IRSTI 28.23.02 https://doi.org/10.26577/jpcsit2024-02i03-02

Zh. Buribayev , A. Yerkos , S. Shaikalamova ,
R. Imanbek* , Zh. Zhetpisbay

Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: imanbek.rustem2000@gmail.com

IMPROVING MEDICAL DIAGNOSIS WITH A HYBRID BALANCING TECHNIQUE

Abstract. The issue of class imbalance in medical data poses a significant challenge for developing robust machine learning models aimed at medical diagnosis. A characteristic feature of such data is the substantial dominance of instances belonging to majority classes (e.g., healthy patients or those with common diseases) over instances representing rare conditions. This disproportion leads to machine learning models trained on such data being prone to systematic classification errors, predominantly predicting the most frequent class. Consequently, the ability of models to accurately identify rare cases is severely diminished. This paper proposes a hybrid class-balancing algorithm that combines Inverse Quadratic Radial Undersampling (IQRBU) and genetic oversampling to address this issue. The integration of these two methods within a single algorithm achieves an optimal balance between preserving information and enhancing the representation of rare classes. Experimental results conducted on several medical datasets demonstrated the effectiveness of the proposed approach. The obtained results showed that the hybrid algorithm significantly improves classification metrics, such as the F1-score and accuracy. These findings underscore the potential of our approach to enhance the reliability and precision of medical diagnostic systems.

Key words: Imbalanced Data; Radial Basis; Hybrid; Undersampling; Oversampling, Genetic Algorithms.

1. Introduction

The technological revolution in medicine has led to a significant increase in the volume of medical data, which is growing at an exponential rate, creating new challenges and opportunities for medical practice and research [1]. These medical data include laboratory test results, radiological images, and patient medical histories, making them a valuable resource for modern healthcare and scientific research. According to information from the international scientific research database Scopus, more than 98,000 scientific publications on the processing and analysis of medical data were published between 2000 and 2024 [2]. Furthermore, it is worth noting that according to Statista, the market volume of artificial intelligence technologies in healthcare, which was valued at $11 billion in 2021, is projected to reach an impressive $187 billion by 2030 [3]. This underscores the urgent need for processing and analyzing such data, driving the application of artificial intelligence methods, particularly machine learning (ML).

However, there are several challenges that hinder the effective functioning of machine learning methods and can lead to incorrect diagnoses. One such challenge is class imbalance, which in the medical field is caused by the uneven distribution of healthy and sick patients. This can result in algorithms favoring the class with a larger number of examples, thereby reducing the accuracy of diagnosing rare diseases.

There are numerous methods aimed at addressing the issue of class imbalance in machine learning, such as oversampling, undersampling, and hybrid methods. Some of the most well-known class balancing techniques are those based on random addition and deletion of data. These methods are relatively simple to implement, but they have drawbacks, such as overfitting and the loss of important information. Another well-known class of balancing algorithms includes those based on distance calculation. The SMOTE oversampling method, for instance, works by increasing the size of the minority class through the generation of synthetic instances based on the nearest neighbors of existing samples. Another method, Tomek Links, is an undersampling algorithm that identifies pairs of instances from different classes that are nearest neighbors with the smallest distance, and then removes the majority class instances involved in these pairs.

This paper proposes a hybrid class-balancing algorithm that includes the IQRBU (Inverse Quadratic Radial-Based Undersampling) algorithm [4]. This algorithm employs an inverse quadratic radial basis function to evaluate the potential of instances and preserve the original data distribution by controlling the standard deviation. Additionally, the proposed approach incorporates a oversampling algorithm, which is a modified version of SMOTE, based on a genetic algorithm.

2. Materials and Methods

This chapter provides a detailed description of the methods used in the conducted research. The chapter is composed of several key sections, namely: data collection, data preprocessing, RBU [4], and genetic algorithm-based oversampling.

2.1 Data collection

In this research study, three different medical datasets containing information on various diseases were used, namely: Alzheimer's disease, diabetes, and cancer. These diseases were selected due to their significant and detrimental impact on public health in modern society. According to the World Health Organization (WHO), more than 55 million people worldwide suffer from dementia, with 60-70% of them being Alzheimer's patients [5]. Cancer also poses a serious challenge, being one of the leading causes of mortality, resulting in 10 million deaths in 2020 [6]. This statistic highlights the importance of research and the development of effective methods for diagnosis and treatment. Additionally, as of 2014, there were 422 million reported cases of diabetes worldwide [7]. These figures indicate that diabetes is also one of the most prevalent and dangerous diseases, necessitating attention and scientific inquiry. The datasets were sourced from the Kaggle platform [8], which is a popular resource for data sharing and research in data science. Table 1 provides information on the datasets used in this study.

