

Genetic algorithm-based approach to improve the prediction of myocardial infarction complications: optimization of class imbalance ratio

Enfoque basado en algoritmos genéticos para mejorar la predicción de las complicaciones del infarto de miocardio: optimización de la relación de desequilibrio entre clases

Abordagem baseada em algoritmo genético para melhorar a previsão das complicações do infarto do miocárdio: otimização da proporção de desequilíbrio entre classe

Maria Isabel Vidal Caicedo¹
Néstor Milciades Díaz Mariño²

Received: March 15th, 2026

Approved: June 18th, 2026

Available: September 1st, 2026

How to cite this article:

M. I. Vidal Caicedo and N. M. Díaz Mariño, "Genetic Algorithm-Based Approach to Improve the Prediction of Myocardial Infarction Complications: Optimization of Class Imbalance Ratio," *Revista Ingeniería Solidaria*, vol. 22, no. 3, 2026. doi: <https://doi.org/10.16925/2357-6014.2026.03.02>

Research article. <https://doi.org/10.16925/2357-6014.2026.03.02>

¹ Ph.D. in Computer Science, Faculty of Electronic Engineering and Telecommunications, Universidad del Cauca, Popayán, Colombia.

Email: mariaividal@unicauca.edu.co

ORCID: <https://orcid.org/0000-0002-5456-3082>

CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000900893

² Department of Systems, Faculty of Electronic Engineering and Telecommunications, Universidad del Cauca, Popayán, Colombia.

Email: nediaz@unicauca.edu.co

ORCID: <https://orcid.org/0000-0002-6074-9764>

CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000674435



Abstract

Introduction: This article is the result of research conducted as part of a doctoral thesis in the Computer Science Ph.D. program at the University of Cauca in 2025. Myocardial infarction (MI) is a condition associated with a high mortality rate during the first year following onset, and life-threatening complications can arise within the first few hours after hospital admission. Therefore, early prediction of these complications improves prognosis and reduces risks.

Problem: Machine learning algorithms predict potential complications; however, class imbalance in the datasets skews the predictions toward the majority class.

Objective: To improve the predictive performance of an **ML** model by rebalancing the class distribution in the training dataset using a data-driven approach.

Methodology: The CRISP-DM methodology was adapted.

Results: We propose a genetic algorithm (GA) to identify the optimal rebalancing ratio from the data. The best-performing individual determined the optimal rebalancing ratio through random oversampling (ROS) using an XGBoost model, which correctly classified 7 out of 8 cases (F1 score: 0.36), outperforming the baseline model, which correctly classified 6 out of 8 cases (F1 score: 0.28).

Conclusions: An approach that optimizes the class-balancing ratio improves the F1 score and detects an additional class compared to the base model without class balancing.

Originality: This article proposes a mechanism to find the optimal proportion with which an imbalanced dataset should be rebalanced.

Limitations: Processing is required when working with larger datasets.

Keywords: Imbalanced Learning, XGBoost, SMOTE, Random Oversampling (ROS), Genetic Algorithm.

Resumen

Introducción: Este artículo es el resultado de investigaciones realizadas en el marco de la tesis doctoral del programa de Doctorado en Ciencias de la Computación de la Universidad del Cauca en 2025. El infarto de miocardio (IM) es una afección de alta mortalidad que presenta complicaciones mortales en las primeras horas tras la hospitalización. La predicción temprana de estas complicaciones mejora el pronóstico y reduce los riesgos.

Problema: Los algoritmos de aprendizaje automático predicen posibles complicaciones; sin embargo, el desequilibrio de clases en los conjuntos de datos sesga las predicciones hacia la clase mayoritaria.

Objetivo: Mejorar el rendimiento predictivo de un modelo de aprendizaje automático reequilibrando la distribución de clases en el conjunto de datos de entrenamiento mediante un enfoque basado en datos.

Metodología: Se adaptó la metodología CRISP-DM.

Resultados: Proponemos un algoritmo genético (AG) para identificar la proporción óptima de reequilibrio a partir de los datos. El individuo con mejor rendimiento determinó la proporción óptima de reequilibrio mediante sobremuestreo aleatorio (ROS) utilizando un modelo XGBoost, que clasificó correctamente 7 de 8 casos (puntuación F1: 0,36), superando al modelo de referencia, que clasificó correctamente 6 de 8 casos (puntuación F1: 0,28).

Conclusiones: Un enfoque que optimiza la proporción de equilibrio de clases mejora la puntuación F1 y detecta una clase adicional en comparación con el modelo base sin equilibrio de clases.

Originalidad: Este artículo propone un mecanismo para determinar la proporción óptima con la que debe reequilibrarse un conjunto de datos desequilibrado.

Limitaciones: Se requiere procesamiento cuando se trabaja con conjuntos de datos más grandes.

Palabras clave: Aprendizaje desequilibrado, XGBoost, SMOTE, sobremuestreo aleatorio (ROS), algoritmo genético.

Resumo

Introdução: Este artigo é resultado de pesquisas realizadas no âmbito da tese de doutorado do programa de Doutorado em Ciências da Computação da Universidade do Cauca, em 2025. O infarto do miocárdio (IM) é uma condição associada a uma elevada mortalidade e pode apresentar complicações potencialmente fatais nas primeiras horas após a hospitalização. A previsão precoce dessas complicações melhora o prognóstico e reduz os riscos.

Problema: Os algoritmos de aprendizado de máquina podem prever possíveis complicações; entretanto, o desequilíbrio entre classes nos conjuntos de dados tende a direcionar as previsões para a classe majoritária.

Objetivo: Melhorar o desempenho preditivo de um modelo de aprendizado de máquina, reequilibrando a distribuição das classes no conjunto de dados de treinamento por meio de uma abordagem orientada por dados.

Metodologia: Foi adaptada a metodologia CRISP-DM.

Resultados: Propõe-se um algoritmo genético (AG) para identificar a proporção ideal de reequilíbrio a partir dos dados. O indivíduo de melhor desempenho determinou a proporção ótima de reequilíbrio por meio de sobreamostragem aleatória (ROS), utilizando um modelo XGBoost, que classificou corretamente 7 de 8 casos, com pontuação F1 de 0,36, superando o modelo de referência, que classificou corretamente 6 de 8 casos, com pontuação F1 de 0,28.

Conclusões: Uma abordagem que otimiza a proporção de equilíbrio entre classes melhora a pontuação F1 e permite detectar uma classe adicional em comparação com o modelo-base sem equilíbrio de classes.

Originalidade: O artigo propõe um mecanismo para determinar a proporção ótima com a qual um conjunto de dados desbalanceado deve ser reequilibrado.

Limitações: É necessário maior processamento quando se trabalha com conjuntos de dados maiores.

Palavras-chave: Aprendizado desbalanceado; XGBoost; SMOTE; sobreamostragem aleatória (ROS); algoritmo genético.

1. INTRODUCTION

Imbalanced learning is a current challenge in machine learning (ML). This problem is common in different areas such as fraud detection [1], text classification [2], and biomedical datasets [3]. Specifically, in the field of biomedicine, datasets are often imbalanced, with one class overrepresented and the minority class, usually the most clinically relevant, underrepresented. This phenomenon, known as imbalanced learning, can significantly compromise the performance of classification models, making it difficult to accurately detect less frequent but clinically critical cases [4]. Researchers have proposed various strategies to address class imbalance in datasets, which can

be categorized into three main approaches: data-level techniques, algorithm-level techniques, and hybrid techniques that combine elements from both [5].

