://newsroom.heart.org/news/more-than-half-of-u-s-adults-dont-know-heart-disease-is-leading-cause-of-death-despite-100-year-reign

[22] admin, I. S. N. (2023, March 30). New global kidney health report sheds light on current capacity around the world to deliver kidney care. *International Society of Nephrology*. https://www.theisn.org/blog/2023/03/30/new-global-kidney-health-report-sheds-light-on-current-capacity-around-the-world-to-deliver-kidney-care/

[23] Baviskar, K., Kshirsagar, A., Raut, H., & Shaikh, M. R. N. (2024). Overview: Global burden of liver disease. *International Journal of Pharmaceutical Chemistry and Analysis*, *11*(1), 1–10. https://doi.org/10.18231/j.ijpca.2024.001

Sources: sj.astanait.edu.kz, idaacs.net, embc.embs.org, jyi.org, mdpi.com, nature.com, sjsu.edu