JHI JOURNAL OF HEALTH INFORMATICS

Prediction of sudden cardiac death for chagasic patients

Predição de morte súbita cardíaca em pacientes chagásicos

Predicción de muerte súbita cardíaca en pacientes chagásicos federados

Pedro Ernesto de Oliveira Primo[1], Weslley Lioba Caldas[2], João Paulo do Vale Madeiro[3], Danielo Gonçalves Gomes[4], Gabriel Suassuna Almeida[5], Luan Pereira de Lima Brasil[6], Carlos Henrique Leitão Cavalcante[7], Roberto Coury Pedrosa[8]

Keywords: Chagas Disease; Electrocardiography; Machine Learning

ABSTRACT

Objective: Identify the risk of patients with Chronic Chagas Cardiomyopathy (CCC) to prevent them from having Sudden Cardiac Death (SCD). **Methods**: We developed an SCD prediction system using a heterogeneous dataset of chagasic patients evaluated in 9 state-of-the-art machine learning algorithms to select the most critical clinical variables and predict SCD in chagasic patients even when the interval between the most recent exams and the SCD event is months or years. **Results**: 310 patients were analyzed, being 81 (14,7%) suffering from SCD. In the study, Balanced Random Forest showed the best performance, with AUC:80.03 and F1:75.12. Due to their high weights in the machine learning classifiers, we suggest Holter - Non-Sustained Ventricular Tachycardia, Total Ventricular Extrasystoles, Left Ventricular Systolic Diameter, Syncope, and Left Ventricular Diastolic Diameter as essential features to identify SCD. **Conclusion**: The high-risk pattern of SCD in patients with CCC can be identified and prevented based on clinical and laboratory variables.

Descritores: Doença de Chagas; Eletrocardiografia; Aprendizado de Máquina

RESUMO

Objetivo: Identificar o risco de pacientes com Cardiomiopatia Chagásica Crônica (CCC) para prevenir a Morte Súbita Cardíaca (MSC). **Métodos:** Desenvolvemos um sistema de MSC usando um conjunto de dados heterogêneo de pacientes chagásicos avaliados em 9 algoritmos de aprendizado de máquina de última geração para selecionar as variáveis clínicas mais críticas e prever MSC em pacientes chagásicos mesmo quando o intervalo mais recente entre os mais recentes exames e o evento MSC é meses ou anos. **Resultados:** Foram analisados 310 pacientes, sendo 81 (14,7%) portadores de CCC. No estudo, o algoritmo Balanced Random Forest apresentou o melhor desempenho, com AUC:80,03 e F1:75,12. Devido ao seu alto peso nos classificadores de aprendizado de máquina, sugerimos Holter - Taquicardia Ventricular Não Sustentada, Extrassístoles Ventriculares Totais, Diâmetro Sistólico do Ventrículo Esquerdo, Síncope e Diâmetro Diastólico do Ventrículo Esquerdo como características essenciais para identificar a CCC. **Conclusão:** O padrão de alto risco de MSC em pacientes com CCC pode ser identificado e prevenido com base em variáveis clínicas e laboratoriais.

Descriptores: Enfermedad de Chagas; Electrocardiografía; Aprendizaje Automático

RESUMEN

Objetivo: Identificar el riesgo de los pacientes con Miocardiopatía Chagásica Crónica (MCC) para evitar que presenten Muerte Cardíaca Súbita (MCS). **Métodos:** Desarrollamos un sistema MCS utilizando un conjunto de datos heterogéneo de pacientes chagásicos evaluados en 9 algoritmos de aprendizaje automático de última generación para seleccionar las variables clínicas más críticas y predecir MCS en pacientes chagásicos incluso cuando el intervalo más reciente entre los más recientes exámenes y el evento MCS es meses o años. **Resultados:** Se analizaron 310 pacientes, siendo 81 (14,7%) con MSC. En el estudio, Balanced Random Forest mostró el mejor desempeño, con AUC:80.03 y F1:75.12. Debido a su alto peso en los clasificadores de aprendizaje automático, sugerimos Holter - Taquicardia ventricular no sostenida, Extrasístoles ventriculares totales, Diámetro sistólico del ventrículo izquierdo, Síncope y Diámetro diastólico del ventrículo izquierdo como características esenciales para identificar la MSC. **Conclusión:** El patrón de alto riesgo de MSC en pacientes con MCC se puede identificar y prevenir con base en variables clínicas y de laboratorio.

[1] *Graduando em Ciência da Computação, Universidade Federal do Ceará - UFC, Fortaleza (CE), Brasil.*
[2] *Doutorado em Ciência da Computação, Universidade Federal do Ceará - UFC, Fortaleza (CE), Brasil.*
[3] *Doutor em Engenharia de Teleinformática, Departamento de Computação, Universidade Federal do Ceará - UFC, Fortaleza(CE), Brasil.*
[4] *Professor Associado do Departamento de Engenharia de Teleinformática, Universidade Federal do Ceará - UFC, Fortaleza (CE), Brasil.*
[5] *Graduando em Ciência da Computação, Universidade Federal do Ceará - UFC, Fortaleza (CE), Brasil.*
[6] *Graduando em Engenharia de Computação,Universidade Federal do Ceará - UFC, Fortaleza (CE), Brasil.*
[7] *Doutorado em Biotecnologia - RENORBIO, Universidade Estadual do Ceará - UECE, Fortaleza (CE), Brasil. Professor Efetivo Instituto Federal do Ceará - IFCE, Fortaleza (CE), Brasil.*
[8] *Professor Permanente da Pós-Graduação do Instituto do Coração Edson Saad, Universidade Federal do Rio de Janeiro - UFRJ, Rio de Janeiro (RJ), Brasil.*

Autor Coorespondente: **Pedro Ernesto de Oliveira Primo**
e-mail: **pedroernesto2406@alu.ufc.br**

Artigo recebido: **30/09/2021**
Aprovado: **15/03/2022**

INTRODUCTION

Chagas Disease, also known as American trypanosomiasis, occurs in more than 21 countries, mainly in Latin America (Brazil and Mexico)[1-3]. The number of infected is around 6 to 7 million people[4], and according to the World Health Organization (WHO), 30% of patients develop some kind of cardiomyopathy (CCC).

