Dataset balancing in disease prediction

Keywords: machine learning; data analysis; health informatics

Abstract: The use of machine learning in the prevention of serious diseases such as cancer or heart disease os getting more and more important. Various studies shows that improved forecasting performance can significantly increase patients' life expectancy. Of course, it is important to have adequate data sets for the use of techniques to classify the patient's clinical situation, for example, to classify the clinical situation of the patient for the prediction of the onset of disease. usually the available dataset are often not balanced, for example, there are more records with positive metric rather than negative, therefore the pre-processing of data plays a very important role. In this paper, using a given set of examples, we aim at evaluating the impact of the machine learning and smoting methods on the prediction performance. We will also show how selecting an appropriate SMOTE process the performance obtained, compared to many metrics, turns out to be much better. However, in some cases depending on the dataset and machine learning methods or the effect of smoothing is barely appreciable.

1 INTRODUCTION

The importance of machine learning (ML) in healthcare is increasingly evident and significant. ML models can analyze large amounts of data, such as medical images, vital signs, and medical histories, to assist physicians in the early and accurate diagnosis of diseases. This can lead to better outcomes for patients, as it allows for the timely and precise identification of conditions. Furthermore, through the analysis of patient data, advanced customization of treatments is possible. Indeed, ML can help develop personalized treatment plans, taking into account individual variations in biological data, test results, and treatment responses, thereby improving the effectiveness of care. Finally, machine learning plays a central role in disease prevention. ML models can identify risk factors for specific medical conditions and help prevent diseases through the early detection of predictive signs and the implementation of preventive interventions (Qayyum et al., 2021).

Since medical datasets are often imbalanced, the problem of dataset balancing is a crucial aspect when addressing disease prediction using machine learning models. A balanced dataset is essential to ensure that the model does not develop biases towards a particular class due to the prevalence of that class in the data.

Dataset balancing is significant in the medical sector for several critical reasons, with the primary concern being patient safety. Improper balancing in medical treatments can jeopardize patients' health and safety. For instance, excessive doses of medication can cause severe side effects or even be lethal, while insufficient doses might not have the desired therapeutic effect. Correct balancing of doses and treatments is essential to ensure that patients receive the right therapy for their condition. Proper balancing can guarantee treatment effectiveness and contribute to faster and more complete healing. Moreover, minimization of side effects is another crucial benefit. Proper attention to balancing can help reduce the side effects of treatments. For example, choosing the correct dosage can minimize the adverse effects of medications, thus improving patients' quality of life. Last but certainly not least, the reduction of healthcare costs and prevention of drug resistance are relevant matters. Accurate balancing can help reduce healthcare costs associated with ineffective or inappropriate treatments and care. Using only the necessary resources can optimize the allocation of financial resources in the medical sector, and improper use of medications, including administering unbalanced doses, can lead to drug resistance. Proper balancing of treatments can help prevent resistance and maintain the effectiveness of drugs over time. Overall, proper balancing in the medical sector is crucial to ensure positive outcomes for patients, reduce side effects, optimize resource utilization, and preserve treatment effectiveness in the long term. For this reason, the tests conducted in this article will focus on datasets related to the medical sector.

There are many characteristics regarding Dataset Balancing that become significant, such as Disease Prevalence, particularly when studying the class of rare diseases which might be underrepresented. To overcome the balancing problem, Synthetic Data

Generation can be utilized. Techniques like generating synthetic data, for example using augmentation algorithms, can be applied to increase the number of examples in underrepresented classes. In such cases, another aspect of interest concerns stratification, ensuring that the dataset is stratified, meaning that the proportion of target classes should be maintained uniformly in both the training and testing sets.

More general, the main issues Associated with Dataset Balancing are:

- Classification Bias: In an imbalanced dataset, models may tend to predict the majority class at the expense of the minority class, causing classification bias. The main bias concerning with:
- Inaccurate Performance: If a class is under represented, the model may not effectively learn from those examples, leading to inaccurate performance for that class.
- Overfitting: In the presence of underrepresented classes, the model may overfit to the training examples but struggle to generalize accurately to new data.

Balancing the dataset is critical in building disease prediction models, as it directly influences the model's ability to generalize correctly and make accurate predictions for all disease classes.

Existing literature is considered in Section 2, while the datasets are introduced in Section 3. In Section 4, the methods and results are discussed in detail. Moreover, a comparative study is presented to point out the appropriateness of the results with respect to several metrics. We finally consider further works and concluding remarks in Section 5.

2 DATA SET BALANCING

As commonly acknowledged, there are numerous methods for balancing a dataset. The goal is to create a balanced dataset that allows the model to learn fairly from the different classes present in the data. In this section, we discuss balancing methods for classification and provide an overview of related work in the literature on methods to balance datasets, particularly focusing on health-related research.

