Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 23, Iss. 2, pp. 11–22, 2025
https://doi.org/10.21528/lnlm-vol23-no2-art2

PREDICTION OF HOSPITALIZATION TIME AND SURVIVABILITY OF PATIENTS WITH CONGESTIVE HEART FAILURE

João Carlos Pereira Alves[1], Ricardo Menezes Salgado[2]
Iago Augusto Carvalho[2], Eric Batista Ferreira[1]

[1] Department of Statistics, Universidade Federal de Alfenas (UNIFAL-MG), Brazil.
[1] Department of Computer Science, Universidade Federal de Alfenas (UNIFAL-MG), Brazil
joao.carlos@sou.unifal-mg.edu.br, {ricardo.salgado, eric.ferreira, iago.carvalho}@unifal-mg.edu.br

Abstract – Congestive heart failure (CHF) is a serious medical condition associated with high mortality rates. To improve prognosis and treatment, exploring new strategies is essential. This study investigates the use of electronic medical records to train machine learning models in predicting survival and hospitalization time for patients with CHF. Using data from 299 patients collected in Faisalabad, Pakistan, a suite of algorithms was evaluated, including MLP, logistic regression, Random Forests, decision tree, k-Nearest Neighbors, Naive Bayes, and Gradient Boosting. The methodology employed the SMOTE technique for class balancing and a rigorous 10-fold stratified cross-validation for performance evaluation. The Random Forest model emerged as the top performer, achieving a mean accuracy of 0.80 (± 0.05) and an F1-Score of 0.80 (± 0.05) in predicting patient survival. Statistical significance tests confirmed that the superiority of the Random Forest over the worst-performing model (MLP) is statistically significant ($p < 0.05$). For the prediction of hospitalization time, an error rate of 26% was observed. These findings underscore the statistically validated potential of machine learning models in predicting clinical outcomes in patients with CHF, representing an innovative approach to improve diagnostic efficiency and reduce the impact of CHF on public health.

Keywords – Congestive Heart Failure, Machine Learning, Clinical Prediction, hospitalization time, survivability.

1. INTRODUCTION

According to the World Health Organization (WHO), congestive heart failure (CHF) is one of the leading causes of death worldwide, responsible for approximately 8.9 million deaths annually. As a severe and growing clinical condition, it is estimated to affect more than 37 million people worldwide. CHF is characterized by the heart's inability to pump enough blood to meet the body's needs and can lead to a series of severe complications, including kidney failure, anemia, and even death. With the ageing population and the increasing number of cases of heart disease, CHF has become a rising public health concern, with high treatment costs and significant impacts on the life expectancy and quality of life of affected individuals [1–3].

This condition is quite complex and can manifest in various forms, resulting in different types of heart failure, such as Heart Failure with Reduced Ejection Fraction (HFrEF) and Heart Failure with Preserved Ejection Fraction (HFpEF). HFrEF, or systolic heart failure, involves significant myocardial loss/dysfunction, potentially resulting from infarction, genetic abnormalities, or toxins, which reduce the heart's pumping ability. On the other hand, HFpEF, or diastolic heart failure, is characterized by stiffness of the cardiac muscle, preventing proper relaxation and filling during diastole, but typically maintaining an ejection fraction above 50%, leading to increased pressure in the vessels and heart, triggering heart failure symptoms [1].

In addition to these types, other cardiac issues may contribute to CHF, such as aortic valve stenosis, where the valve between the left ventricle of the heart and the aorta narrows, hindering blood flow and overloading the heart. Another factor is dysfunction of the heart's electrical system, which regulates heartbeats; disturbances in this system, such as severe arrhythmias, can result in inefficient heart function, contributing to the development of CHF. Cardiac hypertrophy, characterized by abnormal enlargement of the heart muscle, also plays a role in CHF, which may occur in response to long-term high blood pressure or other conditions that place significant strain on the heart, resulting in less effective blood pumping [4–6].

Its diagnosis is based on a comprehensive clinical assessment that includes medical history, physical examination, laboratory tests, and imaging exams. During the physical exams, signs of CHF, such as swelling in the ankles and legs, elevated blood pressure, increased heart rate, abnormal heart sounds, and abnormal lung sounds, are sought. Laboratory tests may include a complete blood count, electrolyte analysis, kidney and liver function tests, and cardiac biomarkers. Imaging exams, such as echocardiography, cardiac magnetic resonance imaging, and myocardial scintigraphy, can help assess cardiac function, identify structural heart problems, and determine the stage of CHF [7].

In recent years, machine learning (ML) techniques have been widely used in medicine to assist in the diagnosis, treatment, and prevention of various diseases, including HF. Awan et al. [8], aiming to predict hospital readmissions or deaths within 30 days after HF discharge, proposed a model based on multilayer perceptron (MLP) neural networks, taking into account the common problem of class imbalance in medical data. The study used administrative data from 10,757 HF patients aged over 65 in Western Australia, with 23.6% experiencing adverse outcomes. The MLP model outperformed classical techniques such as logistic regression and random forests. The superior performance was attributed to proper class weight adjustment, which is essential for evaluating predictive models in clinical contexts with significant data imbalance.

In a more recent study, Xie et al. [9] conducted a comprehensive review on the use of artificial intelligence (AI) techniques for the identification and classification of HF. The study highlights that despite advances in traditional methods, the clinical heterogeneity of HF makes accurate assessment of severity and prognosis challenging. AI emerges as a promising alternative by integrating large volumes of clinical, laboratory, and imaging data, with particular emphasis on supervised and unsupervised algorithms applied to medical exams. The results indicate that ML models serve as valuable allies for early diagnosis and risk stratification in HF patients, especially in complex clinical contexts where conventional methods show limitations.

Aiming to assist cardiologists in predicting the progression of congestive heart failure (CHF), Goretti et al. [10] proposed a Deep Learning-based system that uses longitudinal clinical data to predict future CHF severity. The study employed a dataset comprising 1037 records from patients at different disease stages, extracted from hospital visits between 2012 and 2021. The results demonstrated that it is possible to predict CHF clinical progression with good accuracy using a moderately-sized dataset and visit history, which could represent a relevant advancement for therapy personalization and healthcare resource management.

With the objective of predicting mortality in patients with heart failure (HF) and atrial fibrillation (AF), Izraiq et al. [11] used data from the Jordanian Heart Failure Registry, covering 1571 patients, of which 494 had AF. The authors applied multiple machine learning models, including Random Forest, Logistic Regression, SVM, and XGBoost, to estimate mortality risk. The most relevant predictive variables included elevated creatinine levels, length of hospital stay, and need for mechanical ventilation. The results show that ML models can be effective tools for identifying HF patients with concurrent AF who are at higher mortality risk, aiding clinical decision-making and guiding more assertive interventions.