Data-driven rebalancing techniques employ resampling methods, including oversampling—which increases the proportion of data from the minority class—and undersampling—which reduces the proportion of data from the majority class. In both cases, it is important to determine the frequency with which instances should be drawn. However, there is no consensus on the mechanism for determining the number of instances of each class required in the rebalanced dataset so that it is possible to learn from the minority classes without losing the relevant characteristics of the most frequent class [6]. The SMOTE (Synthetic Minority Oversampling Technique) [7], [8] is a specific oversampling method that generates synthetic instances via nearest-neighbor interpolation rather than duplicating them, as traditional oversampling does, thereby increasing the number of minority-class instances. There are several variations of SMOTE: SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) [9]. This strategy identifies the categorical features and generates new synthetic instances of the minority class by combining close neighbors across numerical features and preserving categories in nominal features; SMOTE-Borderline [10]. This strategy focuses on minority examples near the decision boundary between classes. These points are the most likely to be misclassified, so the algorithm generates synthetic instances around them to reinforce the representation of the minority class in critical areas of the decision space; SMOTEENN [11]. This strategy is a hybrid technique that combines oversampling and undersampling to address class imbalance in datasets. First, it applies SMOTE, generating synthetic instances of the minority class to increase its representation. Then, it uses ENN (Edited Nearest Neighbors), which removes examples (from both the majority and minority classes) that are misclassified by their nearest neighbors, cleaning up the decision space and reducing noise. The result is a balanced dataset with more precise decision boundaries; and finally, ADASYN [10].

Algorithm-oriented balancing techniques apply strategies to algorithms to improve minority-class detection performance rather than directly to the data [12]. These strategies are integrated into the model training process and affect it. In this context, cost-sensitive learning [13] is a strategy that aims to improve model performance in minority classes. This technique assigns differentiated penalties to classification errors so that errors made on the less-represented class have a higher cost than those on the dominant class. By incorporating these penalties into the model's loss function, the algorithm is encouraged to pay more attention to instances of the minority class, thereby reducing bias toward the majority class.

Finally, hybrid techniques combine several approaches to leverage the advantages of each and minimize their weaknesses. For example, SMOTE can be used to generate synthetic data and then train a model with a class penalty to improve sensitivity to the minority class [14].

This paper outlines the steps required to enhance the prediction performance of an XGBoost model, employing a data-centric approach to rebalance the class distribution in the training dataset. To optimize the number of instances for each class, we implemented a genetic algorithm (GA). The approach described in this paper was developed using the CRISP-DM methodology [15], [16], with the following components considered: data understanding, data preparation, modeling, and evaluation. The details of the methodology's implementation will be explored in greater depth in the Materials and Methods section.

The contributions of this article are as follows: (i) It provides a brief explanation of existing data balancing mechanisms; (ii) It reviews the mechanisms and metrics for evaluating the performance of machine learning models; (iii) **it** describes the experiment developed based on CRISP-DM, adapting it into four phases; (iv) **it** performs an exploratory data analysis on the myocardial infarction complications dataset and outlines a data processing scheme that includes missing data handling and data balancing; (v) It proposes a data balancing scheme using three strategies: a traditional balancing scheme that uses oversampling with an equifrequent class ratio, the use of SMOTE Borderline, and the optimization of the data ratio to be used in random undersampling (RUS) by constructing a genetic algorithm. This phase yields different datasets that will be utilized in the modeling phase; (vi) It uses an XGBoost model with adjusted parameters and evaluates the performance of the trained model with the different datasets obtained after data rebalancing; and (vii) Finally, **it** uses four models: XGBoost, Decision Trees, Random Forest, and LightGBM. Each model is trained on both the raw and rebalanced datasets. The rebalancing uses ROS, SMOTE, and SMOTE Borderline methods. The balancing techniques are applied with both an equal-frequency ratio and the optimal ratio identified by the GA. The results are then compared.

1.1. BACKGROUND

Cardiovascular diseases are one of the leading causes of mortality worldwide. Myocardial infarction is one of the diseases classified as a cardiovascular disease and is associated with high mortality [15]; however, the patient's progress can be affected by complications that may arise at different times during the first year after

the infarction. To take preventive measures, there is interest in predicting possible complications that could worsen the patient's condition or even lead to death [16]. The dataset selected for this study is "Myocardial Infarction Complications," which is used to predict myocardial infarction (MI) complications based on patient information at various stages of disease progression and disease characteristics.

Various studies have used the "Myocardial Infarction Complications" dataset to train classification models. Below, we describe the main aspects reported by the authors, with particular focus on the class imbalance in the dataset [17]. In [18], the focus is on the multiclass classification of patients with heart disease, with the Lethal Outcome attribute (8 classes) as the target. The authors compare the performance of machine learning models using five different feature selection techniques. The dataset exhibits a high imbalance, which the authors addressed using SMOTE, thereby bringing the minority classes to the same proportion as the majority class. The study used four (4) classifiers, and hyperparameters were adjusted using grid search. The models' performance was evaluated using the following metrics: Accuracy, Precision, Recall, and F1-Score. The study concludes that feature selection has a high impact on model performance, reducing training time and improving performance metrics.

The authors of the study [19] performed data preprocessing, including imputation of missing values, feature selection, and data balancing. The treatment of missing values consisted of eliminating columns with more than 25% missing values, and the remaining columns with missing values were imputed with the mode for categorical variables and the arithmetic mean for numerical variables.

To address class imbalance, random oversampling (ROS) and stratified cross-validation were used. The data were split into five equal parts for 5-fold cross-validation. The study used four main multilabel classification approaches. The first approach used is Binary Relevance [20], considered the baseline method, which treats each label as a separate binary classification task. The second approach is Classifier Chains [21], which extends the previous method by incorporating label dependence. The third method applied is Label Powerset [22], which solves the challenge of multi-label classification by transforming the problem into a single multi-class classification task. The fourth approach used is RAKEL-D [23], an ensemble technique designed to improve the performance of multi-label classifiers by using multiple label subsets. The study evaluated performance using three different base classification algorithms and metrics, including accuracy, F1-score, Hamming Loss, Jaccard Index, precision, recall, and ROC-AUC.

2. MATERIALS AND METHODS

2.1. DATASET

The dataset on complications of myocardial infarction [17] contains data collected between 1992 and 1995 at the Krasnoyarsk Interdistrict Clinical Hospital (Russia). This dataset has 111 features (93 categorical variables and 17 continuous variables) and a target class (Lethal Outcome), corresponding to the possible complications that may arise based on the patient's information. It is possible to predict these complications at different points in the patient's evolution: (a) upon admission to the hospital, (b) at the end of the first 24 hours after admission, (c) at the end of the second day after admission, and (d) at the end of the third day after admission. This article evaluates potential complications at the end of the first day (24 hours after admission to the hospital), excluding features suggested by the documentation. The prediction will be made in eight (8) categories: class 0: alive, class 1: cardiogenic shock, class 2: pulmonary edema, class 3: myocardial rupture, class 4: progression of congestive heart failure (CHF Progression), class 5: thromboembolism, class 6: asystole, and class 7: ventricular fibrillation. The dataset was obtained from <https://archive.ics.uci.edu/dataset/579/myocardial+infarction+complications>.

2.2. PROPOSED WORK

Our proposal is guided by the CRISP-DM methodology [24], [16] and considers the phases of data understanding, data preparation, modeling, and evaluation. Figure 1 shows the phases of the CRISP-DM methodology used and how each phase aligns with the proposal of this work. Figure 1 shows that the data analysis phase involves identifying categorical and continuous variables, evaluating their importance for classification, and identifying missing values in the dataset. The data preparation phase includes imputing missing values using a Random Forest model to complete the dataset and removing columns with a missing data percentage greater than 60%. This phase uses a genetic algorithm to find the optimal data rebalancing ratio, as explained in depth in the next section. In the modeling phase, the model used is XGBoost. Finally, in the evaluation phase, metrics such as the F1 score, precision, recall, geometric mean (G-mean), balanced precision index, and the confusion matrix are used to evaluate the model's performance.