The problem of data imbalance is particularly relevant in health-related datasets. In many scenarios, the classes of interest, such as those related to rare diseases or clinically significant events, can be significantly underrepresented compared to control or normal classes. This is challenging specially during the training of machine learning models, as models tend to be influenced more by the majority class, thereby overlooking the minority class. Consequently, the model's ability to generalize to new data and correctly identify positive cases in the minority class may be compromised. Therefore, it is crucial to carefully address the issue of data imbalance in health-related datasets to ensure the construction of accurate and reliable models.

2.1 Data Balancing Solution for Classification

There are numerous techniques for balancing datasets. In this article, we will show the advantages of SMOTE (Pradipta et al., 2021)

SMOTE (Synthetic Minority Over-sampling Technique) is a technique used to address the problem of class imbalance in a dataset, particularly when there is a significantly less represented minority class compared to others. This technique is commonly applied in machine learning contexts, including classification models used to predict diseases, frauds, or other rare events. SMOTE operates based on three main items: Minority Definition, Generation of Synthetic Examples and the actual SMOTE Procedure.

1. Minority Definition: Identify the minority class in the dataset, which is the class that has fewer examples compared to the other classes
2. Generation of Synthetic Examples: SMOTE operates by generating synthetic examples of the minority class. These examples are obtained by linearly combining nearby samples in the feature space
3. SMOTE Procedure: For each example in the minority class, SMOTE randomly selects some of its nearest neighbors and creates new synthetic examples through linear combination of the feature values

.

Finally, by adding these synthetic examples to the dataset, SMOTE increases the amount of data available for the minority class, thus helping to balance the dataset.

In addition to SMOTE, there are several other methods to address the issue of class imbalance in datasets. Among them, we mention the following methods, the table1 reports a comparison amomg them:

- Random Undersampling: Randomly removes some examples from the majority class to balance the number of examples between classes. However, this approach can lead to a loss of valuable information.

- Random Oversampling: Randomly replicates some examples from the minority class to increase the number of examples in that class. Although it can be effective, it might lead to overfitting.
- Tomek Links: Removes the so-called "Tomek links," pairs of examples (one from the majority class and one from the minority class) that are very close to each other. This can help improve the separation between classes. item ENN (Edited Nearest Neighbors): Eliminates examples from the majority class that are close to examples from the minority class. It can reduce noise in the dataset.
- SMOTE-ENN: A combination of SMOTE and ENN, which applies SMOTE to generate synthetic examples in the minority class and then uses ENN to remove examples from the majority class that might be misclassified.
- ADASYN (Adaptive Synthetic Sampling): Similar to SMOTE, but assigns different weights to different examples, allowing for a more focused synthetic generation on the minority class.
- ROS (Random Oversampling with Replacement): Randomly replicates examples from the minority class, allowing for the possibility of replicating the same example multiple times.
- Cluster-Based Over-sampling: Groups examples from the minority class into clusters and applies oversampling to these clusters.
- Cost-Sensitive Learning: Assigns different weights to classes during model training, giving more importance to errors on the minority class.

The choice of method depends on the specific characteristics of the dataset and the problem being addressed. In some cases, experimenting with different approaches may be effective in determining which one works best for the specific case. Each technique has its advantages and disadvantages, and the choice depends on the specific characteristics of the dataset and the problem being addressed. Experimentation and validation are crucial to determine the most suitable approach for a particular scenario.

2.2 Related works

Several papers propose various approaches to address class imbalance and feature selection problems in Clinical Decision Support Systems (CDSS). In (Sreejith et al., 2020) the authors introduce a framework that balances the dataset at the data level and employs a wrapper approach for feature selection, utilizing Chaotic Multi-Verse Optimization (CMVO) for subset selection. Performance evaluation using the arithmetic mean of Matthews correlation coefficient (MCC) and F-score (F1) indicates competitiveness of the proposed framework. Paper (Xu et al., 2021) presents a cluster-based oversampling algorithm (KNSMOTE), which combines Synthetic Minority Oversampling Technique (SMOTE) and k-means clustering. This algorithm identifies "safe samples" from clustered classes and synthesizes new samples through linear interpolation, effectively addressing class imbalance.

In a different study, (Li et al., 2021), (Xu et al., 2020) SMOTE is highlighted as a successful method with practical applications, alongside the introduction of a novel oversampling approach called SMOTE-NaN-DE, which improves class-imbalance data by generating synthetic samples. Additionally, a hybrid sampling algorithm named RFMSE, combining M-SMOTE and Edited Nearest Neighbor based on Random Forest, is proposed to enhance sampling effectiveness. Uhe authors in (Jakhmola and Pradhan, 2015) propose an interactive preprocessing algorithm allowing users to customize preprocessing requirements, yielding higher quality data suitable for correlation and multiple regression analysis, as demonstrated on a diabetes dataset.

Finally, in (Khushi et al., 2021) are introduced research investigates class imbalance techniques for lung cancer prediction, employing various methods including under-sampling, over-sampling, and hybrid techniques. Evaluation metrics such as AUC reveal the superiority of over-sampling methods, particularly random forest with random over-sampling, in predicting lung cancer presence.