On the other hand, Chicco and Jurman [12] highlighted the use of ML in predicting the survival of patients with CHF using only two out of thirteen attributes from the dataset, serum creatinine and ejection fraction. The study shows that these two attributes are sufficient to predict the survival of CHF patients; however, it does not address the prediction of hospitalization time. Thus, the study demonstrates the usefulness of ML algorithms in predicting the survival of CHF patients and opens the door for further investigations.

ML can be an essential tool for identifying CHF patterns, assisting in diagnostic decision-making, and selecting treatments. ML shows promise in identifying patients at higher risk of developing the disease or facing serious complications and death. Its use in CHF can improve diagnostic accuracy, increase treatment effectiveness, and predict and prevent serious complications [13].

In this work, we will demonstrate the power of ML techniques in predicting the deterioration and death of patients with CHF. Our models and techniques were developed using a dataset collected at the Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad (Punjab, Pakistan) during the period from April to December 2015. This dataset contains information from 299 CHF patients (105 women and 194 men, ages ranging from 40 to 95 years) with 13 clinical and health features. This dataset has been previously used in other ML models dealing with CHF patients [12].

The main objective of this research is to predict patient survival rate and hospitalization time using the most significant variables associated with risk factors, without requiring all features from the original dataset. This is a crucial metric for assessing patient health status, disease severity, treatment complexity, and recovery, even with a reduced amount of clinical data. As a key differentiator, this work employs class balancing techniques to handle imbalanced data distribution and focuses on predicting both patient hospitalization time and survival rate. This approach can assist healthcare professionals in making informed decisions and allocating hospital resources more efficiently. We evaluated six different multiclass classification models to predict hospitalization time and six different binary classification models to predict patient survival. The results indicate that although machine learning strategies and models for sparse data scenarios are available, only classification with a reduced number of variables was able to provide effective clinical predictions, outperforming approaches that use all features from the original dataset.

2 METHODOLOGY

The methodology adopted for clinical patient predictions partially follows the procedure outlined by [12], serving as a comparative reference point. ML models were used to select the most relevant variables (*serum creatinine* and *ejection fraction*) for predicting patient survival, based on the same dataset used in this study. However, the methodological approach of this work expands by including the prediction of hospitalization duration in patients with CHF and the use of the SMOTE technique for class balancing.

The outline of the methodology used can be observed in the flowchart in Figure 1. First, we apply the Synthetic Minority Over-sampling Technique (SMOTE) [14] to balance the classes of our dataset. Then, we apply feature extraction algorithms to select the most relevant variables of the dataset using a mean decrease accuracy technique and a second technique to reduce the Gini impurity of the data. Next, the balanced dataset and the most important variables are used as input for our ML algorithms. For the survival forecast, we split our dataset into two groups. For the evaluation of the survival forecasting models, a 10-fold stratified cross-validation strategy was adopted. This method is more robust than a single train-test split as it ensures every data sample is used for testing exactly once, providing a more stable and reliable estimate of model performance. The stratified nature ensures that the class proportions are maintained in each fold. Performance was quantified using Accuracy, Precision, Recall (Sensitivity), and F1-Score metrics, with their respective means and standard deviations calculated across the 10 folds.

The remainder of this section shows in detail the actions described in Figure 1. First, Section 2.1 presents our dataset. Then, Section 2.2 gives our implementation of the SMOTE technique. Next, Section 2.3 explains our feature extraction algorithms. Finally, Section 2.4 presents the implementation of our ML models. This methodology was computationally evaluated in Section 3.

https://doi.org/10.21528/lnlm-vol23-no2-art2

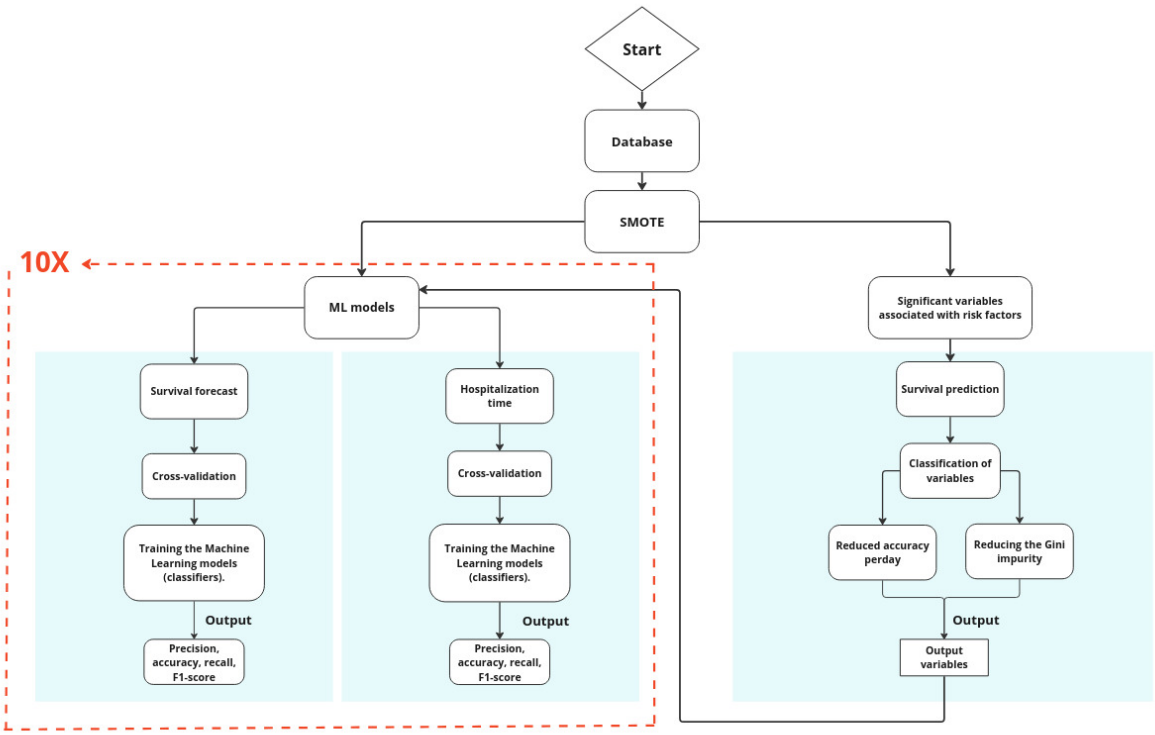

Figura 1: Flowchart of the employed methodology. Source: Authors (2025).

2.1 DATA SET

The dataset[1] analyzed was collected from two medical institutions in Faisalabad, Pakistan, between April and December 2015. It contains information on 299 CHF patients, comprising 105 women and 194 men, aged between 40 and 95 years. All patients had left ventricular systolic dysfunction and diagnosed CHF [15].

The dataset includes 13 features that describe the clinical, bodily, and lifestyle information of the patients, including anemia, high blood pressure, diabetes, gender, and smoking. Some of these features are binary, represented by 0 or 1. For example, a patient is considered anemic if the hematocrit levels are below 36%. However, high blood pressure does not have a clear definition provided in the original article of the dataset.

