Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(5): 9159–9178.
DOI: 10.3934/mbe.2023402
Received: 30 September 2022
Revised: 01 February 2023
Accepted: 20 February 2023
Published: 14 March 2023

Research article

Sudden cardiac death multiparametric classification system for Chagas heart disease's patients based on clinical data and 24-hours ECG monitoring

Carlos H. L. Cavalcante[1,2,*], Pedro E. O. Primo[3], Carlos A. F. Sales[1], Weslley L. Caldas[3], João H. M. Silva[4], Amauri H. Souza[1], Emmanuel S. Marinho[2], Roberto C. Pedrosa[5], João A. L. Marques[6], Hélcio S. Santos[2] and João P. V. Madeiro[3]

[1] Federal Institute of Education and Technology of Ceara, Maracanau, Ceara, Brazil

[2] State University of Ceara - Center for Science and Technology, Fortaleza, Ceara, Brazil

[3] Computer Science Department – Federal University of Ceara, Fortaleza, Ceara, Brazil

[4] Oswaldo Cruz Foundation (Fiocruz), Eusebio, Ceara, Brazil

[5] Edson Saad Heart Institute – Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[6] Laboratory of Applied Neurosciences -University of Saint Joseph, Macau SAR, China

* **Correspondence:** Email: henriqueleitao@ifce.edu.br.

Abstract: About 6.5 million people are infected with Chagas disease (CD) globally, and WHO estimates that >1 million people worldwide suffer from ChHD. Sudden cardiac death (SCD) represents one of the leading causes of death worldwide and affects approximately 65% of ChHD patients at a rate of 24 per 1000 patient-years, much greater than the SCD rate in the general population. Its occurrence in the specific context of ChHD needs to be better exploited. This paper provides the first evidence supporting the use of machine learning (ML) methods within non-invasive tests: patients' clinical data and cardiac restitution metrics (CRM) features extracted from ECG-Holter recordings as an adjunct in the SCD risk assessment in ChHD. The feature selection (FS) flows evaluated 5 different groups of attributes formed from patients' clinical and physiological data to identify relevant attributes among 57 features reported by 315 patients at HUCFF-UFRJ. The FS flow with FS techniques (variance, ANOVA, and recursive feature elimination) and Naive Bayes (NB) model achieved the best classification performance with 90.63% recall (sensitivity) and 80.55% AUC. The initial feature set is reduced to a subset of 13 features (4 Classification; 1 Treatment; 1 CRM; and 7 Heart Tests). The proposed method represents an intelligent diagnostic support system that predicts the high risk of SCD in ChHD patients and highlights the clinical and CRM data that most strongly impact the final outcome.

Keywords: sudden cardiac death; Chagas heart disease; machine learning; ECG

1. Introduction

Identifying patients at high risk for sudden cardiac death (SCD) is a major scientific challenge in cardiac diseases. SCD represents one of the leading causes of death around the world [1]. In particular, SCD corresponds to approximately 65% of the deaths of patients diagnosed with Chagas heart disease (ChHD) [2], a prevalent endemic disease in Brazil and 20 Latin American countries [3]. The rate of SCD in ChHD is 2.4 percent per year, substantially more significant than the SCD rate for the general population [4]. World Health Organization (WHO) estimates that 6–7 million people are infected with Chagas disease (CD). Caused by *Trypanosoma cruzi*, it is a life-threatening and persistent illness, and up to 1.5 million (30%) of them suffer from ChHD in the world [5]. Due to increasing migratory, it has trespassed frontiers, and is rising in non-endemic countries (North America, Europe, Japan, and Australia).

The unpredictability of SCD is widely recognized, making this identification a crucial open problem [1]. In ChHD, although the patients at high risk for SCD can theoretically be identified by their risk factors, in practice, the most significant number of SCD cases occurs in patients not previously determined to be at high risk [6]. Notably, this unpredictability hinders the widespread implementation of effective preventive measures on a large scale against SCD in ChHD patients [7–9]. A significant number of these patients die in this context, which could be empirically avoided if SCD risk was accurately identified and treated with implantable cardioverter defibrillators (ICDs) [10]. Idealistically, accurate and early identification of patients at high risk for SCD could enable the adoption of more advanced but costly treatments.