.

3 DATASET DESCRIPTION

To analyze the issue of balance in the healthcare domain, we will use five different datasets. They share the common denominator of being related to the prediction of medical situations, and involve binary classification.

Despite their differences, all these datasets exhibit imbalance, so we will attempt to analyze how the prediction performs when using the imbalanced dataset, as well as when performing preprocessing for preliminary dataset balancing.

The five datasets, although all involve binary classification, are very different in terms of the number of features, number of observations, and imbalance Ratio.

We will define Imbalance Ratio as the propor-

Method	Advantages	Disadvantages
SMOTE	Preserves information from the minority class, reducing the risk of data loss. Can improve the generalization of the model.	May introduce noise in the synthetic data, especially if the data distribution is complex. Could require more computational time compared to other methods.
Random Undersampling	Simple and fast to implement. Can reduce the training time on very large datasets.	May lead to loss of important information in the majority class, increasing the risk of under representation.
Random Oversampling	Simple to implement. Can improve the accuracy of models on imbalanced datasets.	May lead to overfitting if not used cautiously, especially with excessive replication.
Cluster Based Oversampling	Effective when minority class examples form distinct clusters. Reduces the risk of generating synthetic data in inconsistent regions.	Requires careful parameter tuning and can be computationally expensive.
Tomek Links	Enhances class separation without adding noise.	May not be effective in complex class distributions.
ENN	Can improve model performance by reducing misclassification.	May excessively reduce dataset size, potentially losing important information.
SMOTE-ENN	- Combines benefits of both techniques, enhancing class separation and mitigating overfitting risks.	Computationally intensive, particularly on large datasets.
ADASYN	More effective in complex and non-uniform data distributions.	Requires more computational resources compared to SMOTE.
Random Oversampling with Replacement	Simple to implement. Can enhance model performance on imbalanced datasets.	Risk of overfitting if replication is excessive, especially on small datasets.
Cost-Sensitive Learning	Improves model performance on imbalanced datasets without synthetic data addition.	Requires careful weight selection and may not be universally effective.

Table 1: Comparison of Different Imbalanced Data Handling Methods

tion between the number of examples in the minority class and the number of examples in the majority class. For example, if there are 100 negative examples and 20 positive examples, the imbalance ratio will be $20/100 = 0.2$. Obviously, the more imbalanced the dataset, the closer this value to zero.

Some of The datasets used are very popular whilst others that are less well-known. In particular, we will use the following:

- Breast Cancer Wisconsin Diagnostic (Breast_cancer (Repository,)
- Hearth failure (Chicco and Jurman, 2020a)
- Pima Indians Diabetes Database (Pina)(Sigillito,)
- Differentiated Thyroid Cancer Recurrence dataset (Thyroid) (Borzooei et al., 2023)
- Sepsis Survival Minimal Clinical Records (Sepsis) (Chicco and Jurman, 2020b)

In this section, we analyze the five datasets from several perspectives in order to asses the balancing effect.

3.1 Datasets overview

In this section, we present the datasets that we will use for testing. Some of them are very popular datasets, while others are less so. These datasets vary in size and degree of imbalance. What they have in common is that they are based on binary classification.

The first dataset, Wisconsin Diagnostic Breast Cancer (WDBC), is the well-known dataset that (Repository,) collects data for breast cancer prediction. Since breast cancer is the most common cause of cancer deaths in women and is a type of cancer that can be treated when diagnosed early, prediction is a very important aspect. This dataset (Elter et al.,) has been extensively studied in the literature, which is

why it is utilized in this paper. The dataset is from the University Hospital of California and can be downloaded from both the UCI Machine Learning Repository and Kaggle. It consists of 569 samples and 33 features, computed from a digitized image of a fine needle aspiration (FNA) of a breast mass and related to some characteristics of each cell nucleus (e.g., radius, texture, perimeter, area, etc.). Some of these features are more selective and decisive than others, and the determination of these features significantly increases the success of the models, which is why Feature Selection is applied to select them.

The second dataset, also widely referenced in literature, is the Heart Failure Clinical Records dataset (Chicco and Jurman, 2020a) Cardiovascular diseases (CVDs) are the leading cause of death globally, claiming approximately 17.9 million lives each year, representing 31% of all deaths worldwide. Heart failure, a common occurrence resulting from CVDs, is the focus of this dataset, which comprises 12 features aimed at predicting mortality associated with heart failure. Many CVDs are preventable through addressing behavioral risk factors such as tobacco use, poor diet, obesity, physical inactivity, and excessive alcohol consumption via population-wide interventions. Individuals with existing CVD or those at high cardiovascular risk, often due to hypertension, diabetes, hyperlipidemia, or other established diseases, require early detection and management, where machine learning models can offer significant assistance. This dataset includes medical records from 299 heart failure patients, gathered at the Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad, Punjab, Pakistan, between April and December 2015. It encompasses 13 features encompassing clinical, physiological, and lifestyle-related information.