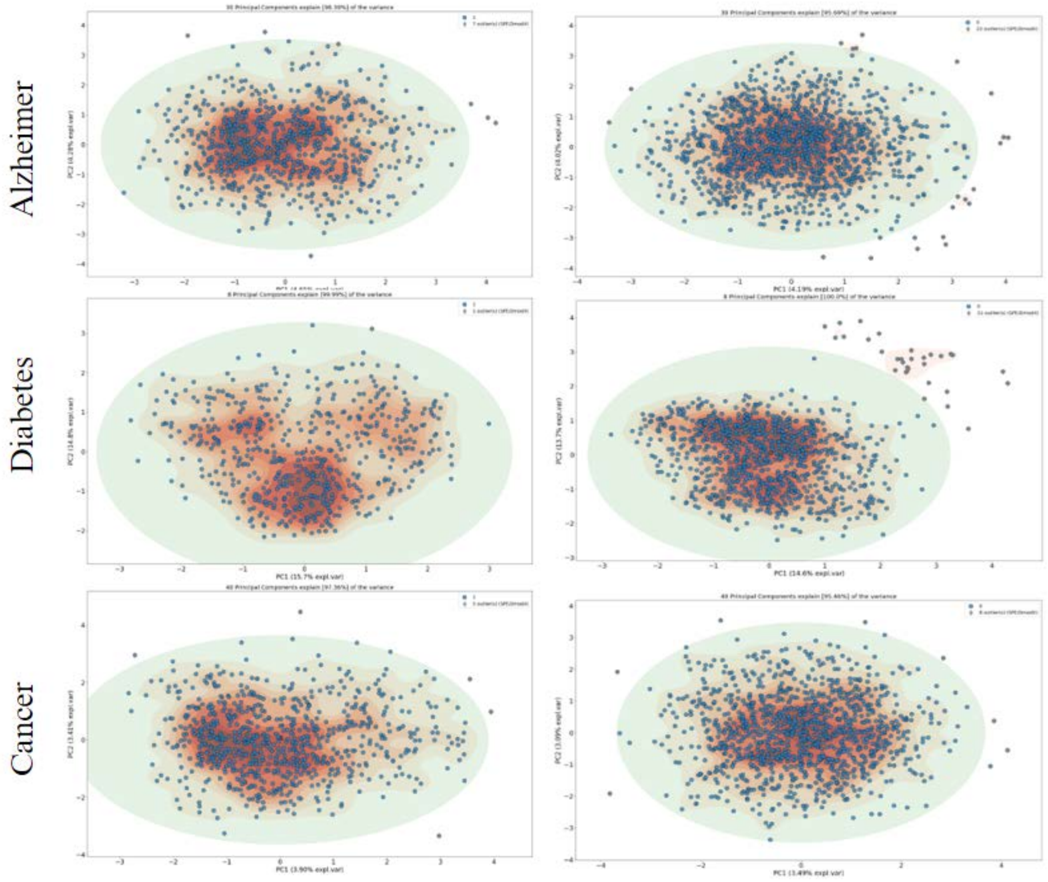
Table 1 – Unbalanced data sets used in this study

Name	objects	Features (with target)	ratio
alzheimer	2149	35	9:5
diabetes	1879	46	3:2
cancer	1500	9	17:10

2.2 Data preprocessing

Data preprocessing is one of the most critical and essential stages in this study, as inadequate or poor-quality preprocessing can lead to incorrect predictions and significantly reduce the effectiveness of machine learning methods. In this research, a linear PCA (Principal Component Analysis) algorithm was used to remove instances belonging to both minority and majority classes, applying the SPE (Squared Prediction Error) measure. SPE is defined as the sum of squared differences between the original data and their reconstructed values obtained after projection onto the principal component space. This allows for the assessment of how well the model reproduces the data and helps to identify anomalies that could negatively impact the quality of predictions.