Some quantitative features include creatine kinase (CPK), ejection fraction, serum creatinine, and sodium. CPK indicates the level of the enzyme in the blood and can be used to identify cardiac injury or failure. Ejection fraction indicates the percentage of blood pumped by the left ventricle with each contraction. Serum creatinine is an indicator of kidney function, while sodium is important for the proper functioning of muscles and nerves.

Although the dataset provides clinically relevant information for predicting outcomes in CHF, it is important to highlight that its sample is relatively small and geographically restricted. This limitation may impact the generalization of developed models, since epidemiological characteristics, comorbidity patterns, and local clinical practices may differ significantly from other regions. Additionally, the patients' age range (40-95 years) is concentrated in older individuals, which may not adequately reflect CHF cases in younger populations. The lack of more detailed information hinders a more precise assessment of this data's representativeness in global contexts. Future studies could enrich the analysis by comparing these parameters with international databases to identify potential biases and calibrate predictive models according to the target population profile.

Given the above, for this specific dataset, the feature *death event*, along with *hospitalization time*, shows significant potential as a target for training AI models, as these two attributes encapsulate crucial information about clinical outcomes and patient progression. The first indicates whether the patient died or survived before the end of the follow-up period, and the latter shows the duration of hospitalization. Unfortunately, there is no information regarding whether any patient had primary kidney disease or the type of follow-up performed. The dataset is imbalanced, with 203 surviving patients and 96 deceased patients, i.e., 67.89% negatives and 32.11% positives.

[1] Available at: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

https://doi.org/10.21528/lnlm-vol23-no2-art2

Tabela 1: Quantitative Description of Categorical variables. Source: Authors (2025).

Categorical Variables	Deceased Patients		Surviving Patients	
	Value	Proportion	Value	Proportion
Anemia (0)	50	52.08	120	59.11
Anemia (1)	46	47.92	3	40.89
Diabetes (0)	56	58.33	118	58.13
Diabetes (1)	40	41.67	85	41.87
High Blood Pressure (0)	57	59.38	137	67.49
High Blood Pressure (1)	39	40.62	66	32.51
Sex (0: female)	34	35.42	71	34.98
Sex (1: male)	62	64.58	132	65.02
Smoking (0)	66	68.75	137	67.49
Smoking (1)	30	31.23	66	32.51

Initially, the descriptive statistical analysis of the data provides an overall idea of the distribution. As seen in Table 1, we have a descriptive statistical analysis of the categorical characteristics (or variables) of the patients, dividing them into groups of patients who survived and those who did not. It presents binary characteristics such as high blood pressure, whether the patient has hypertension, anemia (indicating a decrease in red blood cells or hemoglobin), whether the patient has diabetes, whether the patient smokes or not, and gender. It details the distribution of characteristics such as anemia, high blood pressure, diabetes, gender, and smoking, showing the number and percentage of patients in each category, both in the total group and in the subgroups of patients who survived and those who died. This allows us to observe how these categorical conditions are distributed among the patients and their potential relationship with survival outcomes.

Table 2 provides a quantitative statistical description of the numerical characteristics of the study, divided between the sample of deceased patients and surviving patients. It highlights quantitative characteristics such as the patient's age (in years), creatine phosphokinase (CPK), which indicates the level of CPK enzyme in the blood (mcg/L), ejection fraction, which indicates the percentage of blood that leaves the heart with each contraction, blood platelets, serum creatinine level in the blood (mg/dL), serum sodium level in the blood (mEq/L), and follow-up period (in days).

Tabela 2: Quantitative description of numerical variables. Source: Authors (2025).

Variable	Deceased Patients			Surviving Patients		
	Mean	Median	σ	Mean	Median	σ
Age (years)	65.22	65.00	13.21	60.00	58.76	10.64
CPK (mcg/L)	670.20	259.00	1316.58	540.10	245.00	753.80
Ejection Fraction (%)	33.47	30.00	12.53	40.27	38.00	10.86
Platelets (k plaq/AM)	256.38	258.50	98.53	266.66	263.00	97.53
Serum Creatinine (mg/dL)	1.84	1.30	1.47	1.19	1.00	0.65
Sodium (mEq/L)	135.40	135.50	5.00	137.20	137.00	3.98
Time (days)	70.89	44.50	62.38	158.30	172.00	67.74

By analyzing the data in Table 2, we compare the means and medians of the variables to check for similarities in the results. It was observed that for the variables *ejection fraction*, *age*, and *serum sodium level*, there is a small difference between the mean and the median, indicating that the data distribution is symmetric. However, for the variables *CPK* and *time*, the difference is significant, meaning the data distribution is asymmetric, and consequently, the mean and median differ.

2.2 SMOTE

Upon analyzing the variables *death* and *month*, it was found that they were imbalanced, meaning there was a disproportionate number of observations for each of their respective values. To correct this imbalance, a class balancing technique is needed. which is a class-balancing technique where new synthetic observations are created for the minority class using linear combinations of existing observations.

We employed SMOTE as a class-balancing algorithm. It is an advanced data preprocessing technique used to address class imbalance in ML datasets. The main goal of SMOTE is to create a balance between minority and majority classes by artificially increasing the number of observations in the minority class through the generation of synthetic instances

The SMOTE procedure is carried out by selecting observations from the minority class and generating new synthetic instances that are interpolated between the selected observations and their nearest neighbors. In this way, it ensures that the feature space of

the minority class is better represented, providing a more balanced dataset. This technique differs from traditional oversampling, which simply replicates the minority class observations, potentially leading to overfitting of the model

The selection of the nearest neighbors is done based on a distance metric, typically Euclidean distance, where each example from the minority class is considered along with its k nearest neighbors. For each selected example, a synthetic instance is created by randomly choosing one of the k neighbors and interpolating between the example and the selected neighbor.

The importance of SMOTE lies in its ability to improve the performance of ML models in classification tasks with class imbalance. It has been shown that models trained on SMOTE-balanced datasets tend to exhibit better sensitivity, specificity, and accuracy compared to models trained on imbalanced datasets. In the present study, the SMOTE technique was employed to balance the variable *months*, a multiclass variable representing the number of months the patient was hospitalized.

Despite its advantages, SMOTE is not free from limitations. The generation of synthetic instances can introduce noise if the minority class observations are scattered or if there is overlap between classes. In some cases, this approach may fail to capture the true distribution and characteristics of the minority class, producing observations that do not adequately represent its patterns. This discrepancy between theory and practice can result in ML models that show high efficacy in controlled or testing environments but do not maintain the same level of performance when applied in real-world situations [16].

Despite the challenges associated with it, techniques like SMOTE are of great value for imbalanced datasets. They offer a way to improve the representativeness of the minority class, enabling ML models to learn more complex and subtle patterns that may be overlooked in imbalanced datasets. When applied cautiously and complemented by rigorous evaluation, these techniques can significantly enhance the effectiveness of models in practical applications, making them more robust and adaptable to different scenarios [14].