Several studies of Machine Learning (ML) algorithms applied to cardiology can be found in databases such as PubMed and MEDLINE. Almost all cardiology subfields have applied ML to automatic ECG interpretation, result analysis, monitoring, or diagnostic support systems [11,12]. Concerning the specific context related to SCD, some studies have shown that ML can classify SCD patients with the same or greater accuracy than clinicians [13]. Despite the relevance of the SCD issue, analyzed as a global phenomenon [1], its occurrence in the specific context of ChHD has been poorly exploited, especially considering the use of machine learning computing tools. In parallel, technologies for real-time cardiac activity monitoring has grown significantly in recent decades, being a valuable instrument for use in medicine, considering currently miniaturized devices, wearable or even implantable systems, which are based on highly efficient biosensors, to provide data related to the heart electrical activity, the heart's rate variability and other physiological signals such as blood pressure or body temperature [14,15]. The parameters can be transferred via a wireless or wired link to a microcontroller board, and diagnostic support systems may analyze and share information with specialists [16,17].

Numerous kinds of research have been used to enhance the accuracy of death risk stratification. They have used information from clinical notes or exams from many different sources, such as the electrocardiogram (ECG), 24-hour Holter, Exercise Test, Echocardiography, Cardiac Computed Tomography (CT), Cardiovascular Magnetic Resonance, and myocardial scintigraphy with single photon emission computed tomography (SPECT) or positron emission tomography (PET) [18–20]. However, most of them performed only classical linear analyses and did not fully evaluate the potential of combining different indices for the prognostication of patients with ChHD. This diversity of data and its origins is a favorable scenario for using feature selection (FS) techniques as a pre-processing step in ML to identify the most relevant attributes for the desired classification task. The main goal is to select

the smallest possible subset of features appropriate for the problem, eliminating noise and redundant characteristics, preventing overfitting, and simplifying the development of intelligent systems [21, 22].

This study offers the first evidence to support the use of ML techniques within Patients' Clinical Data and Cardiac Restitution Metrics (CRM) features extracted from ECG records as an adjuvant in the SCD risk assessment in ChHD. CRM features are critical and extensively researched in the context of heart disease [23–25]. These characteristics make the present study innovative since it uses information from non-invasive tests. Also, to the best of our knowledge, at this moment, only two previous studies have added information to these literature gaps*. One study has a limited approach, which only performed classical linear analyses and did not fully assess the potential of combining different parameters [8]. In another research, the authors extracted eight variables by applying heart rate variability (HRV) and heart rate turbulence (HRT) techniques over Holter-ECG records to investigate SCD in ChHD. The set of features was reduced with the forward and backward-stepwise approaches. The left ventricular ejection fraction (LVEF) was also analyzed with these variables. The study used ECG records of patients divided into SCD deaths and alive patients. The work used the k-nearest neighbors classifier, and a leave-one-out cross-validation [26]. However, this approach uses a limited sample of 82 individuals (20 SCD positives), compromising the generalization of the results, and the features are focused only on the metrics extracted from the ECG signal processing. Finally, because of the lack of sensitivity, specificity, and methodologies for leading with unbalanced data, it is impractical to replicate those works in different datasets.

The main goal of this paper is to create an intelligent diagnostic support system for predicting SCD in ChHD patients. We will use an ML process to detect a high predisposition to SCD based on an optimized and reduced set of non-invasive test attributes. Other contributions of this work include: 1) using CRM features extracted from the ECG waveform by an automatic subsystem in conjunction with patient clinical data and different feature selection flows based on exhaustive testing; and 2) the searching for the best combination considering performance rates as high as possible and number of reduced features as low as possible.

2. Materials and methods

2.1. Definitions and background

2.1.1. Sudden cardiac death (SCD)

SCD was defined as an abrupt collapse with documented loss of vital signs that might result in attempts to restore circulation (cardiopulmonary resuscitation). The etiology was only considered cardiac after excluding of SCDs due to vascular non-cardiac disease, acute non-cardiac illnesses, drug overdose, metabolic causes, or terminal disease [27].