To implement the PCA algorithm using SPE, the PCA library of the Python programming language was utilized [9]. Figure 1 presents the results of identifying noisy data in the minority and majority classes across three different datasets. Additionally, during the preprocessing stage, duplicates, features with only one unique value, and features where the number of unique values equaled the number of instances in the dataset were identified and removed. This is because such features do not carry informative value and can confuse the model, thereby reducing its predictive capability.



a) Noise Identification in Minority Class Data

b) Noise Identification in Majority Class Data

Figure 1 – Noise Identification in Data Using PCA

2.3 IQRBU

The IQRBU (Inverse Quadratic Radial-Based Undersampling) algorithm is used as the undersampling technique in this study. This algorithm is designed to reduce the number of majority class instances while maintaining a balance between classes in the training dataset. The core of the algorithm involves the use of an Inverse Quadratic Radial Basis Function (IQ-RBF) to evaluate the potential of each majority class instance. Unlike the RBU method, which employs a Gaussian Radial Basis Function [4], the IQRBU method uses an Inverse Quadratic Radial Basis Function to more accurately preserve the original data distribution.

The form of the Inverse Quadratic Radial Basis Function (IQ-RBF) is expressed by the following formula (1)

$$IQ - RBF(d) = \frac{1}{d^2+c^2} \tag{1}$$

where d is the Euclidean distance between two points, and c is a parameter that controls the shape of the function. The IQ-RBF is characterized by the fact that the function's value smoothly decreases as the distance increases, making it particularly useful for tasks where it is important to consider both nearby and distant objects.

The IQRBU method involves several stages. First, the potential of each majority class instance is calculated based on the distance to instances from both classes. The potential of an instance xxx from the majority class is computed using the following formula (2) [10].

$$\phi(x, K, k, \text{c}) =$$

$$= \sum_{i=1}^{|K|} \frac{1}{\left\|K_i - x\right\|^2 + c^2} - \sum_{j=1}^{|k|} \frac{1}{\left\|k_j - x\right\|^2 + c^2} \tag{2}$$

where K denotes the set of majority class instances, κ represents the set of minority class instances, and c is the parameter controlling the shape of the radial basis function. This calculation allows for determining the extent to which each majority class instance influences the class balance in the dataset.

Subsequently, instances with the highest potential are iteratively removed until the desired class ratio is achieved. After each removal, the potential of the remaining instances is recalculated, taking into account the changes in data distribution (3).

$$\phi_i \leftarrow \phi_i - \frac{1}{\left|\left|K_i' - x\right|\right|^2 + c^2} \tag{3}$$

where K' represents the remaining majority class instance, and x is the removed instance with the highest potential. This recalculation allows for considering the changing influence of each instance and adjusting the removal strategy accordingly.

It is important to note that the IQRBU method places significant emphasis on preserving the original data distribution. The standard deviation of the majority class instances' distribution is controlled to avoid significant distortions. The removal of instances continues until the standard deviation of the remaining instances falls below 1, which helps to minimize information loss and retain the core characteristics of the data.

Thus, the IQRBU method ensures the effective removal of majority class instances, improving class balance while maintaining crucial aspects of the original data distribution. This approach significantly enhances the performance of machine learning models, particularly in tasks related to medical data analysis, where accuracy and reliability are of critical importance.

2.4 Genetic Algorithm-Based oversampling

In this study, a genetic algorithm is employed as the oversampling technique, with the goal of finding an optimal solution to the problem at hand [11]. It is important to note that a modified version of the GenSMOTE algorithm [12] is used, where a different selection method is applied. In this case, it is used to optimize the oversampling frequencies in the SMOTE algorithm. Any genetic algorithm consists of several stages, including population creation, fitness function, selection method, crossover, and mutation.

During the initialization stage of the population, each individual, or instance within the population, represents a set of genes, which in this study are the oversampling frequencies. Next, for each individual, a fitness function is calculated to determine how well the individual meets the objective of the task. In this study, the F1-score was used as the fitness function. The next stage in the genetic algorithm is the selection method. The selBest selection method [13] is applied, where the top mmm individuals from the current population are chosen based on their fitness values. These selected individuals have the highest fitness and, therefore, a higher likelihood of passing their genes on to subsequent generations, facilitating the achievement of an optimal solution during the evolutionary process. The description of this selection method is demonstrated in equation (4):

$$selBest(I, m, f) = Sort_{desc}((I, f)[1{:}m]) \tag{4}$$

where I represents the set of all individuals, m is the number of the best individuals to be selected, and f is the fitness value used as the sorting criterion.