The third dataset used is Pima Indians Diabetes Database (Sigillito,). The Pima Indians Diabetes Database is a well-known dataset in the field of machine learning and healthcare research. It contains medical data from the Pima Indian population, specifically focused on women aged 21 and above from the Gila River Indian Community near Phoenix, Arizona. The dataset includes various health-related attributes such as glucose level, insulin level, BMI (Body Mass Index), age, and the presence or absence of diabetes within a five-year period following the initial examination. This dataset is widely used for developing predictive models to identify individuals at risk of developing diabetes. Due to its large sample size and comprehensive health information, the Pima Indians Diabetes Database has been instrumental in advancing research in diabetes prediction and management. Despite its significance, the dataset also poses challenges due to its inherent class imbalance and missing data, necessitating careful preprocessing and model evaluation techniques. Its availability in the public domain has facilitated numerous studies aimed at improving diabetes diagnosis and treatment strategies, contributing significantly to the broader efforts in public health and medical informatics.

The fourth dataset is a more recent data set. The Differentiated Thyroid Cancer Recurrence dataset is a valuable resource in the domain of thyroid cancer research. It comprises clinical data from patients diagnosed with differentiated thyroid cancer (DTC) who underwent thyroidectomy and subsequent treatment. The dataset includes various demographic and clinical variables such as age, sex, tumor size, histopathological characteristics, treatment modalities, and follow-up information. A key focus of the dataset is to predict the recurrence of thyroid cancer following initial treatment based on these factors. Researchers utilize machine learning and statistical methods to develop predictive models that can identify patients at higher risk of recurrence, thereby aiding in personalized treatment strategies and follow-up care. Due to its specialized nature and importance in thyroid cancer management, the Differentiated Thyroid Cancer Recurrence dataset has garnered attention from researchers worldwide. However, challenges such as limited sample size and data heterogeneity need to be addressed to enhance the robustness and generalizability of predictive models derived from this dataset. Overall, it serves as a valuable tool in advancing our understanding of thyroid cancer recurrence and improving patient outcomes through tailored interventions.

Finally the last dataset used is the Sepsis Survival Minimal Clinical Records. The Sepsis Survival Minimal Clinical Records dataset is an essential and widely used dataset in sepsis study and research, a severe medical condition caused by a systemic inflammatory response to an infection. This dataset contains clinically relevant and minimal information about patients with sepsis, including demographic data, vital signs, laboratory test results, and treatment information. Its simplified structure makes it particularly suitable for developing predictive models of sepsis survival and for evaluating clinical management strategies. Thanks to its availability and focused nature, the Sepsis Survival Minimal Clinical Records dataset has significantly contributed to the understanding of sepsis and the research of effective clinical interventions to improve outcomes for patients with this severe medical condition. However, it is important to consider limitations and potential biases in the data to obtain accurate and generalizable results.

The common feature they share is the presence of only two classes. The main data of the 5 datasets are summarized in the table2 that shows that they exhibit different numbers of features and observations and varying Imbalance Ratios.

3.2 Methods

Supervised Learning is a machine learning technique in which the algorithms are designed to learn by example. In fact, the training data consist of inputs paired with the correct outputs. During training the algorithm will search for patterns in the data that correlate with the desired outputs. After training the algorithm will take in new unseen inputs and will determine which label the new inputs will be classified as, based on prior training data. The objective of a supervised learning model is to predict the correct label for newly presented input data.

We are going to implement the following 10 supervised algorithms:

1. Logistic Regression (**LG**)
2. Support Vector Machine (**SVM**)
3. Gaussian Naive Bayes (**GNB**.)
4. Decision Tree (**DT**)
5. Random Forest (**RF**)
6. Extra Tree (**ET**)
7. K-Nearest Neighbors (**KNN**)
8. Hist Gradient Boosting (**HGB**)
9. Bagging Clsssifier (**BC**)
10. Multilayer Perceptron (**MLP**)

The choice of ten different methods aims to conduct experiments that demonstrate the impact of different types of smoothing in various ways. This will ultimately allow us to assert that it is difficult to claim that there is generally a better smoothing method. As we will see, the same smoothing method that proves to be the best in some cases may even yield worse results than the unbalanced dataset in other cases.

Logistic regression is a machine learning method used for binary classification problems. Its principle of operation is based on estimating the conditional probabilities that an instance belongs to one of the two classes. It uses the logistic function (or sigmoid function) to transform a linear combination of features into a value between 0 and 1, representing the estimated probability. This value is then used as a threshold to assign the instance to one of the two classes. The advantages of logistic regression include: Interpretability and Scalability.

Support Vector Machine operates by seeking to find the optimal hyperplane of separation between classes in the case of binary classification. The separation hyperplane is defined as the hyperplane that maximizes the margin between the nearest class instances, which are called support vectors. SVM can effectively handle datasets with a large number of features and it tends to generalize well on test data, reducing the risk of overfitting.