Concerning the vital status of all participants, a probabilistic linkage was conducted with the Brazilian National Mortality System (SIM-Sistema de Informação sobre Mortalidade, in Portuguese) [28]. The linkage algorithm has been previously validated with a sensitivity and specificity of 94% and 91%, respectively [29]. SIM covers the entire population nationwide. Mortality data are considered reliable from a qualitative point of view, as accurate as those of other countries with a long tradition

*The search strategy for reviewing the literature used the PubMed/MEDLINE, Web of Science, and SciELO databases, and Periodic CAPES portal. The descriptors used were: "Chagas disease"; "Chagas heart disease"; "Electrocardiography"; "Sudden Cardiac Death"; "Machine Learning".

in these statistics [30]. When contact was possible, SCD data were obtained by directly interviewing participants' relatives. In addition, information about SCD was also obtained annually from University Hospital Clementino Fraga Filho of Federal University of Rio de Janeiro (HUCFF-UFRJ), Digital Registry and mobile Emergency Medical Service (SAMU), Brazil. SAMU follows the French prehospital care model, which provides on-scene care for individuals and not just transport to the hospital. The Brazilian government supports this, being available 24 hours a day, and has teams of health professionals, including doctors. The doctors have the responsibility to complete the death certificates. Individuals who were not identified in the SIM were sanctioned in February 2016 (the date of the link).

2.1.2. Chagas heart disease (ChHD)

In the used dataset, the diagnosis of ChHD required at least two positive serology tests for *T. cruzi* antibodies (indirect hemagglutination, indirect immunofluorescence, or enzyme-linked immunosorbent assay) and electrocardiographic changes typical of Chagas disease [18].

2.1.3. Cardiology guidelines classification (CGC)

The New York Heart Association (NYHA) classifies the extent of heart failure. It divides patients into four groups (I, II, III, and IV) based on physical activity limitation (dyspnea) [31]. The Rassi risk score estimates general death risk from ChHD patients in the next 5–10 years. Patients are classified into low, intermediate, and high risk of death based on the sum of the regression coefficient points for six risk factors: NYHA class III or IV (5 points), cardiomegaly on chest radiography (5 points), left ventricular systolic dysfunction (3 points), non-sustained ventricular tachycardia (NSVT) on 24-hour Holter monitoring (3 points), low QRS voltage on ECG (2 points), and male sex (2 points) [4]. Stages based on the severity of cardiac (Guideline 2005) involvement were determined for all participants before the SCD event. The stages are defined by: the presence of abnormalities within ECG (stage A); abnormal ECG and Echo, left ventricular ejection fraction (LVEF) > 45% without symptoms of heart failure (stage B1); abnormal ECG and Echo, LVEF < 45% without symptoms of heart failure (stage B2); abnormal ECG, Echo, LVEF < 45%, and symptoms of heart failure (stage C) [32].

2.1.4. Cardiac restitution metrics (CRMs)

Cardiac restitution is a natural myocardial property that translates the heart's ability to dynamically recover the time interval QT (action potential duration) from one beat to the next. It is an essential mechanical point where the cardiac cycle is shortened at faster heart rates to allow more effective contraction and relaxation for the efficient function of the cardiac pump. The restitution function is nonlinear, highly dynamic, occurs independently from underlying heart rate variability, and varies with normal and abnormal physiological conditions, including the autonomic state. On the electrocardiogram (ECG), this can be estimated by comparing the current QT interval to the previous TQ interval (diastolic interval) [23].

The QT interval is long at slow heart rates (long diastolic interval and TQ), but it shortens at high heart rates or ectopic beats (short diastolic interval and TQ). The magnitude of QT interval shortening with a reduction in the diastolic interval, characterizing the dynamics of electrical restitution, can be described by a plot of QT interval versus diastolic interval (action potential duration restitution), usually assuming a mono-exponential curve. If the action potential duration restitution (APDR) slope

is <1, changes in TQ lead to relatively little change in QT with a balanced QT-TQ relationship. If the APDR slope is ≥1, a slight change in TQ would lead to more extensive changes in QT interval, reducing the action potential duration (APD) and leading to unstable re-entry and cardiac arrhythmias. APDR dynamic is recognized as one of the techniques for measuring the electrical restitution heterogeneity as a result of sympathetic activation [23]. For this work, some CRM was used and defined below. The acquisition process from ECG signals is shown in Figure 1 and explained in the next section.