The selected individuals then proceed to the crossover stage, where the genetic information of randomly selected pairs is combined to create new offspring. Next, the mutation method is applied to alter the genetic structure of the population. In this study, two-point crossover [14] and uniform mutation [15] methods were used. As a result of applying these methods, the genetic algorithm identifies the optimal oversampling frequencies. The DEAP library of the Python programming language [16] was used to implement genetic algorithm-based oversampling in this work.

Algorithm 1 illustrates the working principle of the hybrid class balancing approach.

Algorithm 1: IQRBUwOGA

```
Require: IQRBU, oversampling with GA
Ensure: optimal collection of sampling rates
Function IQRBU (K, k, c):
K' ← K
For every majority object $K_i'$ from K' and its associated potential
$\Phi_i$ do
        $\Phi_i \leftarrow \Phi_i\ (K_i', K, k, c, ratio)$
end for
while $|K| - |K'| < ratio(|K| - |k|)$ do
for every majority object $K_i'$ from K' and its associated potential
$\Phi_i$ do
$\Phi_i \leftarrow \Phi_i - \frac{1}{\|K_i' - x\|^2 + c^2}$
 If $|std(K) - std(K')| \leq 1.0$ then
 drop point x from K'
 else
 break
 end if
end while
        $K \leftarrow K'$
function oversampling with GA (num_iterations):
 The goal is to find the optimal X (sampling rates) with the Genetic Algorithm
 $X = (F_1, F_2, F_3, F_4, \dots F_N)$
 $P = (X_1, X_2, X_3, X_4, \dots X_j)$
 F1_score(SMOTE(P))
 In the case of SMOTE, the synthetic is generated as follows:
        $x_{new} \leftarrow EuclideanDistance(x_1, x_2 * random(0,1))$
make the classification using DecisionTree and get train and test sets calculating the fitness function for each
$X_j$(F1_score)
while $i < num\_iterations$ do
 calculate fitness function
 selBest selection
 two point crossing operation and
 uniform mutation
 i++
end while
return optimal X
```

The IQRBUwOGA algorithm begins with initialization, where each majority class instance is assigned an initial potential. Then, for each instance, the potential is calculated using the Inverse Quadratic Radial Basis Function (IQ-RBF), which allows for the assessment of its impact on class balance. Instances with the highest potential are iteratively removed until the desired class ratio, specified by the parameter $ratio$, is achieved. This parameter defines the target ratio between classes after the removal of instances. The process continues until the standard deviation of the remaining instances falls below a set threshold, helping to preserve important data features.

Next, the oversampling algorithm begins its operation, where $num\_iteration$ is the number of generations is specified. The initial population is initialized. The fitness value for each individual in the population is iteratively calculated by performing SMOTE on the dataset. SMOTE generates synthetic minority class instances by calculating the distance to the nearest neighbor, multiplying it by a random number, and adding it to the original instance. The distance is measured using the Euclidean distance metric. The resulting data are then used to train a decision tree. The fitness function is evaluated using the F1-score. Following this, the selection, crossover, and mutation methods are executed. As a result of genetic algorithm-based oversampling, the individual with the optimal oversampling frequencies is returned.

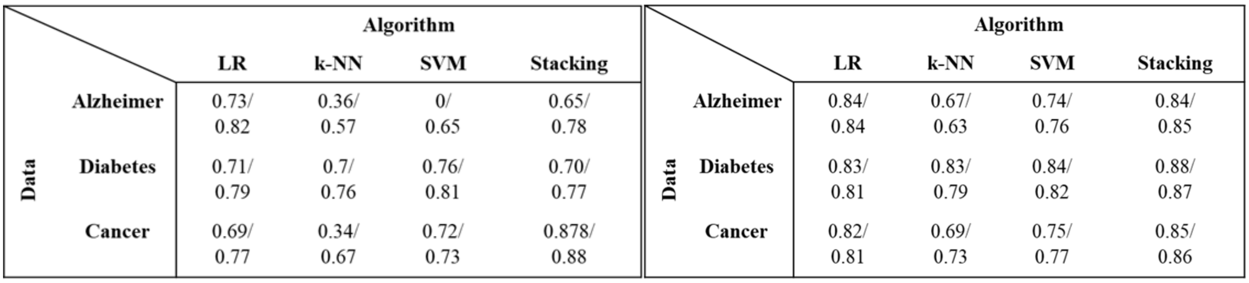
3. Results

This section presents the results obtained during the research. The data were split into 95% for training and validation, with 5% allocated for the test set. A 3-fold cross-validation was used to evaluate the models. The datasets were trained using basic machine learning methods and the ensemble stacking method.