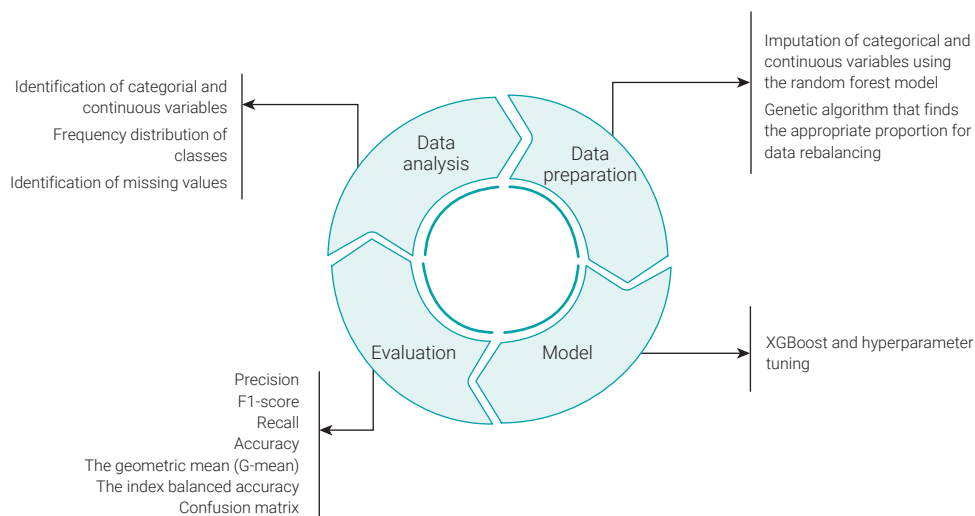


Figure 1. Crisp-DM methodology adaptation

Source: Own work

Data Analysis: Exploratory data analysis of the selected dataset reveals two aspects that may affect classification: (1) data imbalance and (2) missing values. About class imbalance, most cases in the dataset correspond to the class called “alive,” with a total of 1,429 cases, representing 84.1% of the total data, while minority cases, such as thromboembolism, with 12 cases representing 7% of the total data, pulmonary edema, with 18 cases representing 1.1% of the total data, 23 cases of heart failure, representing 1.4% of the total data, 27 cases of asystole and ventricular fibrillation, each representing 1.6% of the total data, 54 cases of myocardial rupture, representing 3.2% of the total data, and, finally, 110 cases of cardiogenic shock, representing 6.5% of the total data.

Figure 2 shows the percentage distribution of classes, indicating that classes with a lethal outcome (cardiogenic shock, pulmonary edema, myocardial rupture, progression of congestive heart failure, thromboembolism, asystole, and ventricular fibrillation) are significantly underrepresented in the dataset, accounting for only 15.9% of total cases. In comparison, the “alive” class accounts for 84.1% of the total dataset. It introduces bias in models’ predictions of minority classes.

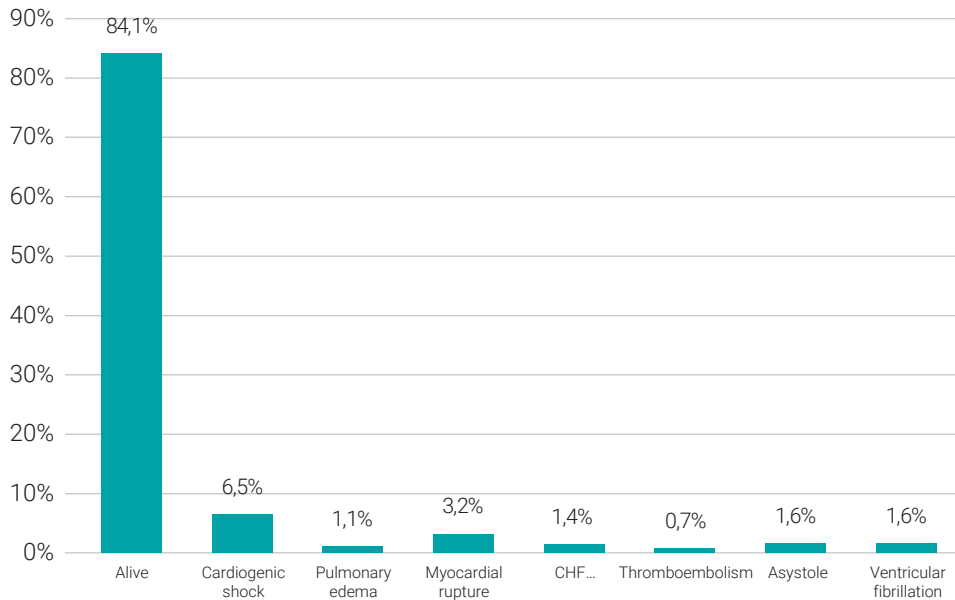


Figure 2. Percentage distribution of class frequency in the dataset on complications of myocardial infarction.

Source: Own work

The second relevant aspect concerns missing values. Missing values are often attributed to human error during data collection and dataset construction [25]. A dataset with missing values can impair the performance of machine learning (ML) models, leading to biased or erroneous analyses and poorer generalization, thereby reducing predictive accuracy. To address this issue, data imputation techniques have been developed to preserve model integrity and avoid erroneous conclusions; however, before applying them, it is important to analyze the dataset's missing values. In the analysis of missing values performed on the myocardial infarction complications dataset, categorical and continuous variables were identified. They were then grouped into the following categories according to the percentage of missing values: low (<10%), moderate (10–60%), and high (>60%).

The analysis shows that, for continuous variables, seven had missing values of less than 10%, seven had missing values between 10% and 60%, and three had missing values of more than 60%. Analysis of the dataset reveals that three continuous variables and one categorical variable have a high level of missing values. For categorical variables, 84 had missing values of less than 10%, 6 had missing values between 10% and 60%, and 1 had missing values of more than 60%.