- Tpeak-Tend: beat-to-beat temporal distance between the peak and the end of each T-wave in ECG or Holter signal. It is associated with the polarization and depolarization time variations between the heart's myocardial, endocardial, and epicardial layers.
- TendQ 5th Percentile (s) and TpeakQ 5th Percentile (s): statistical measures related, respectively, to the beat-to-beat temporal distance between the end (and the peak) of each T-wave and the beginning of the subsequent QRS complex, which are related to the relative refractory period. It has been proposed that arrhythmia vulnerability may increase due to the likelihood of re-entry as the relative refractory period approaches zero. The TQ 5th quantile is hypothesized to quantify the lower boundary where arrhythmia vulnerability is the greatest;
- Percentage (%) of beats with QTend/TendQ > 1 and Percentage (%) of beats with QTpeak/TpeakQ > 1: relative probability of occurrences where a measure of ventricular activity duration (QTend or QTpeak) is higher than a measure of the relative refractory period (TendQ or TpeakQ). This relationship has been proposed to be associated with increased arrhythmia vulnerability by the steepness of the restitution relationship. Therefore, the percentage of beats with a QT/TQ ratio greater than 1 reflects the relative time spent on the restitution curve where there may be increased instability;
- Upper 98% quantile of the Qtend/TendQ ratio and Upper 98% quantile of the QTpeak/TpeakQ ratio: statistical measures related to the most extreme beats for which the percentages QTend/TendQ and QTpeak/TpeakQ present the highest values. This measure reflects the magnitude of the steepness of the restitution relationship. The 98% quantile takes the most extreme beats with the highest likelihood of leading to arrhythmia into account.

The cardiac restitution metrics (CRM) acquisition process from ECG signals is shown in Figure 1 and explained in the next section.

2.1.5. Feature selection (FS)

Feature selection (FS) is choosing a better subset of attributes with more meaningful information for a given context. FS can reduce the set of attributes, remove noisy variables, detect the most relevant variables, or decrease the computational cost of training and testing by improving or without hurting the efficiency of the classification model [33]. This technique aims to obtain the best subset (s) of attributes within a more extensive set (X) of attributes, where $s \subseteq X$. In a dataset with X features that are intended to obtain ranks for a given feature (Y), the FS problem can be defined as a search for a subspace of attributes (R^s) in an X-dimensional space (R^X) that allows finding the best equivalent Y. The total number of subspaces is 2^X, and the number of subspaces with dimensions equal to or less than s is $\sum_{i=1}^{s} \binom{1}{X}$.

Due to many possibilities for finding the best subset, several FS techniques have been developed to decrease the complexity and computational cost. Filters, embedded, wrapper, hybrid, and ensemble

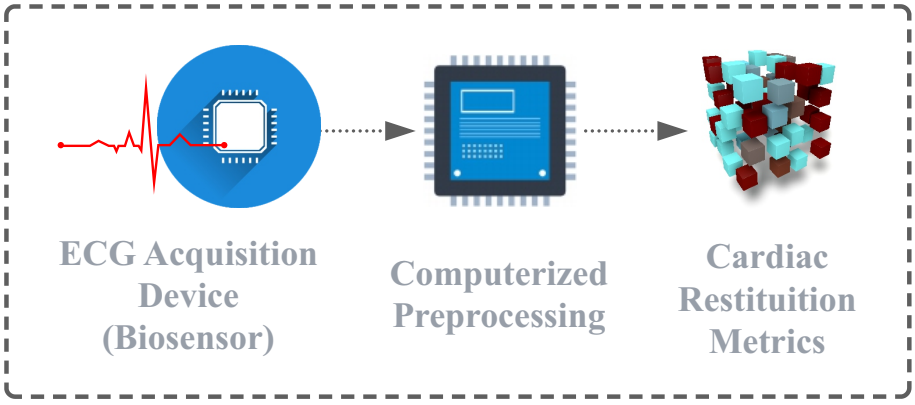

Figure 1. Stages for extracting CRM features: 1) ECG signal acquisition through a biosensor over 24 hours; 2) computerized preprocessing concerning noise removal, QRS detection and delineation, T-wave detection and delineation, and computing of beat-to-beat metrics; 3) output data with the generated CRM features.