Stacking is an ensemble approach in machine learning where multiple different models, referred to as base models, are combined to improve the overall accuracy of predictions. The principle of stacking involves using the predictions of the base models as input data for a meta-model, which is trained on these predictions and makes the final decision. Stacking differs from other ensemble methods in that it offers flexibility in selecting different base models and allows the meta-model to adaptively combine their predictions, capturing diverse aspects of the data. In this study, various machine learning algorithms, such as Logistic Regression, Support Vector Machine, and K-Nearest Neighbors, were used as base models. The meta-model was a Decision Tree algorithm, which integrated and processed the outputs of the base models, optimizing the overall prediction result and enhancing classification accuracy.

The results of F1-score and accuracy metrics before and after hybrid balancing are presented in Table 2.

Table 2 – F1/Accuracy Results Before and After Class Balancing

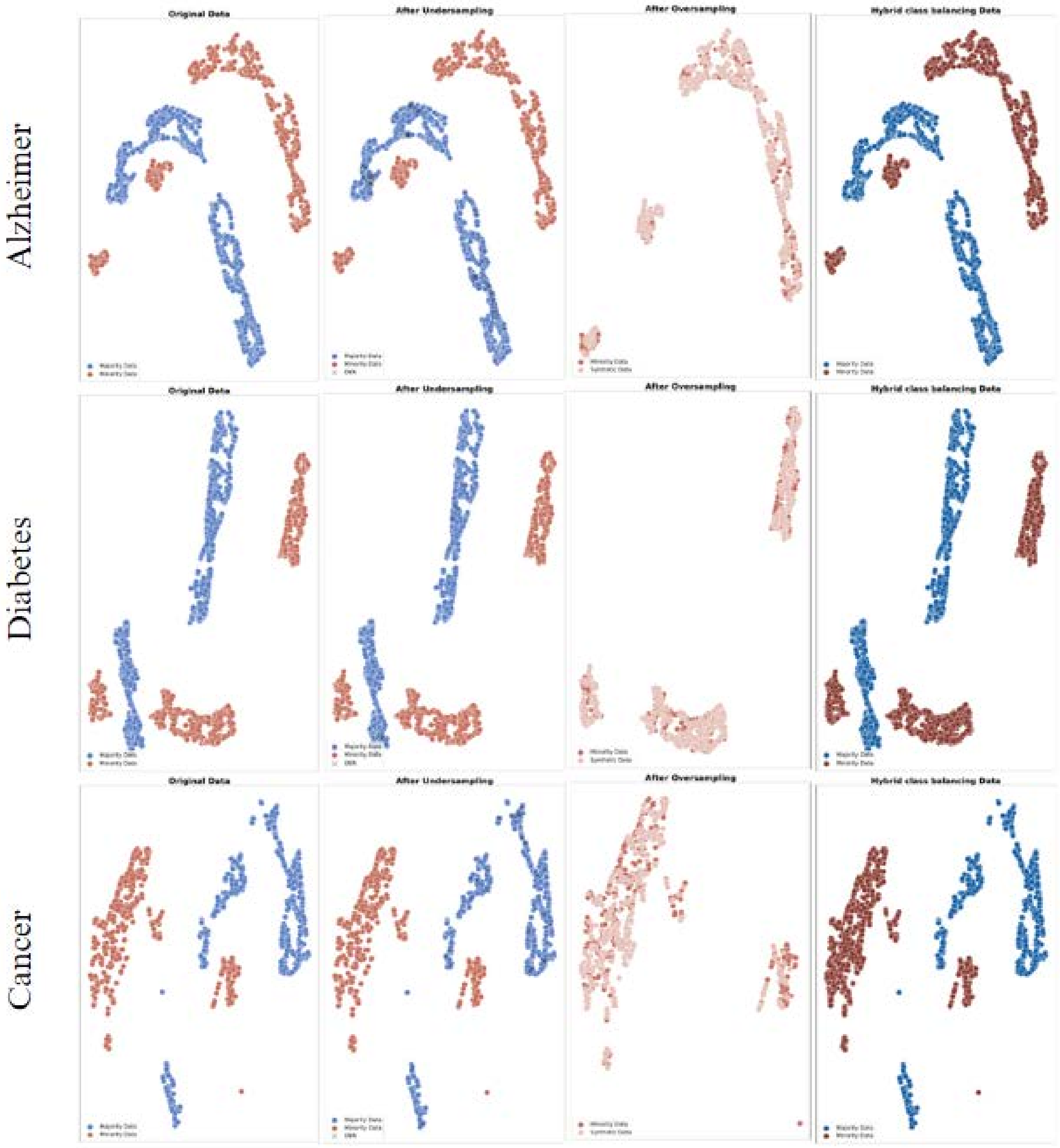


		Algorithm			