2.3 FEATURE SELECTION

Feature selection in ML is a fundamental process aimed at identifying and selecting the most relevant features for building effective predictive models. This process uses fewer features while optimizing results. Among the various techniques employed for this purpose, Mean Decrease Accuracy (MDA) and Gini Impurity stand out for their unique approaches to evaluating feature importance. Both techniques offer methods for quantifying the impact of each feature on model performance, enabling more informed feature selection [17].

The MDA technique works by evaluating the impact of each variable or feature on the model's accuracy. A systematic alteration of a feature's values is made while keeping the other attributes constant to measure how this change affects the model's accuracy. This procedure is repeated for each feature in the dataset. The underlying idea is that if altering a feature's values leads to a significant decrease in the model's accuracy, the feature is considered important. On the other hand, if changing a feature's values does not significantly affect accuracy, the feature may be considered less relevant. The MDA technique is particularly useful for identifying features that positively contribute to model performance, allowing for effective dimensionality reduction without significant information loss.

On the other hand, Gini Impurity is a metric primarily used in decision trees and ensemble models, such as Random Forest, to assess feature importance. Gini Impurity measures the frequency with which a random element from the set would be misclassified if it were randomly labeled according to the distribution of labels in the split. Features that produce splits with low Gini impurity are considered important, as they indicate that the feature is effective in separating the classes. During the training of a decision tree, the reduction in Gini impurity resulting from a split on a particular feature is used to assess that feature's importance. Thus, features that result in larger reductions in impurity are valued more highly.

Both techniques, MDA and Gini Impurity, provide valuable insights into the relevance of features for building predictive models. While MDA focuses on the direct impact of features on model accuracy, Gini Impurity evaluates the ability of features to effectively separate the classes in the dataset. The choice between these techniques depends on the type of model to be built and the specific objectives of the analysis.

2.4 ML MODELS

This work employed several supervised ML models to predict CHF patients' survivability and hospitalization time. The objective of the survivability models was to predict whether the patient would survive or not. Therefore, we constructed eight different supervised binary classification models to predict this outcome: a decision tree, a random forest, a gradient boosting, a naive Bayes algorithm, a k-nearest neighbours (KNN), a multi-layer perceptron neural network (MLP), a logistic regression, and an ensemble of the algorithms mentioned above [18–23]. The ensemble gave the same weight to all ML algorithms employed.

The dataset described the hospitalization time of CHF patients as a number of days by an integer variable. We opted to use a data bucketing preprocessing method to group the days of the dataset into months, whereas one month corresponds to a total of 30 days. Thus, we have that month 0 corresponds to a period of less than 30 days, month 1 corresponds to a period between 30 and 60 days, month 2 corresponds to a period between 60 and 90 days, and so on. Therefore, we replaced the given hospitalization time in days with the corresponding hospitalization time in months. Using this new data, we developed ML models to predict the hospitalization time as described below.

The objective of hospitalization time models is to predict the number of months the patients will be hospitalized. It considers the number of months as different classes of the data instead of a continuous value. Therefore, we could use classification algorithms similar to those developed to predict the patient's survivability. We constructed six different ML models for predicting

patients' hospitalization time: a decision tree, a random forest, a logistic regression, a naive Bayes algorithm, a gradient boosting, and an ensemble of the algorithms.

2.4.1 HYPERPARAMETER OPTIMIZATION

As per the evaluation methodology outlined in Figure 1, hyperparameter optimization was integrated directly into the 10-fold cross-validation pipeline. For each Machine Learning algorithm, a random search strategy [24] was employed to find the most effective combination of parameters.

Specifically, the random search process was configured to evaluate a total of 100 distinct sets of hyperparameters for every model. This search was executed within each of the 10 folds of the main cross-validation. In each iteration, the training portion (corresponding to 90% of the data) was used by the random search to determine the optimal parameter settings through its own internal cross-validation. Subsequently, the model, configured with these optimized hyperparameters, was evaluated on the hold-out test fold (the remaining 10%). This approach ensures that the optimization is performed without exposing the test set of each iteration, providing a robust and unbiased performance estimate for the tuned models.

3 COMPUTATIONAL EXPERIMENTS

For the development of this study, the Python programming language (version 3.11.5) was utilized, along with specialized libraries such as scikit-learn (version 1.2.2). To ensure the replicability and consistency of all experiments, the *random_state* parameter was globally set to 42. To address the class imbalance in the original dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied prior to model training.

A robust 10-fold stratified cross-validation strategy was employed for model evaluation instead of a single data split. This method ensures that each data point is used for testing exactly once, providing a more reliable estimate of generalization performance. The stratified nature of the folds guaranteed that the proportion of classes (patient survival or death) was maintained consistently across each training and testing iteration.

Model performance was quantified using four standard classification metrics: Accuracy, Precision, Recall (Sensitivity), and F1-Score. To determine if the observed differences in these metrics were statistically significant, the Wilcoxon signed-rank test was utilized. This non-parametric statistical test was applied to compare the performance between the best (Random Forest) and worst (MLP) performing models, providing statistical evidence for their ranking.

Furthermore, a detailed error analysis was conducted using aggregated confusion matrices generated via the *cross_val_predict* method. This technique provides a comprehensive view of each model's predictive behavior by summing the outcomes (True Positives, False Positives, etc.) across all 10 folds of the cross-validation. This consolidated analysis offers a more stable and complete picture of the models' classification errors than a matrix from a single test set.

3.1 PREDICTING PATIENT SURVIVABILITY

The methods presented in Section 2 were used to predict patient survival. Each binary classification model for the variable Death was applied, and the scores are reported in Table 3. Table 3 presents the classification models and their respective results for accuracy, precision, sensitivity, and F1 score.

Tabela 3: Results of ML Models for Survival Prediction. Source: Authors (2025).

Model	Accuracy	F1-Score	Precision	Recall
Decision Tree	0.73 (±0.06)	0.73 (±0.06)	0.74 (±0.06)	0.73 (±0.06)
Gradient Boosting	0.80 (±0.04)	0.80 (±0.04)	0.80 (±0.05)	0.80 (±0.04)
Random Forest	0.80 (±0.05)	0.80 (±0.05)	0.81 (±0.05)	0.80 (±0.05)
Naive Bayes	0.69 (±0.05)	0.69 (±0.05)	0.70 (±0.05)	0.69 (±0.05)
Logistic Regression	0.74 (±0.07)	0.74 (±0.07)	0.75 (±0.07)	0.74 (±0.07)
K-NN	0.71 (±0.05)	0.71 (±0.05)	0.71 (±0.05)	0.71 (±0.05)
MLP	0.67 (±0.07)	0.66 (±0.07)	0.68 (±0.07)	0.67 (±0.07)
Ensemble	0.75 (±0.06)	0.75 (±0.06)	0.75 (±0.06)	0.75 (±0.06)

The Random Forest model demonstrated high efficiency in prediction. It achieved 80% accuracy, correctly classifying 80% of the samples. Precision was 81%, while recall was 80%, indicating that it correctly identified 80% of the positive samples and 80% of the positive observations in the test set, respectively. The F1-score, 80%, confirms the model's balanced ability to identify both positive and negative samples.

https://doi.org/10.21528/lnlm-vol23-no2-art2

On the other hand, the model with the poorest performance was the MLP, which correctly classified 67% of the samples (accuracy). The precision was 68%, and the recall and F1-score were 67% and 66%, respectively. The Ensemble model also performed well, as expected, with an accuracy of 75%, precision of 75%, recall of 75%, and an F1-score of 75%. These results indicate that the Ensemble model is a good option for the prediction task.