are methods of FS techniques present in the literature. This research employed the following FS techniques: variance (VAR) and ANOVA as statistical filtering methods, select from model (SFM) as an embedded method, and recursive feature elimination (RFE) as a wrapper method [34,35]. VAR filter eliminates features with similar values and reduces the initial number of features. ANOVA enables the search for feature similarity. A slight variation in the means of two variables suggests that there is likely little difference in the data of the two variables, making them nearly identical and implying that one of them should be removed [36]. SFM technique uses the punctuation approaches capable of generating appropriate coefficients for each attribute that some machine learning (ML) classifiers employ. This argument ranks and selects a set number (N) of relevant features within a context. RFE uses the rank features based on the coefficients or feature importance attributes of the ML model. It removes a small number of features per loop, removing any existing dependencies and collinearities in the model. It finds a subset of features by starting with all of the features in the training dataset and successfully removing features until the desired number is reached [37].

2.2. ChHD database

The research's sample space was a dataset (clinical and cardiac examinations) from 315 patients followed in the protocolized clinical follow-up program of patients of the ChHD outpatient clinic of HUCFF-UFRJ between 1992 and 2016. Only patients who presented hospital episode statistics, ECG, echocardiogram, and Holter with a previous date closest to the SCD event were considered. Exclusion criteria were: patients with hypertensive heart disease, with ICD, cardiac resynchronization therapy or pacemaker implantation, and those with low-quality ECG-Holter signals. After the exclusion criteria, the sample space contains 218 patients with ECG-Holter. The coronary disease was excluded by symptom-limited stress test or invasive coronary angiography.

For the acquisition and use of the information in this research, the authorization process was approved by the local ethics committee, which waived the need for written informed consent under number 45360915.1.1001.5262 conforming to standards currently applied by the Brazilian National

Committee for Research Ethics and the principles outlined in the Declaration of Helsinki. All patients received treatment following norms and protocols previously established by the National Health Surveillance Secretariat/Ministry of Health for ChHD, Brazil [18].

The experiments were performed with 218 patients, 96 (44%) men, and 122 (56%) women. As for the type of death, 77 (35%) had SCD in ChHD, and 141 (65%) did not have SCD in ChHD. The dataset contains 57 attributes constituting the sample space, from which 51 attributes are Clinical Data of Patients and 7 attributes are CRM extracted from ECG Data Processing, described in the next section. The data acquisition is displayed in the top left and bottom left frames of Figure 2. For a better understanding, the applied 57 attributes were grouped into five categories (Table 1), namely: clinical, heart tests, Cardiology Guidelines Classification (CGC), treatments, and cardiac restitution metrics (CRM). Six features from Patients' Clinical Data were discarded to ensure the patients' anonymity.

2.2.1. The ECG data processing

The 24-hour ECG-Holter signal recordings started at 8 a.m. According to a study by Fossa [24], cardiac restitution metrics were used. They were extracted from 04 h (dawn) to 08 h (morning) at a frequency of 128 Hz. In this period, it is acceptable that a more extensive adrenergic discharge occurs which causes heterogeneity in cardiac excitability with a high probability of SCD [38]. The software developed by Madeiro [39] was used to extract the seven CRM features by detecting beat-to-beat QT and TQ intervals related to ventricular activity and refractory period (Figure 1). The inter- and intra-operator variability of the software was evaluated in previous work [25].

2.3. Classification experiments

Classification experiments with machine learning were divided into two scenarios. The first scenario (S1) used patients' clinical data and cardiac restitution metrics (CRM) features extracted from ECG-Holter. The second scenario (S2) used only patients' clinical data, excluding the CRM features. Booth scenarios ran the seven Feature Select (FS) flows (variance, ANOVA, SFM, RFE, Pipeline 1, Pipeline 2, and Pipeline 3) described in Section 2.3.2.

The implementation strategy for the classification process consisted of four steps: standardization process, feature selection (FS) flows, machine learning models (training and testing), and the choice of the ML model with the best performance for predicting SCD in ChHD, as shown in Figure 2. The coding was done in the Python programming language using the scikit-learn library [40].