Gaussian Naive Bayes is based on Bayes' theorem and assumes that features are independent and follow a Gaussian distribution. It's simple, computationally efficient, and works well with high-dimensional data.

Decision Tree recursively splits the dataset into subsets based on the value of features, aiming to maximize the purity of each subset in terms of class labels. It's interpretable and can handle both numerical and categorical data.

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions through voting or averaging. It reduces overfitting and improves accuracy compared to individual decision trees.

Extra Trees, or Extremely Randomized Trees, is similar to Random Forest but introduces additional randomness in the feature selection process. It can be faster to train but may require more trees to achieve similar performance as Random Forest. K-Nearest Neighbors (KNN):

K-Nearest Neighbor, operates by classifying an instance based on the majority class among its k nearest neighbors in the feature space. The distance metric (e.g., Euclidean distance) is used to measure the similarity between instances. The advantages of KNN include: Simplicity, No training phase and Adaptability

Hist Gradient Boosting is a boosting algorithm that builds a series of decision trees sequentially, each one correcting the errors of its predecessors. It uses histogram-based techniques to speed up training.

Bagging, or Bootstrap Aggregating, is an ensemble learning method that trains multiple models on bootstrap samples of the dataset and combines their predictions. It reduces variance and improves stability.

Multilayer Perceptron is a type of artificial neural network consisting of multiple layers of interconnected neurons. It's capable of learning complex patterns and relationships in the data, making it suitable for various tasks including classification and regression.

Dataset	#Instances	#Features	Imbalance Ratio	Feature Types	
				#Numeric	#Symbolic
Breast Cancer Wisconsin (Diagnostic)	699	9	0.59	9	0
Hearth failure	299	12	0.94	12	0
Pima Indians Diabetes Database	768	8	0.54	8	0
Differentiated Thyroid Cancer Recurrence dataset	383	16	0.39	6	10
Sepsis Survival Minimal Clinical Records	137	3	0.21	3	0

Table 2: Dataset Information

4 EXPRERIMENT AND DISCUSSION

To assess the impact of balancing, we will employ various evaluation metrics and different smooting methods.

Specifically, we will utilize the following metrics:

- Accuracy is a general measure of the model's precision and represents the percentage of instances classified correctly out of the total instances. It's calculated as the ratio of the number of correct predictions to the total number of predictions made. Accuracy is particularly useful when classes in the dataset are balanced but can be misleading if the dataset is imbalanced. Area Under the ROC Curve (AUC)
- AUC measures the model's discriminative ability, i.e., its ability to correctly classify positive examples as positive and negative examples as negative. The ROC (Receiver Operating Characteristic) curve represents the true positive rate (TPR) versus the false positive rate (FPR) as the classification threshold varies. AUC is the area under the ROC curve and provides an overall assessment of the model's performance, where a value closer to 1 indicates better discriminative ability.
- Precision measures the fraction of positive instances correctly classified out of the total instances classified as positive. It's calculated as the ratio of the number of true positives to the sum of true positives and false positives. Precision is useful when the goal is to maximize the number of positive instances correctly identified while minimizing the number of false positives.
- Recall measures the fraction of positive instances correctly classified out of the total positive instances in the dataset. It's calculated as the ratio of the number of true positives to the sum of true positives and false negatives. Recall is particularly useful when the goal is to identify all positive instances in the dataset, minimizing the number of false negatives.

We will use four different methods for oversampling, commonly used in dealing with imbalanced datasets.

1. BorderlineSMOTE with 20 neighbors=20) (Smote1)
2. BorderlineSMOTE with neighbors=10 and $sampling\_strategy =' minority'$ (Smote2)
3. SMOTE with 10 neighbors (Smote3)
4. SMOTE with $sampling\_strategy =' minority'$ and neighbors=10(Smote4)

BorderlineSMOTE is a variant of the SMOTE algorithm that generates synthetic samples near the decision boundary (borderline) between the minority and majority classes. It generates synthetic samples by selecting a minority class instance and finding its k nearest neighbors. It then selects one of these neighbors randomly and generates a synthetic sample along the line segment joining the original instance and the selected neighbor. By setting m_neighbors=20, it specifies the number of nearest neighbors to consider when generating synthetic samples.

The second method is a variant of BorderlineSMOTE also generates synthetic samples near the decision boundary between the minority and majority classes. Additionally, it adjusts the sampling strategy to focus on the minority class by specifying sampling_strategy='minority'. By setting m_neighbors=10, it specifies a different number of nearest neighbors to consider when generating synthetic samples compared to the previous variant.