For a detailed error analysis, aggregated confusion matrices were generated from the predictions of the 10-fold cross-validation on the dataset previously balanced with the SMOTE technique (Figure 2). The Random Forest model demonstrated a sensitivity (recall) of 79.8%, correctly identifying 162 out of 203 death events, with 41 False Negatives. Its specificity was 80.3%. In direct comparison, the MLP model presented a markedly lower sensitivity of 56.2%, resulting in more than double the number of False Negatives (89). The comparative analysis highlights that the main distinction between the models lies in the detection capability of the positive class (death). The substantial reduction of False Negatives by the Random Forest in relation to the MLP justifies, based on this error analysis, its selection as the superior performing model for mortality prediction in this study.

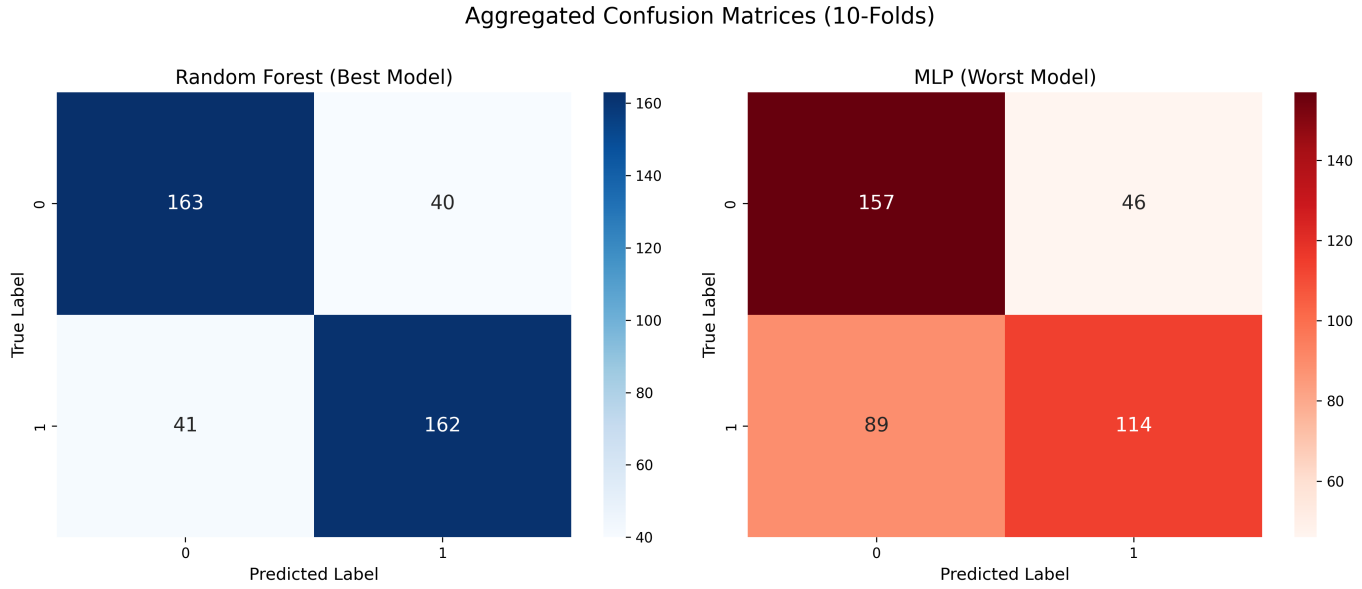

Figura 2: Confusion matrix of the iteration closest to the mean of the Random Forest and MLP model. Source: Authors (2025).

The results highlight that the success of the Random Forest model, based on decision trees and ensemble learning, can be attributed to its ability to combine predictions from multiple trees, reducing overfitting and increasing the stability of the predictions. In contrast, the somewhat lower performance of the MLP model, with the discrepancy between high precision and low recall, suggests that the model is too conservative in its predictions, meaning it prefers not to label an instance as positive unless it is very certain, resulting in many false negatives.

To validate that these performance differences were statistically significant, a focused analysis was conducted comparing the best-performing model (Random Forest model) with the worst-performing one (MLP model). The Wilcoxon signed-rank test for paired samples was applied to each of the four main metrics: Accuracy, F1-Score, Precision, and Recall. The results were conclusive, indicating a statistically significant superiority of the Random Forest over the MLP across all evaluated metrics ($p < 0.05$ for all tests). This rigorous analysis confirms that the performance difference between the models is not due to chance, solidifying Random Forest as the most effective approach and MLP as the least suitable among those tested in this work.

To evaluate the performance of the models employed in this study, a comparison was made between the results obtained and those from the work of [12], in which similar classification models such as Random Forests, KNN, and Naive Bayes were used.

Tabela 4: Comparison of the results of AM models for survival prediction. Source: Authors (2025).

Models	This Study		Reference Study [12]	
	Accuracy	F1-score	Accuracy	F1-score
Random Forests	0.80	0.80	0.74	0.54
Decision Tree	0.73	0.73	0.73	0.55
Gradient Boosting	0.80	0.80	0.73	0.52
Naive Bayes	0.69	0.66	0.62	0.14
KNN	0.71	0.71	0.69	0.36

https://doi.org/10.21528/lnlm-vol23-no2-art2

The results of this comparison are presented in Table 4, and it was observed that significantly better results were presented by this work compared to the reference study. Based on the performance metrics, although there were variations between the models, all showed promising results. These results suggest that the proposed tool could be considerably useful in this situation for predicting patient survival.

3.2 PREDICTING PATIENTS HOSPITALIZATION TIME

As in the previous subsection, models were used to predict the hospitalization time of patients. For this analysis, the Binning technique was initially applied, a process of dividing a continuous variable into intervals or bins, and then replacing them with a representative value or category for each bin. In this way, the numerical variable Time (days) was transformed into a multivariate variable Month, which grouped the data into different value ranges. Subsequently, the variable was balanced using the SMOTE method to avoid the imbalance of the minority class (Figure3).