2.3.1. Standardization process

As for the type, 37 of the attributes were binary, 15 scalar, and 5 categorical (2005 Guideline Classification, Rassi Score, NYHA, Classification, and Diastolic Dysfunction). The One Hot Encoding (OHE) technique transformed the values of the categorical features into binaries. This transformation was necessary because some ML algorithms work only with numerical values [41]. The data were normalized with the Min-Max and Z-score strategies to guarantee that all features were on the same scale [42].

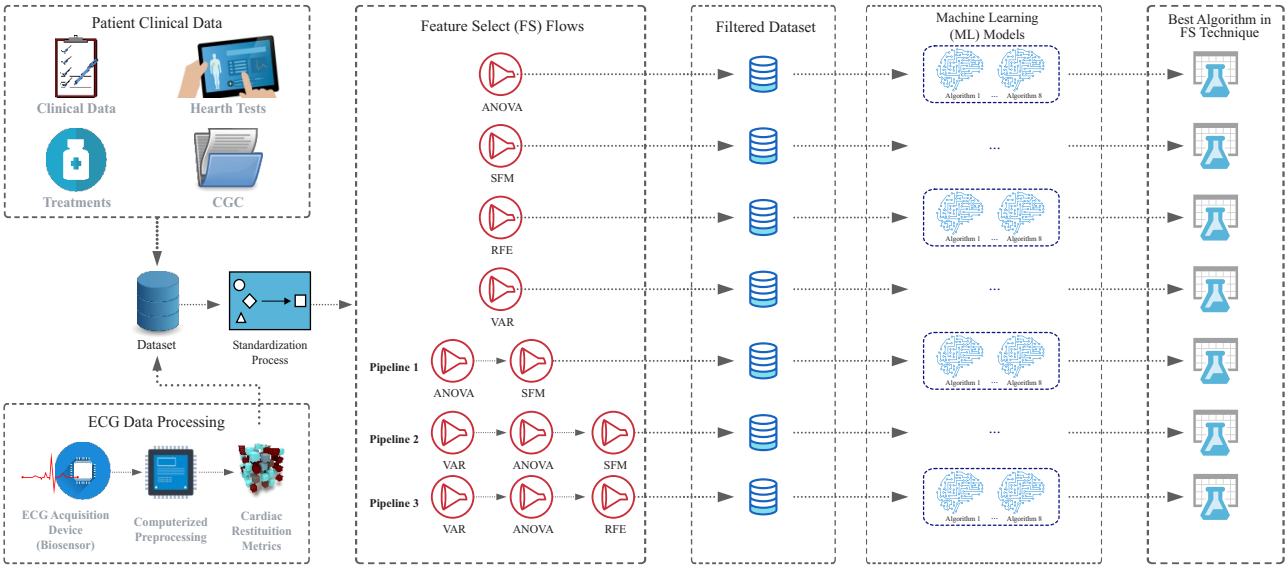

Figure 2. Stages for SCD Classification System in ChHD: 1) CRM features are obtained through an ECG Acquisition Device (Biosensor) and automatic preprocessing of ECG signals over 24 hours; 2) The dataset is built from patient clinical data and CRM features; 3) Each FS flow implements a specific FS technique (VAR, ANOVA, SFM, and RFE) or a pipeline of techniques (Pipeline 1, Pipeline 2, and Pipeline 3) providing seven different filtered datasets; 4) Eight different ML classifiers are evaluated for each filtered dataset and 5) the best algorithm in each flow is selected for analysis and comparison.

2.3.2. Feature selection (FS) flows and filtered datasets

In this study, seven different FS flows were employed. Four applied the FS techniques (VAR, ANOVA, SFM, and RFE) individually, and three flows combined them sequentially in proposals for hierarchical pipelines (Pipeline 1, Pipeline 2, and Pipeline 3). The sequence of FS techniques in the pipeline considered each technique's computational cost and application. The objective was to reduce the number of features between the techniques. Pipeline 1 has two steps, combining the variance (VAR) filter and select from model (SFM). Pipeline 2 has three steps, the first uses the VAR, the second uses the ANOVA filter, and the third one uses the SFM. Pipeline 3 combined VAR, ANOVA, and the RFE.

Each FS Flow generated a filtered dataset with a reduced number of features used as input for the classification process with machine learning algorithms described in the next section. All seven FS flows and their filtered datasets are displayed in the center frames of Figure 2.