CCC is considered an arrhythmogenic condition because of the presence of a variable myriad of life-threatening arrhythmias. In Brazil, the CCC has a mortality annual rate of 24/1000 patients, and sudden cardiac death (SCD) is one of the main modes of death (55%-65%) with an annual mortality rate varying from 0.2% to 19.2%[5-6]. Approximately, 90% are due to sustained ventricular tachycardia (SVT) that degenerates to ventricular fibrillation throughout the various stages of the CCC[5-7]. Concerning this context, a considerable number of these patients die and that could be empirically averted if patients at risk were identified and treated with implantable cardioverter defibrillators (ICDs)[8-9]. Nonetheless, the accurate identification of vulnerable patients susceptible for frequent malignant arrhythmias and SCD who actually benefit from ICD implantation remains a great clinical challenge in CCC[9]. The use of a machine learning based approach can meet this need. However, this technique was very little tested in the universe of CCC patients. We want to evaluate the discriminative ability of the clinical and electrical variables panel to identify individual subgroups with higher risk for SCD in CCC. Therefore, we have the following research question. Considering a dataset with a wide diversity of CCC patients, containing a diversity of clinical data, we may obtain an application based on Machine Learning for classifying these patients as low and high predisposition for SCD, applying also resampling methods and feature selection. In addition, as another challenge, the success for SCD prediction may occur even when the interval between the most recent exams and the SCD event is months or years.

The mentioned research question was applied using Cox proportional-hazards models with clinical variables[10]; Logistic Regression (LR)/Fisher's linear discriminant (LDA) using Heart Rate Turbulence (HRT) and Heart Rate Variability (HRV) features[11], and K-Nearest Neighborhood with HRT and HRV features[12]. Unfortunately, most of these approaches are performed with a small and non-significative sample, compromising the generalization of the results. Besides, algorithms like Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB) typically outperform LR, KNN, and cox models, but weren't tested. Finally, the lack of sensitivity/specificity and methodologies

Table 1 - Attributes used in the experiments

Attributes Group	Variables	Type
Personal Data	Gender	Categorical
	Body Mass Index	Quantitative
Clinical History	Cancer	Categorical
	Systemic Arterial Hypertension	Categorical
	Type 2 Diabetes Mellitus	Categorical
	Other Heart Diseases	Categorical
	Pacemaker	Categorical
	Syncope	Categorical
	Atrial Fibrillation/Flutter	Categorical
	Chronic Kidney Failure	Categorical
	Pulmonary Embolism	Categorical
	Cardiac insufficiency	Categorical
	Ventriculoperitoneal Shunt	Categorical
	Tabagism	Categorical
	Alcoholism	Categorical
	Sedentary Lifestyle	Categorical
ECG	Inactive Electrical Area	Categorical
	Ventricular Extrasystole	Categorical
	Supraventricular Extrasystole	Categorical
	Non-Sustained Ventricular Tachycardia	Categorical
	Pause > 3s	Categorical
	Primary Change	Categorical
	Interventricular Conduction Disturbance	Categorical
	Atrioventricular Conduction Disturbance	Categorical
ECO	Diastolic Dysfunction	Categorical
	Left Atrial Diameter	Quantitative
	Left Ventricular Diastolic Diameter	Quantitative
	Left Ventricular Systolic Diameter	Quantitative
	Segmental Deficit	Categorical
Holter	Atrial Fibrillation/Flutter	Categorical
	Average Heart Rate	Quantitative
	Sinus Node Dysfunction	Categorical
	Sustained Ventricular Tachycardia	Categorical
	Non-Sustained Ventricular Tachycardia	Categorical
	Ventricular Extrasystoles	Categorical
	Total Ventricular Extrasystoles	Quantitative
	Atrioventricular Conduction Disturbance	Categorical

to lead with unbalanced data make it impracticable to reproduce those works in different datasets.

To solve the presented drawbacks, we proposed a new prediction system tested on 310 patient samples using 9 state-of-art algorithms in machine learning and two leading techniques with unbalanced data. As for results, we verify Holter - Non-Sustained Ventricular Tachycardia, Total Ventricular Extrasystoles, Left Ventricular Systolic Diameter, History - Atrial Fibrillation/Flutter and Left Ventricular Diastolic Diameter as important features to predict SCD.

MATERIAL AND METHODS

Database

Clinical and laboratory data were collected from 310 patients at University Hospital Clementino Fraga Filho (HUCFF), at Federal University of Rio de Janeiro, for 26 years, between 1990 and 2016. About 160 patients had two or more medical records. Hence, the database contains 550 samples, from which 232 are male patients and 318 are female patients. About 14,7% of records (81) correspond to patients who suffered SCD as an outcome of Chagas disease. In order to obtain the database used in this paper, the protocol was approved by HUCFF-UFRJ ethics committee, who waived the need for written consent under number 45360915.1.1001.5262, in accordance with the current standards applied by national research ethics committee (Conep) and the principles described in Declaration of Helsinki.