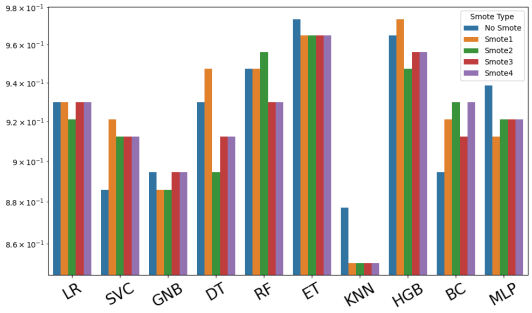
SMOTE (Synthetic Minority Over-sampling Technique) is a popular oversampling technique that generates synthetic samples by interpolating between existing minority class instances. It selects a minority class instance and finds its k nearest neighbors. It then selects one of these neighbors randomly and generates a synthetic sample along the line segment joining the original instance and the selected neighbor. By setting k_neighbors=10, it specifies the number of nearest neighbors to consider when generating synthetic samples.

The fourth method, similar to the third, this variant of SMOTE also generates synthetic samples by interpolating between existing minority class instances. It further adjusts the sampling

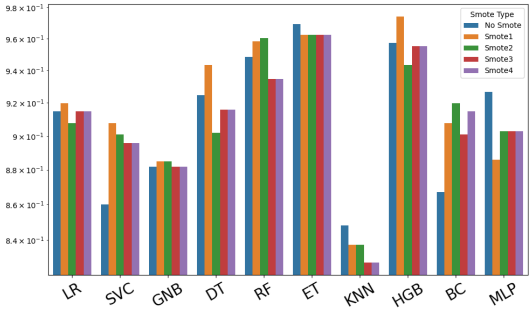
strategy to focus on the minority class by specifying sampling_strategy='minority'. By setting k_neighbors=10, it specifies a different number of nearest neighbors to consider when generating synthetic samples compared to the previous variant.

In general, BorderlineSMOTE focuses on generating samples near the decision boundary, while SMOTE generates samples by interpolating between existing instances. The number of nearest neighbors considered (m_neighbors or k_neighbors) may vary between the methods, affecting the characteristics of the synthetic samples generated. Similarities: Both BorderlineSMOTE and SMOTE aim to address class imbalance by oversampling the minority class. They both use a similar mechanism to generate synthetic samples based on nearest neighbors.

Above we will show the performance of each dataset in term of accuracy anf Auc. The accuracy and the Auc score are shown on a logarithmic scale as the machine learning methods vary across the five datasets. The blue column refers to the analysis of the imbalanced dataset, while the other four refer to datasets obtained with the four balancing methods. It can be observed that while some datasets yield consistent accuracy values across methods, others exhibit significant variability depending on the method used.



(a) Accuracy

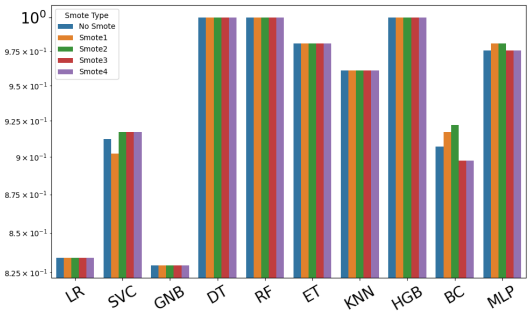


(b) AUC

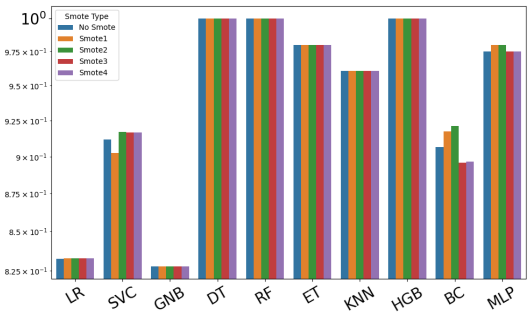
Figure 1: Breast_cancer scores

The result of Breast_Cancer accuracy and Auc score are shown in figure 1. For the Breast_Cancer dataset, the highest accuracy (see figure 1a) values are achieved with the ET and HGB methods, whether smoothing is applied or not. The maximum score of 0.973684 is obtained for ET when no balancing is performed and for HGB when Smote1 is applied. The maximun Auc score (see figure 1b) it obtained (0.974206) for HGB with Smote1. This, considering that AUC is less prone to overfitting, allows us to affirm that smoothing allows us to gain an advantage, albeit small

The result of *Heart_failure* accuracy and AUC score are shown in fig 2. The accuracy values reach 1 (see figure 2a) using various methods, but almost always with smoothing. It is possible to observe that this data set consistently performs quite well in terms of accuracy (the worst value being 0.829268). Analyzing the results for AUC (see figure 2b) allows us to make the same considerations and therefore its study is not particularly significant for our purposes.