A grid of combinations was run on each FS flow to find the best output. The exhaustive tests were carried out with various configurations of each FS technique. VAR was tested with ten different threshold values (0.01, 0.02, ..., 0.09, 1). ANOVA was executed with 15 different sets of best attributes (k = 10, 13 ... 46, 49) with step = 3. The SFM and RFE were tested in Logistic Regression (LR), Balanced Random Forest (BRF), and Support Vector Machine (SVM). SFM was tested with eight different values (threshold = 0.2, 0.3,..., 0.8, 0.9) and RFE with 16 subsets, varying the number of features from 5 to 16, with a step of 1 feature.

2.3.3. Machine learning process

The classification model for SCD and non-SCD in ChHD was constructed using eight ML algorithms: 1) K-Nearest Neighbors (KNN), 2) Gradient Boosting (GB), 3) Logistic Regression (LR), 4) Naive Bayes (NB), 5) Support Vector Machine (SVM), 6) Balanced Random Forest (BRF), 7) Multilayer Perceptron (MLP), and 8) Catboost. Different types were used, such as probabilistic (LR and NB), neural networks (MLP), and tree algorithms (BRF and CatBoost).

For each filtered data set, a holdout of 80% was used for the training base and 20% for the test base. Tables 2 and 3 indicate the results of the test bases. ML algorithms were hyper-parameterized, 10-fold cross-validation techniques were used, and each was run 30 times to calculate the averages of six evaluation metrics: accuracy (ACC), the area under the ROC curve (AUC), precision, recall, F1-score, and their respective standard deviations. Due to the specificity of the data, this work uses only real data without using synthetic data generation techniques like SMOTE [43]. In all experiments, the algorithms were run with the following hyperparameters [40]:

- KNN: numeral of neighbors (3, 5, 7, 9, 11);
- GB: learning_rate (0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2), min_samples_split (12 samples spaced

Table 1. Five Groups and 57 Features divided into 5 groups of the dataset were used for SCD classification in ChHD with ML algorithms. The dataset contains information about Clinical Data, Heart Tests, Treatments, Cardiology Guidelines Classifications (CGC), and Cardiac Restitution Metrics (CRM) from ChHD's patients.

Group	Qnt	Attributes/Features	
Clinical Data	20	Age at Holter, Gender, Body Mass Index (BMI), Cancer, Systemic Hypertension, Type 2 Diabetes Mellitus, Pacemaker (non-dependent), Syncope, Atrial Fibrillation/Flutter, Chronic Renal Insufficiency, Dyslipidemia, Coronary heart disease, Heart failure, Cerebrovascular accident, Peripheral vascular disease, thyroid-stimulating hormone (TSH), Smoking, Alcoholism, Sedentary lifestyle, and Sudden Cardiac Death (SDC).	
Heart Tests	23	ECG	Primary Alteration, Interventricular Conduction Disorder (IVCDs), Atrioventricular Conduction Disorder, ≥ 3 second pause, Supraventricular Extrasystole (SVES), Ventricular Extrasystole (VE), Nonsustained Ventricular Tachycardia (nonsustained VT), and Electrical Inactive Area.
		ECHO	Left Atrial Diameter (LAD), Left Ventricular Diastolic Diameter (LVAD), Left Ventricular Systolic Diameter (LVSD), Ejection Fraction calculated by the method of Teichholz (EF-Teichholz), Classification, Diastolic Dysfunction, and Segmental dysfunction.
		Holter	Atrioventricular Conduction Disorder, Sinus Node Dysfunction, Atrial Fibrillation/Flutter, Average Heart Rate (AHR), Sustained Ventricular Tachycardia (SVT), Non-Sustained Ventricular Tachycardia (NSVT), Ventricular Extrasystole (VE), and Total Ventricular Extrasystole (TVE).
Treatments	3	ICD, Ablations and Amiodarone.	
Cardiology Guidelines Classification (CGC)	4	New York Heart Association (NYHA), Rassi Point, Rassi Escore, and Stage based on severity of cardiac involvement (2005 Guideline Classification).	
Cardiac Restitution Metrics (CRM)	7	TendQ 5th Percentile (s), TpeakQ 5th Percentile (s), Percentage (%) of beats with Qtend/TendQ > 1, Percentage (%) of beats with QTpeak/TpeakQ > 1,Upper 98% quantil of the Qtend/TendQ ratio, Upper 98% quantil of the QTpeak/TpeakQ ratio, and Tpeak-Tend 5th-Percentile (ms).	