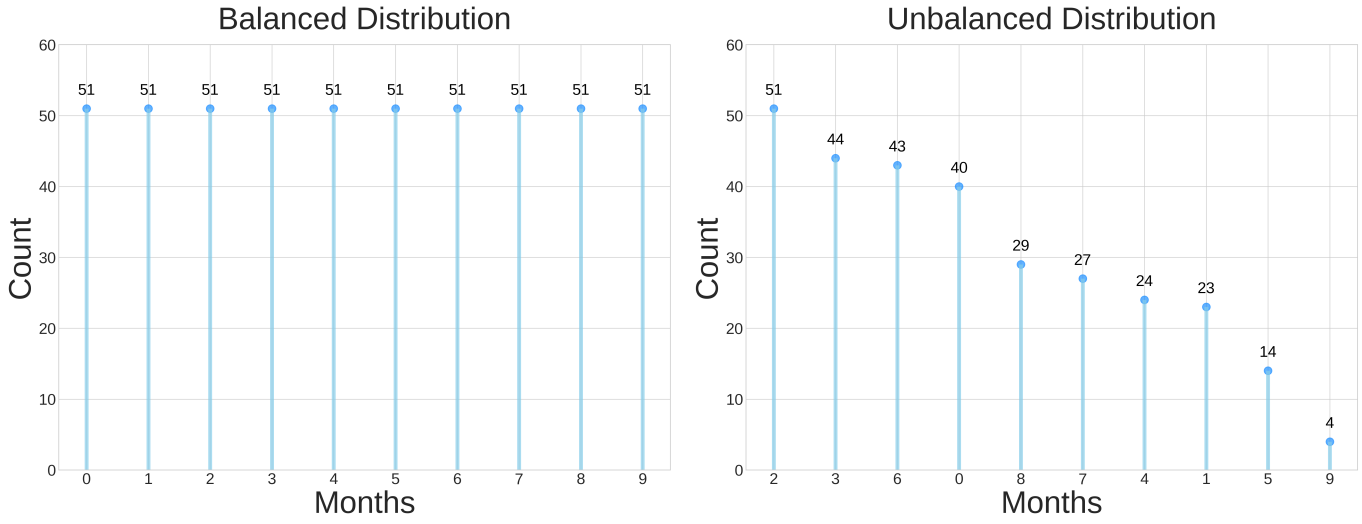

Figura 3: Class distribution before (right) and after (left) the data balancing process. Source: Authors (2025).

The multiclass classification results for hospitalization time are presented in Table 5. The Random Forest model showed the best overall performance, achieving a weighted F1-Score of 0.4873 and an accuracy of 0.5000. Gradient Boosting was the second most effective model (F1-Score: 0.4211), confirming that tree-based ensemble approaches were superior.

Tabela 5: Results of AM models for predicting hospitalization time. Source: Authors (2025).

Model	Accuracy	F1-Score	Precision	Recall
Decision Tree	0.37 (±0.06)	0.36 (±0.06)	0.37 (±0.06)	0.37 (±0.06)
Gradient Boosting	0.43 (±0.07)	0.42 (±0.06)	0.43 (±0.06)	0.43 (±0.07)
Random Forest	0.50 (±0.07)	0.49 (±0.07)	0.50 (±0.07)	0.50 (±0.07)
Naive Bayes	0.25 (±0.05)	0.22 (±0.05)	0.27 (±0.08)	0.25 (±0.05)
Logistic Regression	0.27 (±0.04)	0.24 (±0.04)	0.23 (±0.07)	0.27 (±0.04)
K-NN	0.35 (±0.06)	0.33 (±0.06)	0.34 (±0.07)	0.35 (±0.06)
MLP	0.18 (±0.05)	0.14 (±0.05)	0.13 (±0.04)	0.18 (±0.05)
Ensemble	0.37 (±0.06)	0.36 (±0.05)	0.37 (±0.06)	0.37 (±0.06)

In contrast, the Voting Classifier (Ensemble), which combined three models, obtained an F1-Score of 0.3596, failing to outperform its best individual components (Random Forest and Decision Tree). Models such as MLP (F1-Score: 0.1394) and Naive Bayes (F1-Score: 0.2232) recorded the lowest metrics.

Overall, the metric values, with a ceiling of 0.4873 for the F1-Score in a ten-class problem, indicate that predicting the length of stay with the available features is a task of high complexity.

3.3 FEATURE SELECTION

Initially, the technique of Average Precision Reduction was applied, which evaluates the importance of features based on the *Random Forest* model for survival classification. This evaluation can be visualized in Figure 4 (on the left). Features with lower importance were discarded until the desired number of features was reached or the specified threshold was met. The Gini Impurity Reduction technique was also applied, which is a measure of impurity or uncertainty in a dataset, aiming to assess the

https://doi.org/10.21528/lnlm-vol23-no2-art2

relative importance of each feature in the model and then discard those with minimal importance, as shown in Figure 4 (on the right).

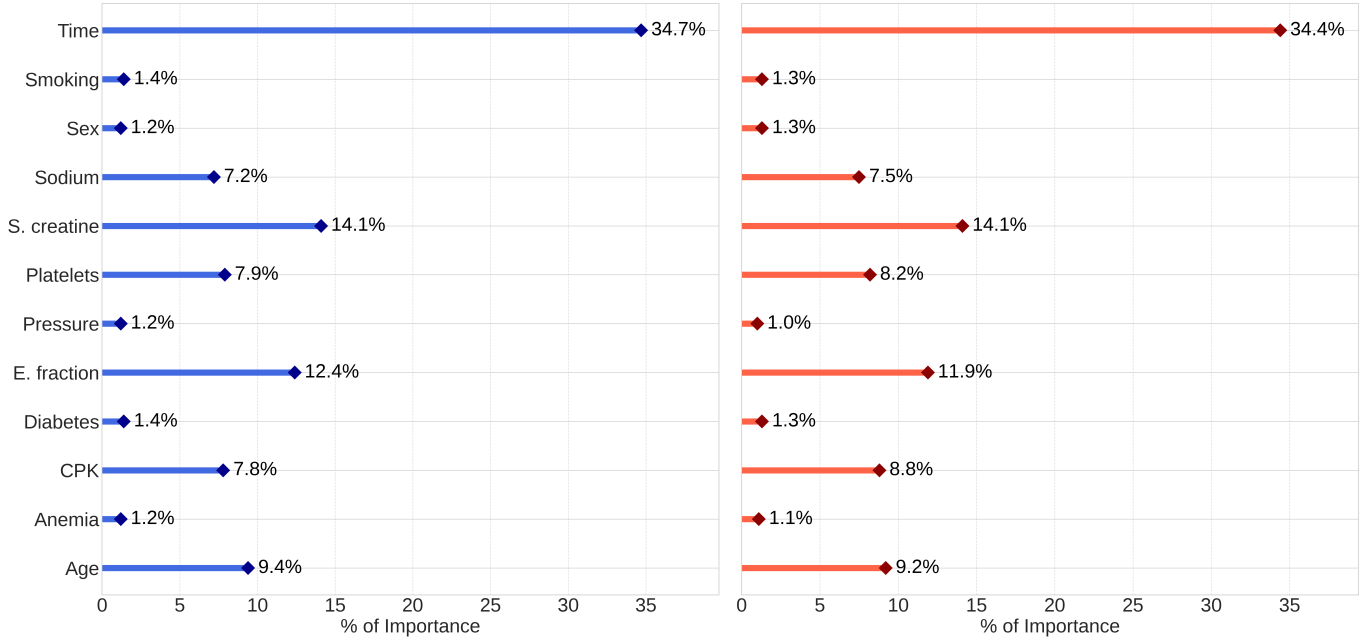

Figura 4: Percentage of the importance of the resource in relation to patient survival. On the left, average reduction in accuracy and, on the right, a reduction in Gini impurity. Source: Authors (2025).