		LR	k-NN	SVM	Stacking
Data	Alzheimer	0.73/ 0.82	0.36/ 0.57	0/ 0.65	0.65/ 0.78
	Diabetes	0.71/ 0.79	0.7/ 0.76	0.76/ 0.81	0.70/ 0.77
	Cancer	0.69/ 0.77	0.34/ 0.67	0.72/ 0.73	0.878/ 0.88

		Algorithm			
		LR	k-NN	SVM	Stacking
Data	Alzheimer	0.84/ 0.84	0.67/ 0.63	0.74/ 0.76	0.84/ 0.85
	Diabetes	0.83/ 0.81	0.83/ 0.79	0.84/ 0.82	0.88/ 0.87
	Cancer	0.82/ 0.81	0.69/ 0.73	0.75/ 0.77	0.85/ 0.86

To visualize the datasets at each stage, the nonlinear dimensionality reduction algorithm UMAP [17] was applied. The results are presented in Figure 2 and show the data distribution at various stages of processing.

The results of binary classification models were further evaluated using ROC curves, which allowed for an analysis of the sensitivity and specificity of the algorithms before and after class balancing. The ROC curve visualization is presented in Figure 3, where it can be observed that significant improvements in performance were achieved after applying the class balancing method compared to the original results. This clearly demonstrates the effectiveness of the proposed class balancing approach and its positive impact on model performance.