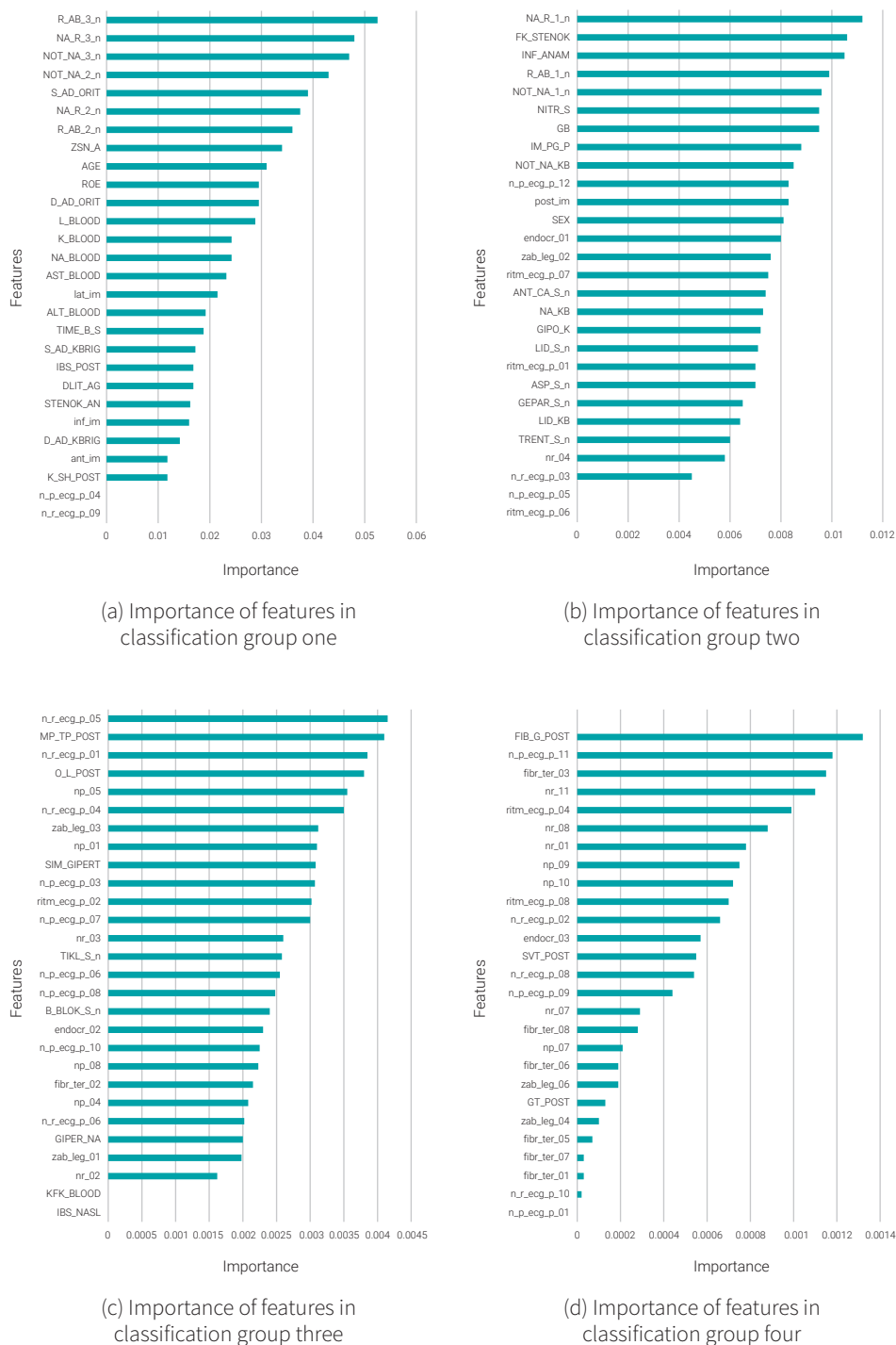


Figure 3. Importance of features in classification

Source: Own work

Once the analysis of missing values is complete, the importance of the features in the classification is evaluated to assess the impact of the variables with the most missing values. To do this, the Random Forest model is used, which, through the `feature_importances_` attribute, indicates the relative importance of each feature in the model's predictions by assigning a numerical value to each feature, with the total sum of these values equaling 1. The study finds that the attributes accounting for more than 60% of the missing values are of little importance for classification, including IBS_NASL (0.0), KFK_BLOOD (0.0), S_AD_KBRIG (0.0168), and D_AD_KBRIG (0.0139). Figure 3 shows the importance of the features for classification. Due to the number of features, they are presented separately across four graphs to facilitate better visualization. This information is relevant to the experiment, as it helps guide decisions during data processing.

Data preparation: Data preparation aims to minimize erroneous data in the training of machine learning models. In this phase, we analyze two key aspects during data preparation: handling missing values and treating imbalanced data.

The first aspect to address is the treatment of missing values. We identified columns with more than 60% missing values. The proposal is to eliminate features with more than 60% missing values. This decision is based on the fact that, in addition to the number of missing values, the analysis from the previous phase indicates that these features have a low impact on classification, as shown in Figure 3. The selected features and their impact measures are as follows: IBS_NASL: 0.0; KFK_BLOOD: 0.0; S_AD_KBRIG: 0.0168; and D_AD_KBRIG: 0.0139. The measure obtained ranges from 0 to 1, where a value close to 0 suggests the feature has little impact on the prediction and does not significantly affect the model's performance. Conversely, a high value indicates that the evaluated feature has a substantial effect on the model's performance. The total importance of all features equals 1.

The second aspect to be addressed is imbalanced data. The imbalance ratio (IR) for survivor patients (class 0) vs. each cause of death is as follows: (class 1) cardiogenic shock, 1:13; (class 2) pulmonary edema, 1:79; (class 3) myocardial rupture, 1:26; (class 4) congestive heart failure, 1:59; (class 5) thromboembolism, 1:118; (class 6) asystole, 1:52; and (class 7) ventricular fibrillation, 1:52. The IR shows that most classes with lethal outcomes are highly underrepresented in the dataset, which generates bias toward predicting the majority class. The selected strategies for data rebalancing were: Equal-frequency ROS, optimal data proportion ROS, and SMOTE Borderline.

The optimal data proportion to balance the dataset is determined using a Genetic Algorithm (GA), which optimizes the proportion of data for each minority class. A GA is

a type of bio-inspired optimization algorithm [26]. Figure 4 shows the life cycle of the proposed genetic algorithm and some details of its implementation.

The proposed individual is represented as a binary list indicating the proportion of data in the minority classes, which constitutes the strategy for class rebalancing. The size of the individual is calculated based on the majority class and the number of minority classes. To ensure the representation of the individual in binary and multi-class classification problems, the algorithm's size is calculated by Equation (1), where the number of minority classes corresponds to the number of unbalanced classes multiplied by the ceiling of the logarithm of the number of instances of the majority class in base 2. In this case, the proposed individual has a chromosome length of 70, and each group of 10 genes represents the proportion of each minority class. The population consists of one hundred (100) individuals, including the individual who initially sets up the population with an equiproportional proportion, i.e., all minority classes have the same size. In this case, the equiproportional proportion is given by: Strategy = {1: 957, 2: 957, 3: 957, 4: 957, 5: 957, 6: 957, 7: 957}.

$$Individuo_size = minority * \lceil \log_2 \max(<class_frequency>) \rceil$$

Equation 1: Calculation of the size of the individual

The aptitude function evaluates the proposed individual, which represents a strategy for balancing the dataset's proportions of minority classes. Next, ROS (Random Oversampling) is applied using a sampling strategy that corresponds to the proposed individual. Once the dataset has been rebalanced, the XGBoost classifier (XGBClassifier) is trained with hyperparameters tuned via grid search. The model then makes predictions on the test set, and the F1 Score (f1_score) is used as the performance metric. The returned value (f1s) represents the quality of that strategy applied to the model.