The scope of this approach is limited to the classification of chagas patients into two classes: SCD and non-SCD. Hence, some considerations about the original database are made. First, the patients who died from causes other than SCD, such as natural causes or other diseases, are allocated into a non-SCD class. Only the most recent record of each patient is used, totalizing 310 unique patients, from which 78 (25,16%) died from SCD and 232 (74,84%) were alive until the last check or died from other causes. Furthermore, only 37 features are considered, referring to previous medical history and data from Electrocardiogram, Echocardiogram and Holter exams. Other features, such as medicament use, are discarded with the aim to avoid bias in the data. Table 1 summarizes the features used in this approach.

To occurrence of death was conducted a probabilistic linkage with the Brazilian National Mortality System (SIM-Sistema de Informação sobre Mortalidade, in Portuguese) using full name, date of birth, mother's name, and municipality of residence as matching variables. The linkage algorithm has been previously validated with a sensitivity and specificity of 94% and 91%, respectively[13]. When contact was possible, SCA data were obtained by direct interview with participants' relatives. In addition, information about SCA was also obtained annually from HUCFF Digital Registry and from mobile emergency care service (SAMU in Portuguese) [SAMU database, available at http://www.saude.gov.br/samu], Brazil. Non-witnessed cases by SAMU crew were excluded. SAMU follows the French pre-hospital care model that provides on-scene care for individuals and not just transport to the hospital. This is supported by the Brazilian Government and is available 24h a day, and consists of teams of health professionals that include medical doctors. It is the medical doctors' responsibility to complete the death certificates. Individuals not identified in the mortality database (SIM) were censored in February, 2020 (date of the linkage). As such, vital status was determined for all participants irrespective of whether contact was possible at the follow-up visit.

SCD was defined as an abrupt collapse with documented loss of vital signs that might result in attempts to restore circulation (cardiopulmonary resuscitation). The aetiology was only considered cardiac after the exclusion of SCAs due vascular non-cardiac disease, acute non-cardiac illnesses, drug overdose, metabolic causes or terminal disease[14].

It is important to emphasize that this database is constantly updated, despite the clinical data having been collected until 2016. Information like date and death occurrence are collected annually. In this approach, the last update was on 10/02/2020, and there were no new deaths until that day.

Related Works

Different approaches were proposed to identify the clinical and laboratory features of patients with high SCD propensity[9, 11-12]. The approach developed by[10] used Cox proportional-hazards model to evaluate the relationship between the risk factors with CCC and SCD, creating a risk score, analyzing ROC curves and Kaplan Meier survival curves to evaluate the scores' predictive performances. Although they have a sample of 373 examples (43 examples of SCD), they did neither explore some scenarios, such as the use of data resampling techniques, nor provided details about sensibility and specificity.

The work in[11] proposed to extract features from Heart Rate Turbulence (HRT) and parameters from time-domain Heart Rate Variability (HRV) of ECG signals divided into two 12-hours periods. These parameters were used as input to two multivariate linear models - Logistic Regression (LR) and Fisher's linear discriminant (LDA). However, although stratified, the approach had a limited number of 22 samples. Finally, in another approach[12], applied HRT and HRV techniques to extract features from Holter ECG signals and investigate possible associations with SCD events, considering 3 different scenarios: a 24h complete signal, just the 12 hours of daylight, and the other 12 hours. Forward and backward feature selection methods were used to reduce the number of parameters, K-Nearest Neighbors to do the classification and Leave-One-Out to do cross-validation. However, this approach uses a limited sample of 82 patients (20 SCD positives) and the features are focused on the metrics extracted from the ECG signal processing. In addition, considering that a great amount of holter signals from CCC patients contain significant occurrences of ectopic/arrhythmic beats, we found that the methodology for feature extraction applying HRV techniques is impaired.

Table 2 - Comparison between all approaches.

	work	Sample Size	Lead with unbalanced	Use Clinical Data	Present precision/sensitivity/sensibility/f1-score	Use State of Art Machine Learning?
1	[10]	333/43	NO	YES	NO	NO
2	[11]	11/11	NO	YES	YES	NO
3	[12]	62/20	NO	NO	YES	NO
4	Proposed Work	232/78	YES	YES	YES	YES

Even though all mentioned approaches proposed a system to predict SCD, they differ from the proposed work in several aspects as you can see in Table 2. First, our system uses a bigger sample than approaches[11] and[12], providing more substantial results than the previous works. Also, different from[10], we use the appropriate metrics to evaluate our approach, not only accuracy, but precision, sensitivity, sensibility and f1-score which gives more information about the model's effectiveness.

In this work, we developed a model tested with 9 state-of-the-art machine learning algorithms in four different scenarios leading with imbalanced data (undersampling and oversampling techniques), yielding a robustness system to predict SCD and indicating which variables are more important for this prediction. These algorithms were optimized over a sample of 310 patients, from which 78 suffered SCD, aiming at a computer-aided diagnosis model of early identification of patients with high SCD propensity. We emphasize three innovative points over which the proposed application is developed: using a dataset with a wide diversity of clinical data for CCC patients, including clinical notes, heart tests, treatments and classifiers, and a strong heterogeneity concerning the temporal distance between the most recent exams and the SCD event (which may reach months or even years); comparison of the performance of nine Machine Learning models in different scenarios concerning the application of feature selection techniques and resampling methods; and, finally, the use of the parameters of the models to rank the clinical features concerning the prediction of SCD events.