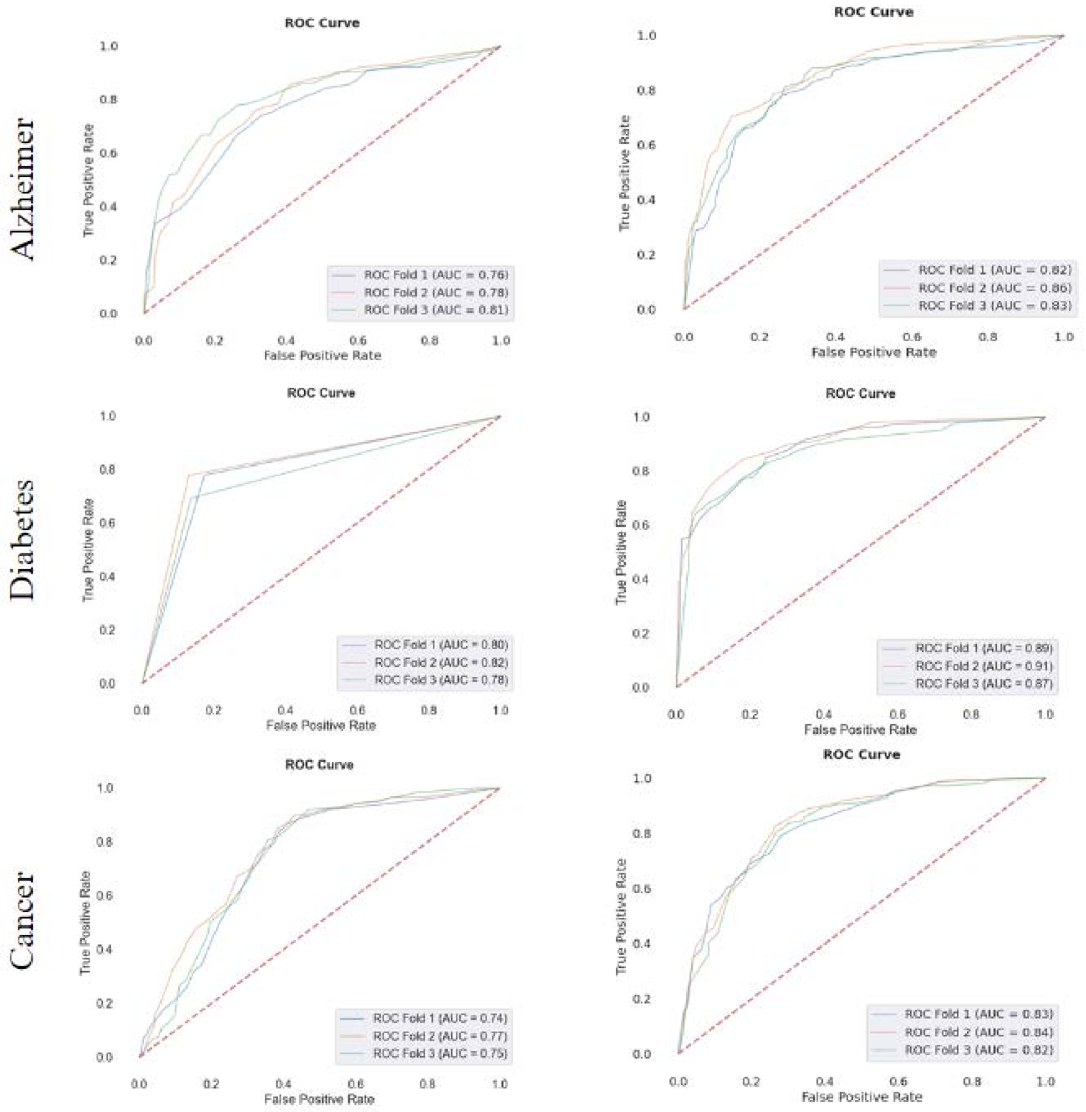
a) Original Dataset

b) Dataset After Undersampling Method

c) Minority Class Data After Oversampling Method

d) Dataset After Hybrid Class Balancing Method

Figure 2 – Visualization of Datasets Before and After the Hybrid Class Balancing Method



a) Results Before Hybrid Class Balancing

b) Results After Hybrid Class Balancing

Figure 3 – Visualization of ROC Curve Before and After Hybrid Class Balancing

4. Discussion

The proposed hybrid algorithm effectively addresses class imbalance by selectively removing excessive instances from majority classes. Genetic oversampling further enhances performance by synthesizing representative instances for minority classes. Experiments have demonstrated the superiority of our approach over traditional methods. Feature selection significantly impacts the results, making the search for optimal feature selection strategies a promising research direction. Additionally, exploring combinations of undersampling and oversampling methods presents an interesting area for further study.

5. Conclusions

This work presents a novel hybrid class-balancing algorithm that successfully addresses the issue of imbalance in medical data. Experimental results confirm its significance in enhancing the effectiveness of machine learning models on imbalanced datasets. The combination of the

IQRBU method and genetic oversampling represents a powerful tool for handling class imbalance in various medical applications. The results of experiments conducted on datasets related to Alzheimer's disease, diabetes, and cancer demonstrate that the proposed approach not only effectively mitigates imbalance but also significantly improves classification metrics, such as the F1-score and accuracy. For instance, in the case of Alzheimer's data, the F1-score increased from 0 to 74 when using the SVM method, indicating substantial progress. UMAP visualization clearly illustrates the effectiveness of the hybrid approach in class separation and reducing overlap, while ROC curves confirm the improved classification of models trained on balanced data. Despite these promising results, there is potential for further research, including the exploration of various feature selection methods and the combination of undersampling and oversampling approaches. Thus, this work makes a significant contribution to the field of machine learning for medical data by offering an effective solution to the class imbalance problem and opening new avenues for future research.