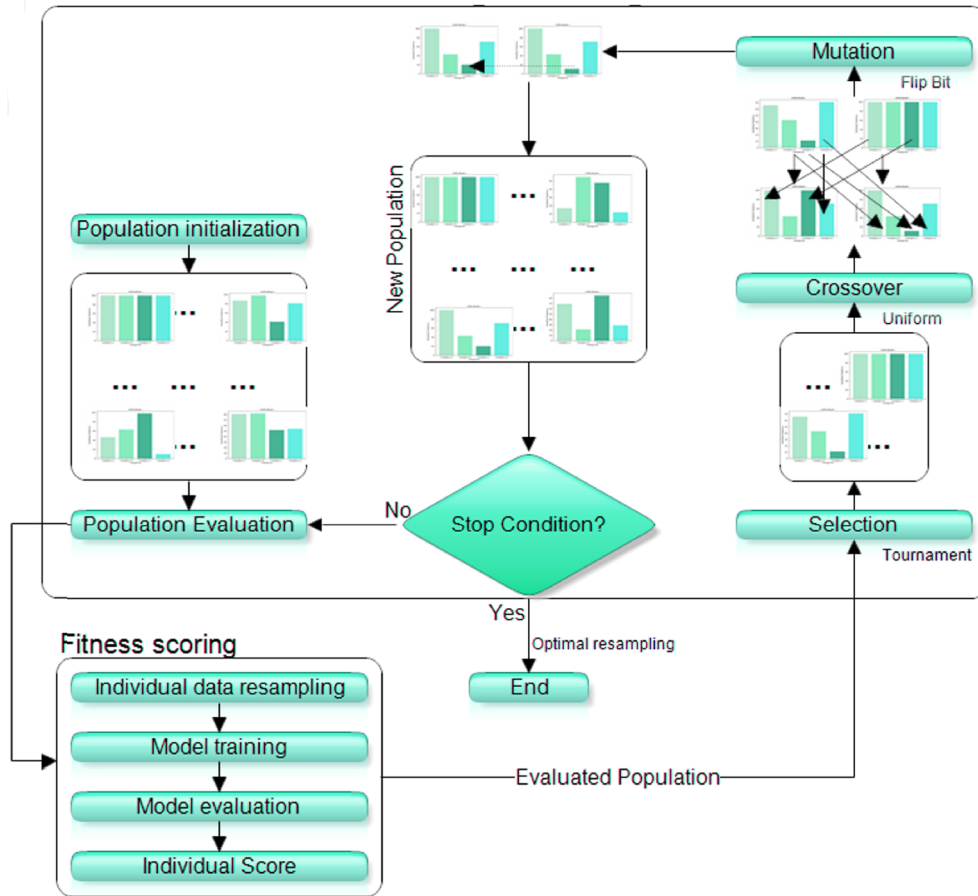


Figure 4. Genetic Algorithm
Source: Own work

The proposed algorithm incorporates an elitism mechanism via a hall of fame. This approach ensures that the best individuals found are preserved in each generation, preventing them from being modified by genetic operators such as selection, crossover, or mutation; that is, they are excluded from crossover and mutation. In each generational cycle, new individuals are selected to form the offspring, genetic operators are applied, their fitness is evaluated, and the elite individuals are reintegrated directly into the population. This maintains diversity in the population. Evolutionary progress is also recorded in a logbook, and the hall of fame is dynamically updated with the best individuals found in each generation.

The crossover operator used is cxUniform, with a probability of 0.2, which randomly exchanges genes between two individuals. Once the two individuals to be crossed have been selected, a crossover probability is randomly generated for each gene pair, creating a crossover mask in which genes are exchanged only if the

probability is less than 20%. Figure 5 shows the work schedule for the cross-over operator.

P1	1	0	1	1	0	1	0	0	1	1
P2	0	1	1	0	1	0	0	1	0	0
PB rand	0.9	0.1	0.4	0.3	0.05	0.8	0.15	0.6	0.1	0.7
CX S1	N	Y	N	N	Y	N	Y	N	Y	N
S1	1	1	1	1	1	1	0	0	0	1
CX S1	N	Y	N	N	Y	N	Y	N	Y	N
S2	0	0	1	0	0	0	0	1	1	0

Figure 5. Crossover operator cxUniform
Source: Own work

The selected mutation operator is mutflip-bit, with a probability of 0.2. For each gene of the individual, a random probability of change is generated; if it is less than 20%, the gene changes to its opposite value. This operator introduces controlled genetic diversity into the population to prevent genetic stagnation during individual evolution. Figure 6 shows the work schedule for the mutation operator.

I1	0	1	1	0	1	0	0	1	0	0
PB rand	0.9	0.1	0.4	0.3	0.05	0.8	0.15	0.6	0.1	0.7
CX	N	Y	N	N	Y	N	Y	N	Y	N
S1	1	1	1	1	1	1	0	0	0	1

Figure 6. Mutation operator mutflip-bit
Source: Own work

The selection operator used is selTournament with size 3. This operator selects individuals through small tournaments: a small group of randomly selected individuals compete against each other, thereby selecting the best (according to their fitness) to reproduce.

The GA is executed using the DEAP v1.4.3 framework, with 30 generations run. The result obtained from the genetic algorithm is the optimal ratio used for class re-balancing. The dataset is divided into training data, which accounts for 70%, and test

data, which accounts for 30%. Data rebalancing is performed using three strategies, generating three new datasets. It is essential to clarify that the rebalanced data are generated solely for training purposes. The first strategy employed was rebalancing the dataset to an equal-frequency distribution using ROS. The second strategy used Borderline-SMOTE, and the third strategy combined ROS with the optimal class ratio determined by the genetic algorithm (1: 349, 2: 235, 3: 933, 4: 239, 5: 870, 6: 685, 7: 423).

Model: Four models were used in this study: XGBoost, Decision Trees, Random Forest, and LightGBM. For each model, the hyperparameters were tuned using Optuna [27]. Optuna is an open-source framework for automating hyperparameter optimization in machine learning models. A baseline model for each algorithm was trained on the original data without rebalancing, while six additional models per algorithm, using the same configuration, were trained on datasets rebalanced to either equal-frequency or GA-determined proportions to reduce majority-class bias. Table 1 presents the trained models, the datasets, and the proportion of data used for oversampling for every algorithm. The optimal ratio is (1: 349, 2: 235, 3: 933, 4: 239, 5: 870, 6: 685, 7: 423).

Table 1. Models

Model	Balancing Strategy	Data Proportion
Baseline model	NA	Raw data
Model 1	ROS	Equi-frequent
Model 2	ROS	Optimal ratio
Model 3	SMOTE Borderline	Equi-frequent
Model 4	SMOTE Borderline	Optimal ratio
Model 5	SMOTE	Equi-frequent
Model 6	SMOTE	optimal ratio

XGBoost (Extreme Gradient Boosting) is an ML algorithm based on decision trees and the boosting technique, combining multiple weak trees to form a stronger model. According to research, it is recommended for use with large and highly unbalanced datasets because this algorithm avoids overfitting, which is common when the model is forced to learn instances of classes that are underrepresented or belong to the minority class, and provides a mechanism for adjusting class weights, which improves model performance, among other advantages [28].

Decision Trees: A decision tree is a non-parametric supervised learning method used for both classification and regression tasks. Its primary goal is to create a model that predicts the value of a target variable by learning simple decision rules (of the “if-then” type) inferred directly from the data features. One of its advantages is that, for some authors, the model is considered a “white box” model, meaning that its logic is easily interpretable and visualizable, reflecting the decision-making process through Boolean logic. Furthermore, decision trees require little data preparation and can handle both numerical and categorical variables [29].

Random Forest is a supervised learning algorithm. It combines multiple decision trees to deliver a robust and accurate predictive model. It benefits from a lower risk of overfitting, high accuracy, and resilience to noise and outliers. Still, if hundreds of trees are used, interpretation becomes significantly more difficult than with a single tree [30].

LightGBM (Light Gradient Boosting Machine) is a highly efficient implementation of the Gradient Boosting Decision Tree (GBDT) algorithm. It is an algorithm that some authors consider efficient, scalable, and memory-efficient [31].

3. RESULTS

Table 2 shows the ML algorithms, the balancing techniques, the data ratio used in the rebalancing, and the metrics used.

To evaluate model performance, we use the confusion matrix, which allows us to visualize the performance of a classification model by showing correct and incorrect predictions, and the precision metric, which indicates the proportion of positive predictions made by the model that were correct, i.e., it measures how accurate the positive predictions made by the model are. Recall (sensitivity), known as the true positive rate, evaluates how well a model correctly identifies positive instances; specificity assesses the model’s ability to identify negative instances correctly; the F1-score is defined as the harmonic mean that combines the accuracy of positive predictions (precision) with the model’s ability to detect all actual positive cases (recall); and the Index Balanced Accuracy (IBA), which is the average recall per class with a penalty factor that considers the imbalance between classes and the dispersion of errors. This metric permits us to evaluate not only how well the model correctly identifies each class but also how it distributes its errors, favoring models that do not lean excessively toward the majority classes [32].