(a) Accuracy

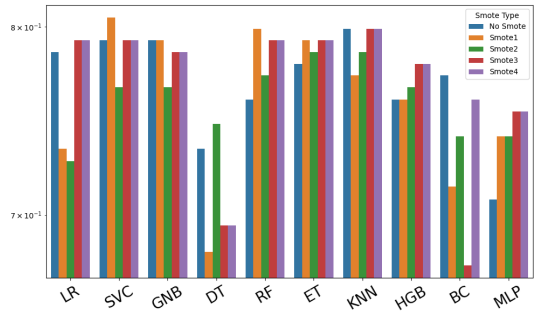


(b) AUC

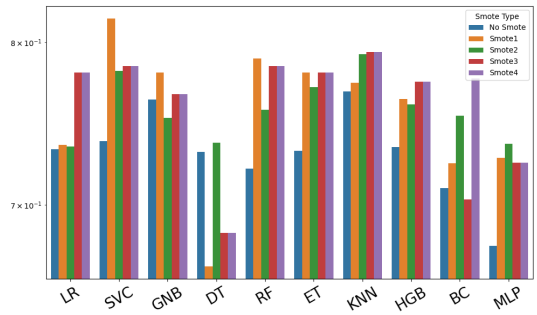
Figure 2: Heart_failure scores

For the third dataset (Pima), all methods perform better with appropriate balancing, both in terms of accuracy and AUC score 3. In this case, several methods with SMOTE allow achieving an accuracy of 0.892857 (3a). As you can see in the figure 3b the best AUC (DT or ET method) is consistently obtained with datasets that have had SMOTE2 applied (0.856522).

This allows us to assert that for this dataset with an imbalance ratio of 0.54, balancing is often beneficial.



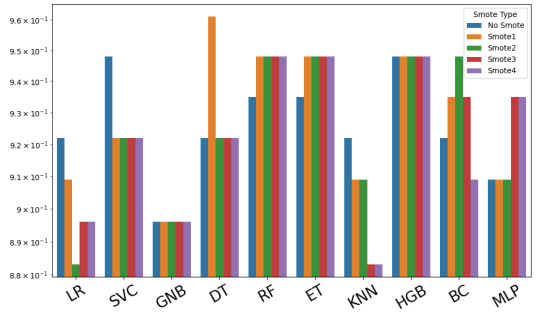
(a) Accuracy



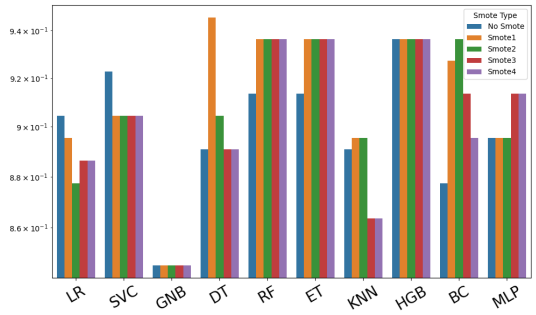
(b) AUC
Figure 3: Pima scores

For the fourth dataset (Differentiated Thyroid Cancer Recurrence dataset) 4, the best accuracy value (0.961039) is obtained with DT using Smote1 (see figure 4a. However, it can be observed that all methods tend to perform better with balancing. It is interesting to note that this dataset exhibits a higher imbalance ratio compared to the previous three, and therefore, the effect of smoothing is generally positive. The analysis of the AUC highlights more prominently the differences between the various methods but confirms which are the most performant solutions (see figure 4b. The best score for AUC is 0.945455 and is achieved with the DT method with Smote1.

Finally, for the fifth dataset (the most imbalanced), the behavior is nearly equivalent for any method, as adopting the appropriate smoothing method (not always the same) 5. The best choice allows achieving a value of 0.892857. A particular situation occurs for the MLP method, which performs poorly using any of the four balancing methods. In terms of AUC score, the results are still extremely diverse depending on the method and smoothing used, much like with accuracy. However, it should be noted that for this dataset, which is the most imbalanced one, balancing can be crucial as it allows us to achieve the best result. At the



(a) Accuracy



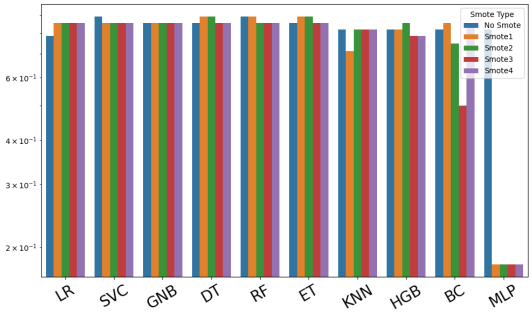
(b) AUC
Figure 4: Thyroid score

same time, if used inadequately, it can even worsen the results. In term of ACU score the best score is obtaine with DT and ET using Smote2.