Methodology

The proposed methodology can be divided in 4 steps: data resampling, attribute normalization, selection, and classification. First, data resampling is necessary due to the significant difference in number of the samples labeled as SCD and non-SCD. For this, both oversampling and undersampling techniques were applied to the data in order to improve the predictive power of the final classifier[15]. For oversampling, Synthetic Minority Oversampling Technique (SMOTE)[16] was adopted, which generates synthetic samples for the minority class from the existing samples. For undersampling, random under sampling was adopted, which removes samples from the majority class randomly.

After this, all attributes are rescheduled between 0 and 1 to ensure there is no discrepancy in magnitudes. Then, redundant or unnecessary attributes, which can impair the interpretation and results of the models, are removed. For this, the K-Best method was applied for feature selection. K-Best chooses a number K of best features according to a score from an evaluation metric. In this approach, the metrics below were chosen:

1 Chi2: The Chi-Squared statistic, calculated by the formula:

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right)$$

where O is the observed frequency, and E is the expected frequency of a category.

2 f_classif: This method calculates the F-value based on Analysis of Variance (ANOVA). This calculation is done by dividing the variance into groups via the internal variance of these groups.

Lastly, the remaining attributes are used as input to the classification algorithm. In this approach, the selected algorithms are: K-Nearest Neighbors (KNN), Gradient Boosting (GB), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), Balanced Random Forest (BRF), Multilayer Perceptron (MLP), Bagging Classifier (BGC) and Extra Trees Classifier (ETC).

An overview of the proposed sequence of steps is shown in Figure 1, which represents an implementation of this methodology using Scikit-Learn library for Python [Müller, 2016].

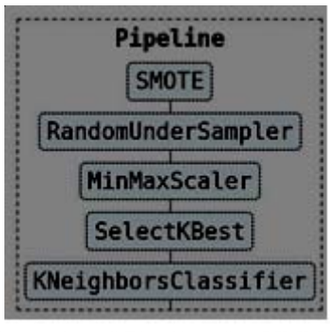

Figure 1 - Example of Pipeline for KNN using Scikit-Learn library and imblearn pipeline[17]

EXPERIMENTS AND RESULTS

In order to evaluate the performance of the proposed methodology, four experiment scenarios were done with several state-of-the-art machine learning algorithms: in Scenario 1, there was no feature selection or data resampling; in Scenario 2, there was feature selection, but there was no data resampling; in Scenario 3, there was data resampling but no feature selection. Finally, in Scenario 4, there was both data resampling and feature selection. Table 3 summarizes each scenario and its best performing algorithm.

Table 3 - Selected Scenarios

	Feature Selection	Resampling	Algorithm
Scenario 1	No	No	BRF
Scenario 2	Yes	No	BRF
Scenario 3	No	Yes	KNN/BRF
Scenario 4	Yes	Yes	BRF

Regarding the training methodology, 80% of samples were randomly selected for the training set and 20% for the testing set in a stratified way, to ensure the same proportion of classes in the training and testing sets. Within the training set, a stratified 5-fold method was used to estimate the hyperparameters, among those mentioned below.

1 GradientBoosting

1.a.i.a Learning rate: 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2

1.a.i.b Number of estimators: 10, 30, 70, 100

1.a.i.c Minimum number of samples to split a node: 12 equally spaced samples from 0.1 to 0.5

1.a.i.d Minimum number of samples to be a leaf: 12 equally spaced samples from 0.1 to 0.5

1.a.i.e Maximum number of features: binary logarithm, square root

1.a.i.f Maximum depth: 3, 5, 8

1.a.i.g Function to measure the quality of a split: mean squared error, Friedman mean squared error

1.a.i.h Subsamples for fitting: 0.5, 0.618, 0.8, 0.85, 0.9, 0.95, 1

1 Balanced Random Forest

1.a.i.a Number of estimators: 100 equally spaced samples from 151 to 1200

1.a.i.b Minimum number of samples to split a node: 5, 7, 10, 14

1.a.i.c Minimum number of samples to be a leaf: 4, 6, 8, 12

1.a.i.d Maximum number of features: binary logarithm, square root, all features

1.a.i.e Maximum depth: 10 equally spaced samples from 10 to 1200

1.a.i.f Function to measure the quality of a split: Gini index, information gain

1 Multilayer Perceptron

1.a.i.a Learning rate: constant, adaptative

1.a.i.b Hidden layer sizes: (200, 50, 30), (100, 50, 10), (100, 50), (200, 100), (500, 250), (20,), (50,), (100,), (10,), (200,)

1.a.i.c Activation function: hyperbolic tangent, rectified linear unit

1.a.i.d Solver for weight optimization: stochastic gradient descent, Adam

1.a.i.e Regularization parameter: 0.0001, 0.005, 0.05

1 Logistic Regression

(a) Regularization parameter: 0, 0.01, 0.1, 1.0, 10, 100

2 K-Nearest Neighbors

(a) Number of neighbors: 3, 5, 7, 9, 11

3 Support Vector Machine

1.a.i.a Kernel: RBF, linear

1.a.i.b Gamma (only RBF): 2^{i*15}, for i from 0 to 19, step 2

1.a.i.c Regularization parameter: 2^{i*5}, for i from 0 to 21, step 2

1 Gaussian Naive Bayes: No hyperparameters

2 Bagging Classifier

1.a.i.a Base estimator: Logistic Regression, K-Nearest Neighbors, GradientBoosting, Gaussian NB