Author Contributions

Conceptualization, Z.B.; Methodology, A.Y.; Software, S.S. and R.I.; Validation, S.S.; Formal Analysis, Z.B. and A.Y.; Investigation, Z.B., A.Y., S.S., R.I. and Z.Z.; Resources, Z.B.; Writing – Original Draft Preparation, A.Y., S.S., R.I. and Z.Z.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Austin, C., & Kusumoto, F. (2016). The application of Big Data in medicine: current implications and future directions. Journal of Interventional Cardiac Electrophysiology, 47, 51-59.
2. Scopus – Processing of Medical Data'. Retrieved 30 July 2024 (https://www.scopus.com/results/results.uri?sort=plf-f&src=s&st1=processing+medical+data&sid=25c570f1a0b175e3d0ebf331d4dfd979&sot=b&sdt=cl&sl=38&s=TITLE-ABS-KEY%28processing+medical+data%29&origin=resultslist&editSaveSearch=&yearFrom=2000&yearTo=2024&sessionSearchId=25c570f1a0b175e3d0ebf331d4dfd979&limit=10).
3. Stewart, C. (2023). AI in healthcare market size worldwide 2021-2030.
4. Koziarski, M. (2020). Radial-based undersampling for imbalanced data classification. Pattern Recognition, 102, 107262.
5. 'Dementia'. Retrieved 30 July 2024 (https://www.who.int/news-room/fact-sheets/detail/dementia).
6. 'Cancer'. Retrieved 30 July 2024 (https://www.who.int/news-room/fact-sheets/detail/cancer).
7. 'Diabetes'. Retrieved 19 March 2024 (https://www.who.int/health-topics/diabetes).
8. 'Rabie El Kharoua | Master'. Retrieved 30 July 2024 (https://www.kaggle.com/rabieelkharoua/competitions).
9. 'PCA Documentation! — Pca Pca Documentation'. Retrieved 30 July 2024 (https://erdogant.github.io/pca/pages/html/index.html).
10. 'Radial basis function', *Wikipedia*. Jun. 12, 2024. Accessed: Jul. 30, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Radial_basis_function&oldid=1228720002
11. Albadr, M. A., Tiun, S., Ayob, M., & Al-Dhief, F. (2020). Genetic algorithm based on natural selection theory for optimization problems. *Symmetry*, *12*(11), 1758.
12. Buribayev, Z., Shaikalamova, S., Yerkos, A., & Imanbek, R. (2024). EKMGS: A HYBRID CLASS BALANCING METHOD FOR MEDICAL DATA PROCESSING. Scientific Journal of Astana IT University, 5-16.
13. 'Deap.Tools.Selection — DEAP 1.4.1 Documentation'. Retrieved 10 August 2024 (https://deap.readthedocs.io/en/master/_modules/deap/tools/selection.html#selBest)
14. Xue, Y., Zhu, H., Liang, J., & Słowik, A. (2021). Adaptive crossover operator based multi-objective binary genetic algorithm for feature selection in classification. Knowledge-Based Systems, 227, 107218.
15. Rajakumar, B. R. (2013). Static and adaptive mutation techniques for genetic algorithm: a systematic comparative analysis. International Journal of Computational Science and Engineering, 8(2), 180-193.
16. 'DEAP Documentation — DEAP 1.4.1 Documentation'. Retrieved 10 August 2024 (https://deap.readthedocs.io/en/master/index.html).
17. 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — Umap 0.5 Documentation'. Retrieved 30 July 2024 (https://umap-learn.readthedocs.io/en/latest/).

Information About Authors

Zholdas Buribayev is a PhD, Computer Science department at Al-Farabi Kazakh National University (Almaty, Kazakhstan, zhburibaev@gmail.com). His research interests include the development of class balancing algorithms in data processing. ORCID ID: 0000-0002-3486-227X.

Yerkos Ainur is a PhD student of Computer Science department at Al-Farabi Kazakh National University (Almaty, Kazakhstan, yerkosova@gmail.com). Her research interests include the development of class balancing algorithms in data processing. ORCID iD: 0000-0001-5949-6942

Shaikalamova Saida is a Bachelor of Information and Communication Technology of Computer Science department at Al-Farabi Kazakh National University (Almaty, Kazakhstan, shaikalamova02@gmail.com). Her research interests include the development of class balancing algorithms in data processing. ORCID iD: 0009-0002-9966-508X

Imanbek Rustem is a Master of Engineering Science of Computer Science department at Al-Farabi Kazakh National University (Almaty, Kazakhstan, imanbek.rustem2000@gmail.com). His research interests include the development of class balancing algorithms in data processing. ORCID iD: 0009-0008-7261-4382

Zhibek Zhetpisbay is a bachelor student of Computer Science department at Al-Farabi Kazakh National University (Almaty, Kazakhstan, av88276@gmail.com). Her research interests include the development of class balancing algorithms in data processing. ORCID iD: 0009-0005-7807-1444

Submission received: 19 August, 2024.
Revised: 5 September, 2024.
Accepted: 5 September, 2024.