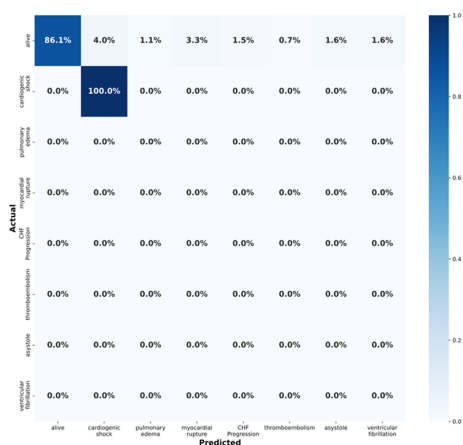
Table 2. Results

Model	Balancing Strategy	Data Proportion	MCC	ImbAcc	Acc	Prec	Recall	F1-Score
LightGBM	NA	Raw data	0.26	0.16	0.85	0.18	0.16	0.16
	ROS	Equi-frequent	0.36	0.29	0.8	0.32	0.29	0.3
	ROS	Optimal ratio	0.4	0.26	0.84	0.43	0.26	0.28
	SMOTE Borderline	Equi-frequent	0.41	0.25	0.86	0.41	0.25	0.29
	SMOTE Borderline	optimal ratio	0.33	0.19	0.85	0.26	0.19	0.2
	SMOTE	Equi-frequent	0.33	0.19	0.85	0.26	0.19	0.2
	SMOTE	optimal ratio	0.35	0.18	0.86	0.21	0.18	0.10
Random Forest	NA	Raw data	0.34	0.17	0.86	0.22	0.17	0.18
	ROS	Equi-frequent	0.30	0.2	0.86	0.26	0.2	0.21
	ROS	Optimal ratio	0.32	0.18	0.86	0.25	0.18	0.19
	SMOTE Borderline	Equi-frequent	0.37	0.18	0.86	0.21	0.18	0.18
	SMOTE Borderline	optimal ratio	0.26	0.16	0.85	0.21	0.16	0.16
	SMOTE	Equi-frequent	0.35	0.19	0.86	0.24	0.19	0.2
	SMOTE	optimal ratio	0.25	0.16	0.85	0.23	0.16	0.16
Decision Tree	NA	Raw data	0.34	0.19	0.85	0.19	0.18	0.18
	ROS	Equi-frequent	0.34	0.19	0.85	0.19	0.18	0.18
	ROS	Optimal ratio	0.34	0.19	0.85	0.19	0.18	0.18
	SMOTE Borderline	Equi-frequent	0.34	0.19	0.85	0.19	0.18	0.18
	SMOTE Borderline	optimal ratio	0.03	0.17	0.69	0.15	0.17	0.15
	SMOTE	Equi-frequent	0.08	0.2	0.47	0.16	0.2	0.14
	SMOTE	optimal ratio	0.09	0.22	0.41	0.21	0.21	0.15
XGBoost	NA	Raw data	0.37	0.06	0.86	0.23	0.17	0.18
	ROS	Equi-frequent	0.38	0.25	0.8	0.29	0.3	0.28
	ROS	Optimal ratio	0.42	0.34	0.83	0.51	0.34	0.36
	SMOTE Borderline	Equi-frequent	0.34	0.13	0.85	0.24	0.22	0.23
	SMOTE Borderline	optimal ratio	0.34	0.12	0.85	0.28	0.22	0.24
	SMOTE	Equi-frequent	0.29	0.08	0.84	0.21	0.17	0.18
	SMOTE	optimal ratio	0.42	0.27	0.83	0.5	0.34	0.23

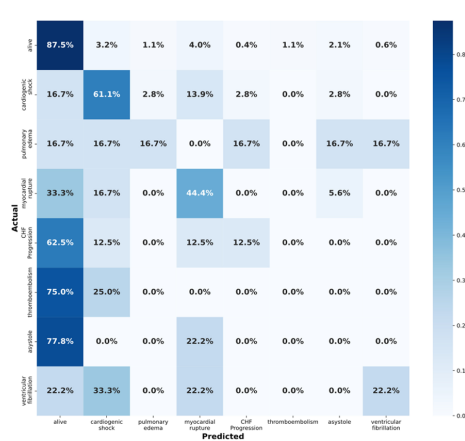
Source: own work

To assess significant differences in classification model performance across different data rebalancing approaches, the Friedman test was applied. This procedure is appropriate when comparing multiple algorithms under homogeneous conditions. It uses repeated metrics as comparison blocks. The Friedman test is based on ranks, not absolute values, allowing detection of differences without assuming normality. This makes it particularly suitable for learning contexts with imbalanced classes. Application of the test yielded a statistic of 103.05 and a p-value of 1.2×10^{-20} , indicating highly significant differences between the models. Consequently, it is confirmed that some models consistently outperform others. The ROS_OPTIMAL and SMOTE_OPTIMAL variants with XGBoost stand out, achieving the highest MCC and F1 scores. These models establish themselves as the most competitive candidates within the analyzed set.

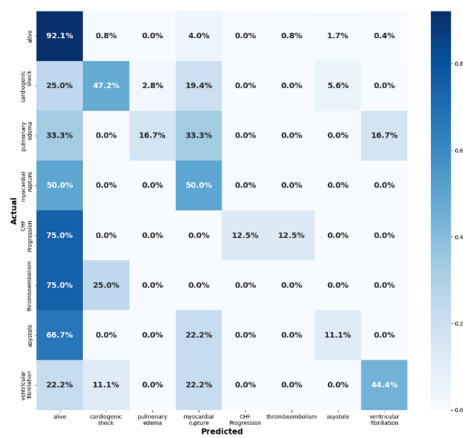
The post-hoc analysis using the Nemenyi test shows that, although the overall differences were significant, not all pairwise comparisons reached statistical significance, suggesting that the leading models cluster at the same performance level. Taken together, these results indicate that model choice matters: ROS_OPTIMAL and SMOTE_OPTIMAL with XGBoost offer advantages in metrics critical for imbalanced data, while the rank analysis confirms their relative consistency compared to less robust alternatives. Below, we present the results obtained using the XGBoost model.



(a) Scaled confusion matrix Baseline model



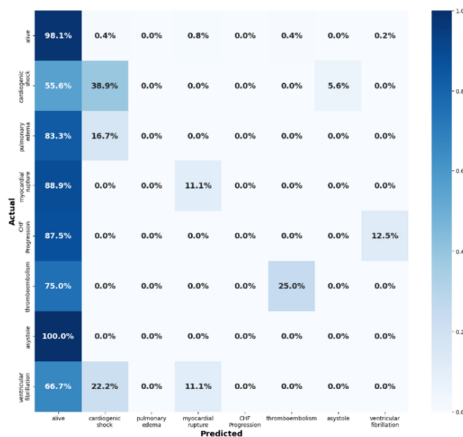
(b) Scaled confusion matrix model 1 ROS Equi-frequent



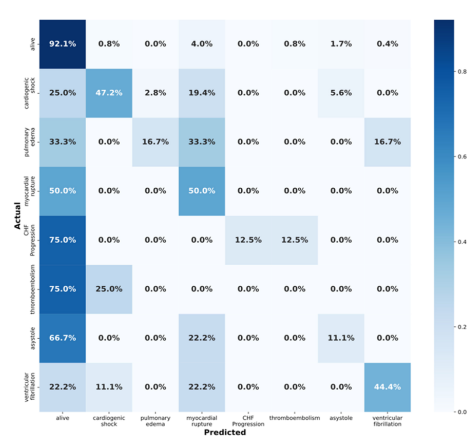
(c) Scaled confusion matrix model 2
ROS optimal proportion



(d) Scaled confusion matrix Model 3
SMOTE Borderline Equi-Frequent



(e) Scaled confusion matrix Model 4
SMOTE Borderline optimal proportion



(f) Scaled confusion matrix Model 5
SMOTE Equi-frequent proportion

Figure 7. Scaled Confusion Matrix.