In Figure 4, it is possible to observe the feature selection using two distinct methods: average precision reduction and Gini impurity reduction. For the first method, a threshold of 0.1 importance was defined, which led to the selection of only the features with importance equal to or greater than 10%. As a result, three features were selected: *ejection fraction*, *serum creatinine*, and *Time*. These variables were considered the most relevant for the classification task, as they had the greatest impact on predicting the clinical outcome.

The choice of the 10% threshold was not arbitrary, but rather based on the methodology adopted by Chicco and Jurman (2020), who, although they did not use a fixed cutoff, consistently identified the variables *ejection fraction* and *serum creatinine* as the most relevant ones by applying several ranking techniques. Thus, the use of a percentage threshold was adopted in this study as a practical strategy to reproduce the selection performed in the reference study, considering the variable *Time*, which is one of the focuses of this research, enabling a direct comparison of results and maintaining consistency with the literature.

In the second method, Gini impurity reduction, results similar to those of the average precision reduction method were obtained. The same three features were selected: *ejection fraction*, *serum creatinine*, and *Time*. These variables presented the highest importance percentages, being considered the most relevant for the classification task. This selection occurred because the information provided by these features has the potential to directly influence the prediction of the clinical outcome for CHF patients.

An investigation into the importance of the variables for the multivariate class of hospitalization time was also conducted. For this purpose, the average precision reduction (with the same 10% importance threshold) and Gini impurity techniques were applied to classify the follow-up time, as presented in Figure 5.

In both cases, six features were selected: *age*, *CPK*, *ejection fraction*, *platelets*, *serum creatinine*, and *sodium*. It was observed that there were more relevant features for the classification of hospitalization time than for the classification of survival.

Among the predictive models for clinical outcomes in heart failure, the Seattle Heart Failure Model (SHFM) and the Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) score stand out, both widely used for risk stratification across different patient profiles. Based on the analysis of these scores, it is observed that *ejection fraction* and *age* are included in both models, highlighting their prognostic relevance. *Serum creatinine* is explicitly included in the MAGGIC score, whereas the SHFM, although not listing it directly among its 14 original components, considers *renal function* as a determining factor. Finally, variables such as *creatine phosphokinase (CPK)* and *platelet count* are not considered by either model [25].

https://doi.org/10.21528/lnlm-vol23-no2-art2

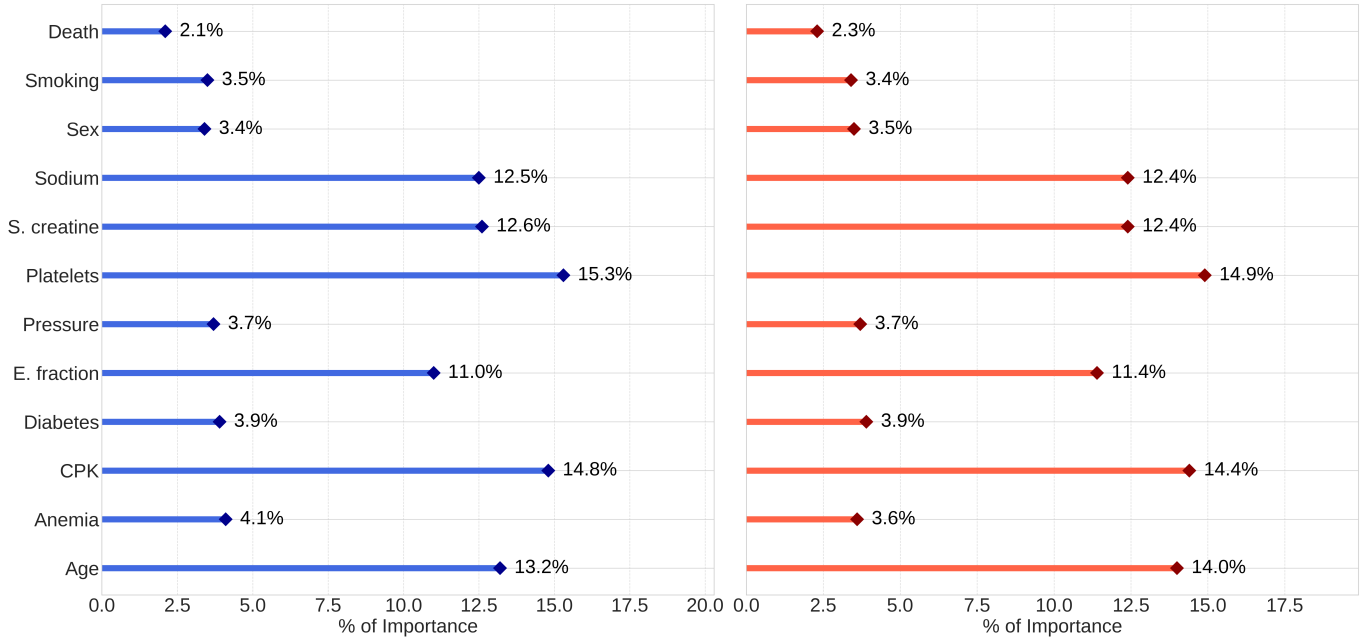

Figura 5: Percentage of importance of the resource in relation to the length of stay. On the left, average reduction in accuracy and, on the right, reduction in Gini impurity. Source: Authors (2025).

3.4 CLINICAL PREDICTION USING ONLY CLASSIFIED CHARACTERISTICS

For the prediction of survival in patients with CHF, using the selected features, specifically *ejection fraction*, *serum creatinine*, and *time*, the results were compared with those obtained in the work of [12], where serum creatinine and ejection fraction were used, excluding the length of hospitalization, which was also performed in this study. Each binary classification method (Random Forests and Gradient Boosting) used in the work of [12] was applied, and the results are presented in Table 6.

Tabela 6: Comparison of the results of AM models for survival prediction. Source: Authors (2025).

Models	This Study		Reference Study [12]	
	Accuracy	F1-score	Accuracy	F1-score
Random Forests	0.73	0.73	0.58	0.75
Gradient Boosting	0.71	0.72	0.58	0.75

The results, using the selected variables ejection fraction and serum creatinine, indicate that the *Random Forest* and *Gradient Boosting* models, with the balancing of the ”Death”variable, were the most efficient in the prediction task analyzing the F1-Score metric, when compared to the reference study [12]. With a remarkable F1-score of 72% and 73% for the Random Forest and Gradient Boosting respectively, these models stood out in terms of efficiency.