1.a.i.b Number of estimators: from 1 to 10

1 Extra Trees Classifier

1.a.i.1.a Criterion: Entropy, Gini index

1.a.i.1.b Maximum depth: None and 10 equally spaced samples from 10 to 1200

1.a.i.1.c Maximum number of features: None, binary logarithm, square root, auto

1.a.i.1.d Minimum number of samples to split a node: 5, 7, 10, 14

1.a.i.1.e Minimum number of samples to be a leaf: 4, 6, 8, 12

1.a.i.1.f Number of estimators: from 1 to 10

It is important to note that the resampling, feature selection and normalizing techniques were applied using only the training test as base, i.e. there is no resampling in the test, and the values used to normalize the testing set were extracted from the training set. For the data resampling techniques, the used hyperparameters were:

1. SMOTE

(a) Final resampling ratio of the minority class over the majority class: 30%,

40%, 50%, 60%

2. Random Undersampling

(a) Final resampling ratio of the majority class over the minority class: 130%, 120%, 110%, 100%

After performing the experiments, the results were grouped and tabulated. Table 4 presents the mean and standard deviation of Accuracy, ROC curve, F1 score, Precision and Sensitivity for each algorithm, after 30 runs over the data set for the first scenario. As can be seen, excepting Naive Bayes, all algorithms presented accuracy rate greater than 80% but, due to the data imbalance, the sensitivity was affected in most classifiers, except in NB and BRF, which had 83.33±9.62% and 84.58±9.49%, respectively.

The addition of feature selection without any data balancing methodology did not provide a significant improvement to the results, as shown in Table 5. NB and BRF had a slight improvement in their sensitivity at the cost of their accuracy/specificity, suggesting that the classifier does not increase its predictive power. We can only hypothesize that the poor distribution of data negatively affects the feature selection.

In scenario 3, according to Table 6, the addition of oversampling and undersampling techniques resulted in the significant improvement of algorithms' performance except BRF and NB, which had a reduction in sensitivity, probably due to some interference in its own internal

Table 4 - Scenario 1: No resampling and no feature selection

	Accuracy		ROC Curve		F1		Precision		Sensitivity	
	Mean	**STD**	**Mean**	**STD**	**Mean**	**STD**	**Mean**	**STD**	**Mean**	**STD**
BRF	83.76%	± 4.8%	84.03%	± 5.09%	72.76%	± 6.77%	64.38%	± 8.17%	84.58%	± 8.49%
NB	72.59%	± 7.34%	76.13%	± 6.66%	61.06%	± 7.15%	48.62%	± 7.24%	83.33%	± 9.62%
GBC	87.3%	± 4.45%	81.73%	± 6.13%	73.69%	± 8.98%	78.44%	± 10.34%	70.42%	± 11.12%
KNN	86.08%	± 4.24%	80.78%	± 6.61%	71.53%	± 9.51%	74.31%	± 9.16%	70.0%	± 12.76%
LR	85.34%	± 4.08%	78.91%	± 5.76%	69.29%	± 8.24%	75.44%	± 10.92%	65.83%	± 12.25%
MLP	85.61%	± 3.68%	79.57%	± 5.37%	70.14%	± 7.77%	75.46%	± 10.34%	67.29%	± 11.8%
SVM	84.92%	± 4.2%	78.69%	± 5.65%	68.8%	± 8.22%	73.59%	± 9.63%	66.04%	± 11.8%
BGC	86.35%	± 3.27%	80.55%	± 4.78%	71.75%	± 6.72%	76.82%	± 9.96%	68.75%	± 10.51%
ETC	86.77%	± 3.55%	80.69%	± 5.65%	72.09%	± 7.99%	78.56%	± 10.29%	68.33%	± 12.38%

Table 5 - Scenario 2: Feature selection and no resampling

	Accuracy		ROC Curve		F1		Precision		Sensitivity	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
BRF	81.27%	± 1.28%	83.39%	± 1.07%	70.42%	± 1.54%	58.85%	± 1.89%	87.71%	± 1.14%
NB	61.9%	± 0.0%	70.35%	± 0.0%	53.85%	± 0.0%	38.89%	± 0.0%	87.5%	± 0.0%
GBC	85.03%	± 2.45%	80.62%	± 4.33%	70.67%	± 5.72%	69.91%	± 3.73%	71.67%	± 8.33%
KNN	80.95%	± 0.0%	70.74%	± 0.0%	57.14%	± 0.0%	66.67%	± 0.0%	50.0%	± 0.0%
LR	85.71%	± 0.0%	82.18%	± 0.0%	72.73%	± 0.0%	70.59%	± 0.0%	75.0%	± 0.0%
MLP	83.86%	± 0.73%	78.54%	± 1.44%	68.04%	± 1.93%	68.39%	± 0.94%	67.71%	± 2.88%
SVM	82.54%	± 0.0%	75.93%	± 0.0%	64.52%	± 0.0%	66.67%	± 0.0%	62.5%	± 0.0%
BGC	81.64%	± 1.85%	74.98%	± 3.08%	62.86%	± 4.26%	64.66%	± 3.95%	61.46%	± 6.37%
ETC	80.32%	± 2.45%	73.69%	± 2.54%	60.86%	± 3.86%	62.1%	± 5.95%	60.21%	± 5.31%