Source: Own work

Figure 7 shows the scaled confusion matrix obtained by running the models on the different datasets resulting from data processing and preparation. The scaled confusion matrix for the baseline model, as shown in Figure 7(a), indicates that 86.1% of the data corresponding to the “alive” class and 100% of the data corresponding to the “cardiogenic shock” class were classified correctly. However, for the remaining classes, the model identified no instances, indicating a bias toward the “alive” and

“cardiogenic shock” classes. This bias is due to the fact that the remaining classes have few instances.

Figure 7(b), the scaled confusion matrix for Model 1, shows improved predictions compared to Figure 7(a). This suggests that the model has begun to recognize multiple classes, albeit with varying levels of accuracy, and that 87.5% and 61.1% of the data corresponding to the “alive” and “cardiogenic shock” classes, respectively, were classified correctly, indicating a tendency to classify many instances as belonging to those classes.

Figure 7(c), which shows the confusion matrix for Model 2 (ROS with the optimal ratio), indicates that the model correctly identifies 7 of the 8 classes in the dataset. This result suggests greater sensitivity in detecting positive cases in the minority classes. The other models do not outperform this one.

Table 2 shows the metrics: precision (Pre), recall (Rec), F1-score (F1), geometric mean (Geo), and Index Balanced Accuracy (IBA). Regarding precision, analysis of minority classes such as pulmonary edema, myocardial rupture, and CHF progression shows that the baseline model fails to detect any patients, while Models 1 and 3 improve precision. Regarding the F1-score metric, all models achieve high F1-scores for the “alive” class, indicating that they predict this class well. However, the baseline model has the lowest IBA (0.17), indicating a strong bias toward the majority class and poor performance in the other classes. The baseline model and Model 3 tend to overfit to the “alive” class, potentially masking deficiencies in detecting minority classes.

Table 3. Evaluation metrics for the target feature of models XGBoost

Class 0: Alive (471 Patients)	pre	rec	F1	geo	IBA
Model raw data	0.86	1	0.93	0.39	0.17
Model ROS equipfrequent data	0.93	0.87	0.9	0.76	0.6
Model SMOTE Borderline	0.89	0.97	0.93	0.59	0.37
Model ROS Optimal Proportion	0.92	0.92	0.92	0.74	0.56
Class 1: Cardiogenic shock (36 patients)	pre	rec	F1	geo	IBA
Model train raw data	1.0	0.39	0.56	0.62	0.37
Model ROS equipfrequent data	0.48	0.61	0.54	0.76	0.56
Model SMOTE Borderline	0.55	0.47	0.51	0.68	0.74
Model ROS Optimal Proportion	0.68	0.47	0.58	0.68	0.44
Class 2: Pulmonary edema (6 Patients)	pre	rec	F1	geo	IBA
Model raw data	0	0	0	0	0
Model ROS equipfrequent data	0.14	0.17	0.15	0.41	0.15

(continúa)

(viene)

Model SMOTE Borderline	0.17	0.06	0.08	0.23	0.22
Model ROS Optimal Proportion	0.25	0.5	0.31	0.69	0.45
Class 3: Myocardial ruptura (18 patients)	pre	rec	F1	geo	IBA
Model raw data	0	0	0	0	0
Model ROS equipfrequent data	0.22	0.44	0.29	0.65	0.4
Model SMOTE Borderline	0	0	0	0	0.0
Model ROS Optimal Proportion	0.22	0.50	0.31	0.69	0.45
Class 4: CHF Progression (8 Patients)	pre	rec	F1	geo	IBA
Model raw data	0	0	0	0	0
Model ROS equipfrequent data	0.2	0.12	0.15	0.35	0.11
Model SMOTE Borderline	0	0	0	0	0
Model ROS Optimal Proportion	1	0.12	0.22	0.35	0.11
Class 5: Thromboembolism (4 Patients)	pre	rec	F1	geo	IBA
Model raw data	0	0	0	0	0
Model ROS equipfrequent data	0	0	0	0	0
Model SMOTE Borderline	0.33	0.25	0.29	0.5	0
Model ROS Optimal Proportion	0	0	0	0	0
Class 6: Asystole (9 Patients)	pre	rec	F1	geo	IBA
Model raw data	0	0	0	0	0
Model ROS equipfrequent data	0	0	0	0	0
Model SMOTE Borderline	0	0	0	0	0
Model ROS Optimal Proportion	0.09	0.11	0.1	0.33	0.1
Class 7: Ventricular fibrillation (9 Patients)	pre	rec	F1	geo	IBA
Model raw data	0	0	0	0	0
Model ROS equipfrequent data	0	0	0	0	0
Model SMOTE Borderline	0	0	0	0	0
Model ROS Optimal Proportion	0.57	0.44	0.5	0.66	0.42

DISCUSSION AND CONCLUSIONS

Based on the overall metrics, we conclude that applying balancing strategies improves the model's performance in detecting the minority classes. The XGBoost algorithm in the baseline model, trained on unprocessed data, exhibits low metrics, such as the G-mean (0.12) and the IBA (0.06), due to majority-class bias. The model trained with balanced data using the optimal ratio and the ROS strategy performs better than the other models, achieving an accuracy of 0.50, a recall of 0.34, a G-mean of 0.48, and an IBA of 0.27. The model trained with balanced data using equal-frequency ROS also improves substantially over the baseline, though with

metrics lower than those achieved by the model trained with the optimal proportion. According to the F1-score, this model achieves the highest performance in detection and precision for the minority classes, balancing accurate detection with minimal false positives. Regarding the G-mean metric, Models 1 and 3 are the most balanced across all classes, suggesting that equal-frequency balancing and balancing with the optimal data ratio for minority classes could be suitable solutions to the problems posed by imbalanced data.

An approach that determines the optimal ratio for working with imbalanced classes, focused on developing a genetic algorithm (GA), yields predictive models that increase the F1-score to 0.36 and detect an additional class compared to the baseline model. Overall, the XGBoost algorithm trained with the optimal ratio is more robust, with a high F1-score, good precision and recall, and balanced G-mean and IBA. Taken together, the results suggest that optimally adjusting the proportion of minority classes and balancing with ROS are promising strategies for improving classification in biomedical contexts.

With regard to models that did not achieve superior results during training on resampled data using the optimal ratio, it is evident that, when applying synthetic oversampling techniques (SMOTE and SMOTE Borderline) with the optimal ratio, the Decision Tree model experiences a significant drop in performance. Its IBA drops from 0.34 to levels of 0.03 and 0.09, and its F1-score falls to 0.15. This suggests that generating synthetic data at this specific ratio introduces “noise” or overfitting, thereby confusing the tree’s decision boundaries.

For future work, in addition to the genetic algorithm, the use of ROSE (Random Over-Sampling Examples) could be included. It combines random oversampling and synthetic noise generation to help avoid overfitting that can occur with direct data duplication, as ROS does [33]. Regarding the genetic algorithm, one could experiment with alternative optimization methods, such as Particle Swarm Optimization (PSO), Evolutionary Strategies (ES), or other methods, to investigate whether better results can be obtained.

REFERENCES

- [1] A. Kavita, P. Sonal, and T. D. L. Nguyen, “Comparative analysis of K-NN, naïve Bayes, and logistic regression for credit card fraud detection,” *Revista Ingeniería Solidaria*, vol. 19, no. 3, pp. 1–22, 2023, doi: <https://doi.org/10.16925/2357-6014.2023.03.05>.