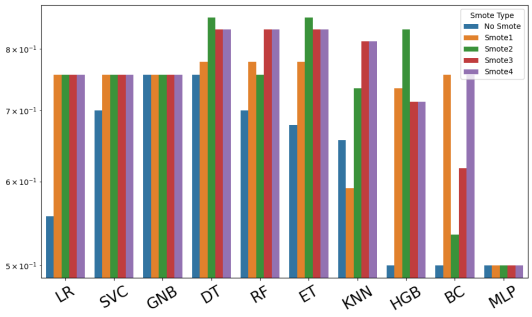
The positive effect of smoothing is better appreciated by analyzing AUC, which is less influenced by overfitting compared to accuracy. For datasets where the choice of method substantially alters accuracy values, the impact of data balancing can be significant. The table summarizes the minimum and maximum accuracy values for each dataset and the method yielding the best performance. Therefore, it can be stated that there is no single most effective balancing method, but rather, the ¡method, balancing¿ pair yielding better performance should be sought.

5 CONCLUSION

We conducted a detailed analysis on five different datasets, using a variety of machine learning methods that included classification, regression, and clustering models. For each dataset, we explored the behavior of ten distinct algorithms, each with its own characteristics and tuning parameters. In order to assess the impact of data balancing, we performed the analysis both with and without data balancing techniques, considering four different smoothing approaches to handle the presence of underrepresented classes. The re-



(a) Accuracy



(b) AUC

Figure 5: Sepsis scores

sults obtained highlighted a significant variation in model performance based on the different combinations of machine learning method, data balancing, and smoothing techniques.

It became clear that the choice of machine learning method and the application of balancing strategies must be closely integrated to achieve optimal results. In particular, we observed that while data balancing can significantly improve model performance on heavily imbalanced datasets, inadequate implementation could lead to inferior results. Furthermore, we recognized that parameter optimization for datasets characterized by imbalance requires a particularly careful and targeted approach, as the specific dataset characteristics can significantly influence the effectiveness of proposed solutions.

While various approaches exist, comparing them can be challenging due to numerous tuning parameters and variations within articles. However, ongoing research suggests the importance of exploring diverse classifiers and imbalance techniques, including deep learning models, to enhance prediction capabilities and address imbalance issues effectively. In conclusion, our analysis underscored the importance of carefully considering the specific context of each dataset and adopting a flexible and targeted approach to address the issue of data imbalance in machine learning contexts.

From the analysis conducted, it emerged that in the vast majority of cases, the solutions' accuracy and AUC with the application of balancing are better. Nonetheless, as future work, the analysis will be extended to a greater number of datasets and balancing methods. Another activity for future work concerns the application to datasets that involve non-binary classification to analyze whether balancing is advantageous in this case as well. Finally, precision and recall analysis could be conducted to add further confidence in the quality of the results.

ACKNOWLEDGEMENTS

The work is partially supported by UDMA project, CUP: G69J18001040007

REFERENCES

Borzooei, S., Briganti, G., and Golparian, M. e. a. (2023). Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. *European Archives of Oto-Rhino-Laryngology*.

Chicco, D. and Jurman, G. (2020a). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20(16).

Chicco, D. and Jurman, G. (2020b). Survival prediction of patients with sepsis from age, sex, and septic episode number alone. *Sci Rep*, 10:17156.

Elter, M., Schulz-Wendtland, R., and Wittenberg, T. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process.

Jakhmola, S. and Pradhan, T. (2015). A computational approach of data smoothening and prediction of diabetes dataset. In *Proceedings of the Third International Symposium on Women in Computing and Informatics*, WCI '15, page 744–748, New York, NY, USA. Association for Computing Machinery.

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., and Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9:109960–109975.

Li, J., Zhu, Q., Wu, Q., Zhang, Z., Gong, Y., He, Z., and Zhu, F. (2021). Smote-nan-de: Addressing the noisy and borderline examples problem in imbalanced classification by natural neighbors and differential evolution. *Knowledge-Based Systems*, 223:107056.

Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., and Ismail, M. (2021). Smote for handling imbalanced data problem : A review. In *2021 Sixth International Conference on Informatics and Computing (ICIC)*, pages 1–8.

Qayyum, A., Qadir, J., Bilal, M., and Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14:156–180.

Repository, U. M. L. UCI Machine Learning Repository: Mammographic Mass Data Set. `http://archive.ics.uci.edu/ml/datasets/mammographic+mass`.

Sigillito, V. National Institute of Diabetes and Digestive and Kidney Diseases, note = Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu) Research Center, RMI Group Leader Applied Physics Laboratory The Johns Hopkins University Johns Hopkins Road Laurel, MD 20707 (301) 953-6231 (c) Date received: 9 May 1990.

Sreejith, S., Khanna Nehemiah, H., and Kannan, A. (2020). Clinical data classification using an enhanced smote and chaotic evolutionary feature selection. *Computers in Biology and Medicine*, 126:103991.

Xu, Z., Shen, D., Nie, T., and Kou, Y. (2020). A hybrid sampling algorithm combining m-smote and enn based on random forest for medical imbalanced data. *Journal of Biomedical Informatics*, 107:103465.

Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N., and Han, X. (2021). A cluster-based oversampling algorithm combining smote and k-means for imbalanced medical data. *Information Sciences*, 572:574–589.