Table 6 - Scenario 3: Resampling and no feature selection

	Accuracy		ROC Curve		F1		Precision		Sensitivity	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
BRF	86.03%	± 4.54%	85.0%	± 5.72%	75.12%	± 7.53%	69.58%	± 8.72%	82.92%	± 11.0%
NB	75.34%	± 5.91%	74.82%	± 8.41%	59.3%	± 13.86%	50.36%	± 12.39%	73.75%	± 18.74%
GBC	85.82%	± 4.86%	83.49%	± 5.58%	73.94%	± 7.81%	70.78%	± 9.87%	78.75%	± 10.33%
KNN	86.24%	± 3.87%	85.42%	± 5.06%	75.54%	± 6.59%	69.58%	± 7.48%	83.75%	± 10.06%
LR	85.66%	± 4.87%	84.62%	± 4.99%	74.74%	± 7.18%	69.19%	± 9.42%	82.5%	± 9.2%
MLP	86.83%	± 4.47%	83.54%	± 6.02%	74.67%	± 8.13%	74.68%	± 10.88%	76.88%	± 13.04%
SVM	85.98%	± 3.91%	84.56%	± 5.32%	74.67%	± 7.03%	69.69%	± 7.93%	81.67%	± 10.88%
BGC	85.19%	± 4.53%	82.99%	± 7.01%	72.54%	± 9.02%	69.24%	± 9.11%	78.54%	± 15.46%
ETC	83.12%	± 5.57%	81.75%	± 5.41%	70.61%	± 7.48%	65.08%	± 9.81%	78.96%	± 10.82%

balancing methodology. Note, however, that BRF, SVM, LR and KNN obtained sensitivity over 80%.

In addition, MLP stood out for presenting the best accuracy with 86.83±4.47% and F1 score 74.67±8.13%. This suggests that data balancing is more important than feature selection to better generalize the data prediction.

Lastly, Table 7 presents the results of the last scenario. Except for the NB classifier, which had a significant improvement in almost every metric (probably because feature selection helps improve independence between the attributes, a hypothesis assumed by NB), no significant improvement was observed in the mean of any of the metrics used for the other classifiers. For this group of experiments, the best algorithm was BRF, with a result of 70.04 ± 3,6%, very similar to the result obtained in the first scenario, 72.76 ± 6.77%. However, it is noted that the standard deviation decreased drastically for all metrics used. A hypothesis for this is that the reduction of attributes can help to reduce data variance, resulting in better model consistency. Similar results can be observed for the remaining algorithms.

Among the four scenarios presented, BRF was the algorithm that outperformed the others in most scenarios. However, in scenario 3, SVM and LR performed very close to BRF, and KNN outperformed BRF with a better F1 score considering all experiments.

Here, we face a trade-off of interpretability/accuracy of results. BRF presents more consistent results, and also presents coefficients for the weights that can be used to interpret the importance of each feature, while KNN presents minimally superior results, but does not present any easy way to interpret the features. In order to verify the real difference between KNN and BRF, and to see if the better effectiveness of KNN justifies their lack of interpretability, we applied the Kolmogorov-Smirnov test to compare both of classifiers and assess if the performance of any classifier is significantly different based on their F1-score.

For a p-value 0.03458, we can not reject the null hypothesis that the two empirical data distributions are the same with 95% level confidence. Thus, the indicated algorithm is the BRF, which obtained superior or statistically similar results in all scenarios.

In order to better interpret the results, we summarized a table containing all the feature coefficients provided by the BRF, and also we formulated a ranking for all analyzed features. The rank consists of the weighted average of the features' coefficients of the 4 classifiers obtained in the 4 scenarios presented. The extracted coefficients and ranks from BRF are presented in Table 8, where 0.0"

Table 7 - Scenario 4: Resampling and feature selection

	Accuracy		ROC Curve		F1		Precision		Sensitivity	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
BRF	82.12%	± 2.04%	82.24%	± 3.06%	70.04%	± 3.6%	61.03%	± 3.31%	82.5%	± 6.44%
NB	72.91%	± 5.88%	75.8%	± 5.72%	59.7%	± 10.09%	49.38%	± 8.19%	81.67%	± 18.78%
GBC	80.79%	± 5.33%	79.71%	± 4.27%	67.53%	± 5.82%	60.6%	± 7.74%	77.5%	± 7.45%
KNN	81.8%	± 2.76%	81.69%	± 3.01%	69.46%	± 3.77%	61.04%	± 4.81%	81.46%	± 7.43%
LR	80.79%	± 3.96%	79.98%	± 3.29%	67.64%	± 4.68%	60.17%	± 7.08%	78.33%	± 6.71%
MLP	82.28%	± 4.11%	80.43%	± 4.44%	68.86%	± 6.24%	62.78%	± 7.22%	76.67%	± 6.55%
SVM	81.9%	± 4.12%	81.62%	± 4.58%	69.56%	± 6.19%	61.32%	± 6.87%	81.04%	± 8.12%
BGC	80.74%	± 4.54%	80.91%	± 3.68%	68.44%	± 5.29%	59.62%	± 7.04%	81.25%	± 6.36%
ETC	81.8%	± 2.66%	81.69%	± 2.63%	69.51%	± 3.52%	60.89%	± 4.64%	81.46%	± 5.31%