- [2] H. Allam, L. I. Makubvure, B. K. Gyamfi, K. N. Graham, and K. Akinwolere, "Text classification: How machine learning is revolutionizing text categorization," *Information*, vol. 16, no. 2, Art. no. 130, 2025, doi: <https://doi.org/10.3390/info16020130>.
- [3] M. Manoj Yadav and A.-H. Ronak, "Resampling imbalanced healthcare data for predictive modelling," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 2, Art. no. 36, 2025, doi: <https://doi.org/10.14569/IJACSA.2025.0160204>.
- [4] X. Xiaoying, L. Huawen, Z. Shouzheng, L. Lingbin, and L. Wen, "A novel progressively undersampling method based on the density peaks sequence for imbalanced data," *Knowledge-Based Systems*, vol. 216, Art. no. 106689, 2021, doi: <https://doi.org/10.1016/j.knosys.2020.106689>.
- [5] S. Tyagi and S. Mittal, "Sampling approaches for imbalanced data classification problem in machine learning," in *Proc. International Conference on Research in Intelligent Computing (ICRIC 2019)*, 2020, pp. 209–221, doi: https://doi.org/10.1007/978-3-030-29407-6_17.
- [6] D. Elreedy and A. F. Atiya, "A comprehensive analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, 2019, doi: <https://doi.org/10.1016/j.ins.2019.07.070>.
- [7] A. Fernández, S. García, and F. Herrera, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018, doi: <https://doi.org/10.1613/jair.1.11192>.
- [8] V. Rattan, R. Mittal, J. Singh, and V. M. (ND), "Analyzing the application of SMOTE on machine learning classifiers," in *Proc. International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2021, pp. 692–695, doi: <https://doi.org/10.1109/ESCI50559.2021.9396962>.
- [9] M. Mukherjee and M. Khushi, "SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features," *Applied System Innovation*, vol. 4, no. 1, Art. no. 18, 2021, doi: <https://doi.org/10.3390/asi4010018>.
- [10] D. Ishani and P. Vibha, "A comparative study of SMOTE, Borderline-SMOTE, and ADASYN oversampling techniques using different classifiers," in *Proc. 3rd International Conference on Smart Data Intelligence (ICSMDI)*, Trichy, India, 2023, doi: <https://doi.org/10.1109/ICSMDI57622.2023.00060>.
- [11] G. Husain, D. Nasef, R. Jose, J. Mayer, and K. Bekbol, "SMOTE vs. SMOTEENN: A study on the performance of resampling algorithms for addressing class imbalance in regression models," *Algorithms*, vol. 18, no. 1, Art. no. 37, 2025. (ND)

- [12] M. Altalhan, A. Algarni, and M. T.-H. Alouane, "Imbalanced data problem in machine learning: A review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: <https://doi.org/10.1109/ACCESS.2025.3531662>.
- [13] J. C. Rella, G. Claeskens, R. Cao, and J. M. Vilar, "Instance-dependent cost-sensitive parametric learning," *Neurocomputing*, vol. 615, Art. no. 128875, 2025, doi: <https://doi.org/10.1016/j.neucom.2024.128875>.
- [14] A. S. Desuky and S. Hussain, "An improved hybrid approach for handling class imbalance problem," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3853–3864, 2021, doi: <https://doi.org/10.1007/s13369-021-05347-7>.
- [15] A. Newaz, M. S. Mohosheu, and M. A. Al Noman, "Predicting complications of myocardial infarction within several hours of hospitalization using data mining techniques," *Informatics in Medicine Unlocked*, vol. 42, Art. no. 101361, 2023, doi: <https://doi.org/10.1016/j.imu.2023.101361>.
- [16] H. Shang, J. Wang, Y. Cong, H. Chen, and Y. Cui, "Risk prediction models for heart failure after myocardial infarction: A systematic review," *Medicine*, vol. 104, no. 30, 2025, doi: <https://doi.org/10.1097/MD.00000000000043480>.
- [17] UCI Machine Learning Repository, "Complications of myocardial infarction: A database for testing recognition and prediction systems," 2020. [Online]. Available: <https://archive.ics.uci.edu/> (SD)
- [18] M. Pandey and S. S. Routaray, "Feature optimized hybrid model for prediction of myocardial infarction," *F1000Research*, vol. 14, Art. no. 78, 2024, doi: <https://doi.org/10.12688/f1000research.160393.1>.
- [19] W. H. Elashmawi, A. Tamer, and A. Nasr, "Multi-target classification for predicting myocardial infarction complications: A comparative analysis of various machine learning techniques," in *Proc. International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 2024, doi: <https://doi.org/10.1109/MIUCC62295.2024.10783524>.
- [20] O. D. Luaces, J. R. Quevedo, J. Díez, C. J. Bahamonde, and A. Alonso, "Binary relevance efficacy for multilabel classification," *Progress in Artificial Intelligence*, vol. 1, no. 4, pp. 303–313, 2012, doi: <https://doi.org/10.1007/s13748-012-0030-x>.
- [21] D. J. Touw, "Classifier chain networks for multi-label classification," *Expert Systems with Applications*, vol. 286, (ND), 2024, doi: <https://doi.org/10.48550/arXiv.2411.02638>.

- [22] E. A. Cherman, M. C. Monard, and J. Metz, "Multi-label problem transformation methods: A case study," *CLEI Electronic Journal*, vol. 14, no. 1, 2011. (SD)
- [23] A. G. C. de Sá, P. G. P. Gomes, L. P. (SD), A. A. (SD), and F. (SD), "A robust experimental evaluation of automated multi-label classification methods," in *Proc. Genetic and Evolutionary Computation Conference (GECCO)*, 2020, pp. 175–183, doi: <https://doi.org/10.1145/3377930.3390231>.
- [24] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proc. 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.
- [25] E. Emmanuel et al., "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, Art. no. 140, 2021, doi: <https://doi.org/10.1186/s40537-021-00516-9>.
- [26] S. Katoch, S. Chauhan, and V. Kumar, "A review on genetic algorithm: Past, present, and future," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 8091–8126, 2021, doi: <https://doi.org/10.1007/s11042-020-10139-6>.
- [27] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019, pp. 2623–2631, doi: <https://doi.org/10.1145/3292500.3330701>.
- [28] Y. Zhang, J. Peng, Y. Sun, and Y. (ND), "Research and application of XGBoost in imbalanced data," *International Journal of Distributed Sensor Networks*, vol. 18, no. 6, 2022, doi: <https://doi.org/10.1177/155013292211069>.
- [29] B. Taha Jijo and A. Mohsin Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, 2021, doi: <https://doi.org/10.38094/jastt20165>.
- [30] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in *Proc. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI)*, 2018, pp. 758–763, doi: https://doi.org/10.1007/978-3-030-03146-6_86.
- [31] A. Choudhury, A. Mondal, and S. Sarkar, "Searches for the BSM scenarios at the LHC using decision tree-based machine learning algorithms: A comparative study and review of Random Forest, AdaBoost, XGBoost, and LightGBM frameworks," *The European Physical Journal Special Topics*, vol. 233, no. 15, pp. 2425–2463, 2024, doi: <https://doi.org/10.1140/epjs/s11734-024-01308-x>.

- [32] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2021, doi: <https://doi.org/10.1016/j.aci.2018.08.003>.
- [33] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 92–122, 2014, doi: <https://doi.org/10.1007/s10618-012-0295-5>.
- [34] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying the CRISP-DM process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, doi: <https://doi.org/10.1016/j.procs.2021.01.199>.