The results obtained in this study suggest that age, CPK, ejection fraction, platelets, sodium, serum creatinine and time are important variables for predicting survival and length of hospital stay in patients with CHF. Ejection fraction and serum creatinine are the most relevant in this prediction task.

4 CONCLUSION

This study demonstrated the application of machine learning models on two critical fronts in the management of patients with congestive heart failure (CHF): predicting survival and forecasting the length of hospitalization.

For survival prediction, the models yielded robust results, where the SMOTE balancing technique proved effective in enhancing predictive performance. Feature selection enabled the identification of key risk factors, reinforcing the approach’s potential as an accurate clinical decision support tool for patient stratification.

Forecasting the length of hospitalization, an intrinsically more complex multiclass task, proved to be a significant challenge. Although ensemble models like Random Forest outperformed others, the modest overall performance indicates that predicting the exact duration of hospitalization with high accuracy remains a barrier to overcome. This finding does not diminish the study’s value; rather, it quantifies the problem’s difficulty and highlights the most promising modeling approaches.

In conclusion, the ML approach is a tool with great potential. The ability to predict survival already points toward a tangible aid in personalizing treatments. The prediction of hospitalization length, while at a more exploratory stage, establishes a foundation for future work that could lead to the optimization of hospital resources. For both applications, external validation across diverse, multi-center cohorts is a crucial next step to mitigate biases and ensure the generalizability of the models before their integration into clinical practice.

ACKNOWLEDGMENT

This study was financed in part by the CAPES - Finance Code 001.

REFERENCES

[1] B. A. Borlaug and M. M. Redfield. "Diastolic and systolic heart failure are distinct phenotypes within the heart failure spectrum". *Circulation*, vol. 123, no. 18, pp. 2006–2014, 2011.

[2] T. A. McDonagh, R. S. Gardner, A. L. Clark and H. Dargie. *Oxford textbook of heart failure*. Oxford University Press, 2011.

[3] N. C. C. for Chronic Conditions (Great Britain). *Chronic Heart Failure: national clinical guideline for diagnosis and management in primary and secondary care*. Royal College of Physicians, 2003.

[4] R. F. Lee, T. K. Glenn and S. S. Lee. "Cardiac dysfunction in cirrhosis". *Best Practice & Research Clinical Gastroenterology*, vol. 21, no. 1, pp. 125–140, 2007.

[5] C. M. Otto and B. Prendergast. "Aortic-valve stenosis—from patients at risk to severe valve obstruction". *New England Journal of Medicine*, vol. 371, no. 8, pp. 744–756, 2014.

[6] I. Shimizu and T. Minamino. "Physiological and pathological cardiac hypertrophy". *Journal of molecular and cellular cardiology*, vol. 97, pp. 245–262, 2016.

[7] M. Gomberg-Maitland, D. A. Baran and V. Fuster. "Treatment of congestive heart failure: guidelines for the primary care physician and the heart failure specialist". *Archives of Internal Medicine*, vol. 161, no. 3, pp. 342–352, 2001.

[8] S. E. Awan, M. Bennamoun, F. Sohel, F. M. Sanfilippo and G. Dwivedi. "Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics". *ESC heart failure*, vol. 6, no. 2, pp. 428–435, 2019.

[9] Y. Xie, L. Zhang, W. Sun, Y. Zhu, Z. Zhang, L. Chen, M. Xie and L. Zhang. "Artificial Intelligence in Diagnosis of Heart Failure". *Journal of the American Heart Association*, vol. 14, no. 8, pp. e039511, 2025.

[10] F. Goretti, B. Oronti, M. Milli and E. Iadanza. "Deep learning for predicting congestive heart failure". *Electronics*, vol. 11, no. 23, pp. 3996, 2022.

[11] M. Izraiq, R. I. Alawaisheh, I. Hamam, M. Hajjiri, I. K. Jarrad, Q. Albustanji, Y. B. Ahmed, O. A. Abu-Dhaim, I. Zuraik, A. A. Toubasi *et al.*. "Machine Learning-Driven Mortality Prediction in Heart Failure Patients with Atrial Fibrillation: Evidence from the Jordanian Heart Failure Registry". *Research Reports in Clinical Cardiology*, pp. 35–44, 2024.

[12] D. Chicco and G. Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". *BMC Medical Informatics and Decision Making*, vol. 20, pp. 1–16, 2020.

[13] V. Jahmunah, S. L. Oh, J. K. E. Wei, E. J. Ciaccio, K. Chua, T. R. San and U. R. Acharya. "Computer-aided diagnosis of congestive heart failure using ECG signals–A review". *Physica Medica*, vol. 62, pp. 95–104, 2019.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[15] T. Ahmad, A. Munir, S. Bhatti, M. Aftab and M. Raz. "Heart Failure Clinical Records Dataset". UCI Machine Learning Repository, 2020.

[16] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh and A. Almuhaimeed. "Stop oversampling for class imbalance learning: A review". *IEEE Access*, vol. 10, pp. 47643–47660, 2022.

[17] H. Han, X. Guo and H. Yu. "Variable selection using mean decrease accuracy and mean decrease gini based on random forest". In *2016 7th IEEE Int. Conf. Software Eng. Service Sci. (ICSESS)*, pp. 219–224. IEEE, 2016.

[18] A. DeMaris. "A tutorial in logistic regression". *Journal of Marriage and the Family*, pp. 956–968, 1995.

[19] J. R. Quinlan. "Induction of decision trees". *Machine Learning*, vol. 1, pp. 81–106, 1986.

[20] L. Breiman. “Random forests”. *Machine Learning*, vol. 45, pp. 5–32, 2001.

[21] J. H. Friedman. “Greedy function approximation: a gradient boosting machine”. *Annals of Statistics*, pp. 1189–1232, 2001.

[22] E. Fontana. “IntroduÃ§Ã£o aos algoritmos de aprendizagem supervisionada”. Technical report, Departamento de Engenharia QuÃmica, Universidade Federal do ParanÃ¡, 2020.

[23] G. E. Hinton. “Connectionist learning procedures”. In *Machine Learning*, pp. 555–610. Elsevier, 1990.

[24] J. Bergstra and Y. Bengio. “Random search for hyper-parameter optimization”. *Journal of Machine Learning Research*, vol. 13, no. 2, 2012.

[25] J. D. Rich, J. Burns, B. H. Freed, M. S. Maurer, D. Burkhoff and S. J. Shah. “Meta-Analysis Global Group in Chronic (MAGGIC) heart failure risk score: validation of a simple tool for the prediction of morbidity and mortality in heart failure with preserved ejection fraction”. *Journal of the American Heart Association*, vol. 7, no. 20, pp. e009594, 2018.