Table 8 - Normalized attribute weight coefficients for the main models ranked

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Score	Ranking
Holter – Non-Sustained Ventricular Tachycardia	0.40925743	0.15267842	0.14246527	0.15267842	0.17929491425398647	1
Total Ventricular Extrasystoles	0.10960927	0.15148947	0.14723054	0.15148947	0.14696716603599536	2
Left Ventricular Systolic Diameter	0.03674904	0.10134342	0.08076052	0.10134342	0.08883627806210703	3
Syncope	0.00249959	0.08604691	0.09936974	0.08604691	0.08171139343906424	4
Left Ventricular Diastolic Diameter	0.03237515	0.10480306	0.05127252	0.10480306	0.08135053957410422	5
Supraventricular Extrasystoles	0.05034312	0.04543784	0.06812225	0.04543784	0.05336466353264556	6
Primary Change	0.01187713	0.06092105	0.05098271	0.06092105	0.0530220621022198	7
Left Atrial Diameter	0.05081418	0.05134889	0.03801602	0.05134889	0.04767526615226073	8
ECG – Ventricular Extrasystoles	0.02560497	0.04011265	0.0520354	0.04011265	0.0425433270367147	9
Interventricular Conduction Disturbance	0.02570511	0.03955867	0.03899984	0.03955867	0.038225456651444306	10
Inactive Electrical Area	0.07287255	0.01891564	0.04698278	0.01891564	0.03366404364940776	11
Diastolic Dysfunction	0.02371363	0.03142102	0.03236045	0.03142102	0.031150614767228225	12
Holter – Ventricular Extrasystoles	0.00755758	0.02153298	0.01889865	0.02153298	0.019381599417631977	13
Chronic Kidney Failure	0.01335745	0.01305777	0.02402586	0.01305777	0.01657875761418014	14
Average Heart Rate	0.05771106	0.0	0.0197916	0.0	0.012446962286701103	15
Cardiac insufficiency	0.00904056	0.0115332	0.0197892	0.0	0.010485695181165527	16
Segmental Deficit	0.0	0.01341656	0.00716877	0.01341656	0.01014230625890852	17
Gender	0.00540964	0.01189868	0.007859	0.01189868	0.01005296295900673	18
Sedentary Lifestyle	0.00573929	0.00786285	0.01172643	0.00786285	0.008887360307147102	19
Body Mass Index	0.01612447	0.01490231	0.00751533	0.0	0.008497855657897097	20
ECG – Non-Sustained Ventricular Tachycardia	0.01166123	0.00374021	0.00793616	0.00374021	0.0059374862657359025	21
Systemic Arterial Hypertension	0.00490441	0.00459214	0.00617821	0.00459214	0.005156149580690402	22
Holter – Atrioventricular Conduction Disturbance	0.00098582	0.00365439	0.00635823	0.00365439	0.004223614390267675	23
Type 2 Diabetes Mellitus	0.00417701	0.00386169	0.00421512	0.0	0.0028979345914201555	24
Sinus Node Dysfunction	0.00207107	0.00221521	0.00338666	0.00221521	0.002579453522876169	25
Holter – Atrial Fibrillation/Flutter	0.00339937	0.00106094	0.0038891	0.00106094	0.0021970594000050815	26
ECG – Atrioventricular Conduction Disturbance	0.00248463	0.00146184	0.00131983	0.00146184	0.0015447859680376619	27
Ventriculoperitoneal Shunt	0.00067337	0.00113219	0.00065296	0.0	0.000609867545382944	28
Pacemaker	0.00224245	0.0	0.00069086	0.0	0.0004596520343358914	29
History – Atrial Fibrillation/Flutter	0.00103943	0.0	0.0	0.0	0.0001147717614314631	30
Pulmonary Embolism	0.0	0.0	0.0	0.0	0.0	31
Sustained Ventricular Tachycardia	0.0	0.0	0.0	0.0	0.0	32
Other Heart Diseases	0.0	0.0	0.0	0.0	0.0	33
Tabagism	0.0	0.0	0.0	0.0	0.0	34
Cancer	0.0	0.0	0.0	0.0	0.0	35
Alcoholism	0.0	0.0	0.0	0.0	0.0	36
Pause > 3s	0.0	0.0	0.0	0.0	0.0	37

represents a zero value importance assigned to the variables that were not selected in the feature selection step.

Random Forest is composed of several decision trees, evaluating multiple scenarios with or without a certain feature. The importance coefficient measures the loss of accuracy in trees that do not contain a certain feature. In other words, this is an average of the importance (loss of accuracy when not present) of that feature among all decision trees. That means, we can interpret the average of the coefficients of the 4 BRF's for the different scenarios as an approximation of the composition of all decision trees of these scenarios.

According to the Table 8, we can see the huge importance of the variables: Holter Non-Sustained Ventricular Tachycardia and Total Ventricular Extrasystoles, followed by Left Ventricular Systolic Diameter, History - Atrial Fibrillation/Flutter and Left Ventricular Diastolic Diameter, indicating that the variables are strongly related in the identification of sudden death in patients with Chagas. In addition, although on a smaller scale, the other attributes still added relevant information with the exception of the following variables: Sustained Ventricular Tachycardia, Pause > 3s, Alcoholism, Tabagism, Cancer, Pulmonary Embolism and Other Heart Diseases, which had zero importance in all scenarios.

CONCLUSION

The main contribution of this paper is to provide a series of computer-aided diagnosis models using a significant dataset, as well as a proper and original balancing system and metrics. In addition, these models contain different interpretation levels about clinical variables which can be used as an adjuvant in the SCD risk assessment in a wide spectrum of patients treated contemporaneously with CCC. All algorithms were evaluated for accuracy, precision, sensitivity and F1 score. KNN, BRF, LR and SVM presented the best results for sensitivity without reduction in other metrics. Another important finding was that, looking at the weight coefficients of each attribute provided by the models, we noticed that the variables Holter - Non-Sustained Ventricular Tachycardia, Total Ventricular Extrasystoles, Left Ventricular Systolic Diameter, Syncope, History - Atrial Fibrillation/Flutter and Left Ventricular Diastolic Diameter had great impact on the decision of sudden cardiac death. Also, we provide a rank for 37 variables about their importance for the prediction of sudden death in patients with Chagas.

In addition, 4 methodologies were presented for evaluating the effectiveness of feature selection and data resampling techniques for this approach. It was found that the combination of SMOTE with random under sampling ensures a better result for sensitivity. On other hand, feature selection slightly affected sensitivity and precision negatively. However, it contributed to a drastic reduction in variance of the results. For future studies, we will extract relevant features from the electrocardiogram signals improving the predictive performance of the current model and also propose a real-time diagnostic system.

ACKNOWLEDGMENTS

The authors thank the support of Federal University of Ceara (PIBIC-UFC), Ceará Foundation for Scientific and Technological Development Support (BICT-FUNCAP process IC7-0170-00004.01.93/20 and process PS1-0186-00439.01.00/21) and CNPq (processes 426002/2016-4, 440092/2020-5 e 310317/2019-3).

REFERENCES

1. WHO (2021). Chagas disease (american trypanosomiasis). https://www.who.int/healthtopics/chagas-disease.
2. Borges-Pereira, J., Coura, J. R., Zauza, P. L., Pirmez, C., and Xavier, S. S. (2020). Chagas disease in virgem da lapa, minas gerais, brazil: left ventricle aneurysm and the risk of death in the 24-year interval. Memórias do Instituto Oswaldo Cruz, 115.
3. Arnal, A., Waleckx, E., Rico-Chavez, O., Herrera, C., and Dumonteil, E. (2019). Estimating the current burden of chagas disease in mexico: A systematic review and meta-analysis of epidemiological surveys from 2006 to 2017. PLoS neglected tropical diseases, 13(4):e0006859.
4. Silva, L. E. V., Moreira, H. T., Bernardo, M. M. M., Schmidt, A., Romano, M. M. D., Salgado, H. C., Fazan Jr, R., Tinos, R., and Marin-Neto, J. A. (2021). Prediction of´ echocardiographic parameters in chagas disease using heart rate variability and machine learning. Biomedical Signal Processing and Control, 67:102513.
5. Nunes, M. C. P., Beaton, A., Acquatella, H., Bern, C., Bolger, A. F., Echeverria, L. E., Dutra, W. O., Gascon, J., Morillo, C. A., Oliveira-Filho, J., et al. (2018). Chagas cardiomyopathy: an update of current clinical knowledge and management: a scientific statement from the american heart association. Circulation, 138(12):e169–e209.
6. Marin-Neto, J. A., Cunha-Neto, E., Maciel, B. C., and Simões, M. V. (2007). Pathogenesis of chronic chagas heart disease. Circulation, 115(9):1109–1123.
7. Simões, M. V., Romano, M. M. D., Schmidt, A., Martins, K. S. M., and Marin-Neto, J. A. (2018). Chagas disease cardiomyopathy. International Journal of Cardiovascular Sciences, 31:173–189.
8. Pavao, R. C., Licia, M. et al. Electrical Storm in Chagas Cardiomyopathy: Clinical Predictors, Outcome, and Arrhythmic Characteristics in a Prospective Registry. Clinical Electrophysiology, v. 6, n. 10, p. 1238-1245, 2020.
9. Rassi, F. M. et al. Systematic review and meta-analysis of clinical outcome after implantable cardioverter-defibrillator therapy in patients with Chagas heart disease. JACC: Clinical Electrophysiology, v. 5, n. 10, p. 1213-1223, 2019.
10. de Souza, A. C. J., Salles, G., Hasslocher-Moreno, A. M., de Sousa, A. S., do Brasil, P. E. A. A., Saraiva, R. M., and Xavier, S. S. (2015). Development of a risk score to predict sudden death in patients with chaga's heart disease. International journal of cardiology, 187:700–704.
11. Alberto, A. C., Limeira, G. A., Pedrosa, R. C., Zarzoso, V., and Nadal, J. (2017). Ecg-based predictors of sudden cardiac death in chagas' disease. In 2017 Computing in Cardiology (CinC), pages 1–4. IEEE.
12. Alberto, A. C., Pedrosa, R. C., Zarzoso, V., and Nadal, J. (2020). Association between circadian holter ecg changes and sudden cardiac death in patients with chagas heart disease. Physiological Measurement, 41(2):025006.
13. Capuani, L., Bierrenbach, A. L., Pereira Alencar, A., Mendrone Jr, A., Ferreira, J. E., Custer, B., P. Ribeiro, A. L., and Cerdeira Sabino, E. (2017). Mortality among blood donors seropositive and seronegative for chagas disease (1996–2000) in são paulo, brazil: A death certificate linkage study. PLoS neglected tropical diseases, 11(5):e0005542.

14. Hinkle Jr LE, Thaler HT. Clinical classification of cardiac deaths. Circulation. 1982 Mar;65(3):457-64.
15. Hernandez, J., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. (2013). An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets. In Iberoamerican Congress on Pattern Recognition, pages 262–269. Springer.
16. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357.
17. Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17):1–5.