evenly from the range starts from 0.1, ends at 0.5), min_samples_leaf (12 samples spaced evenly from the range starts from 0.1, ends at 0.5), max_depth (3, 5, 8), max_features (log2, sqrt), criterion_quality_split (friedman_mse, mean_squared _error), subsample (0.5, 0.618, 0.8, 0.85, 0.9, 0.95, 1.0), and n_estimators (10, 30, 70, 100);

- LR: C - regularization parameter (0, 0.01, 0.1, 1.0, 10, 100);
- NB: 100 samples spaced evenly on a log scale (starts from 0, ends at 9.);
- SVM: kernel (rbf, linear), C - regularization parameter (2^{i-5}, for i from 0 to 21, step 2), and gamma (only RBF Kernel: 2^{i-15}, for i from 0 to 19, step 2);
- BRF: criterion_quality_split (entropy, gini); max_depth (10 samples spaced evenly from the range starts from 10, ends at 1200), max_features (square root, binary logarithm, and all features), min_samples_leaf (4, 6, 8, and 12), min_samples_split (5, 7, 10, and 14), and n_estimators (10 samples spaced evenly from the range starts from 15, ends at 1200);
- MLP: hidden_layer_sizes (tests with 3, 2, and 1 hidden layers, varying the number of neurons in each, e.g., (200, 50, 30), (100, 50, 10), (100, 50), (200,100), (500, 250), (20,), (50,), (100,), (10,), (200,)), activation_function (tangent, rectified linear unit), solver_weight_optimization (stochastic gradient descent, Adam), alpha (0.0001, 0.005, 0.05) and learning_rate (constant, adaptive)
- Catboost: verbose: (0), depth (3, 4, 5, 8, 12), learning_rate (0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2), l2_leaf_reg (1, 4, 9), and iterations (70, 100, 150, 200, 300).

For each of the datasets filtered by the seven FS flows (variance, ANOVA, SFM, RFE, Pipeline 1, Pipeline 2, and Pipeline 3) used in this study, all 8 ML algorithms were executed, and the results of their metrics (ACC, AUC, recall, precision, and F1) were recorded. Altogether, the results of the 56 classification models (7 FS flows × 8 ML algorithms) were obtained and compared. Figure 2 shows the ML models in the second-to-rightmost frame.

Only the best-performing algorithms in each FS flow (variance, ANOVA, SFM, RFE, Pipeline 1, Pipeline 2, and Pipeline 3) were selected and compared. Table 2 shows the results obtained in scenario 1, and Table 3 shows the results obtained in scenario 2. The recall (sensitivity) was the principal metric for the choice because it identifies the highest number of hits for patients prone to SCD in ChHD. Accuracy and the smallest number of used attributes were also considered. Accuracy keeps patients from getting unnecessary treatments, and having the fewest possible attributes makes the proposed model easy to use.

Table 2. The results (ACC, AUC, recall, precision, and F1) of the best-performing machine learning models from feature selection (FS) flows in scenario 1 with patients' clinical data and CRM features (57 attributes in total) for predicting the risk of SCD in ChHD.

FS Flow	Algorithm	Features	ACC	AUC	Recall	Precision	F1
ANOVA	NB	24	76.36 ± 6.55	78.53 ± 6.5	86.46 ± 8.2	63.11 ± 7.09	72.79 ± 6.9
Pipeline 1	NB	16	77.12 ± 5.94	79.66 ± 5.8	88.96 ± 9.2	63.70 ± 6.64	73.93 ± 6.3
Pipeline 2	NB	6	77.80 ± 6.47	78.99 ± 6.9	83.33 ± 14.5	66.67 ± 9.29	72.88 ± 8.6
Pipeline 3	NB	13	77.80 ± 6.08	80.55 ± 5.6	90.63 ± 7.3	64.32 ± 7.42	74.98 ± 6.2
RFE	NB	13	77.73 ± 6.44	79.69 ± 5.7	86.88 ± 10.2	65.59 ± 8.39	74.08 ± 6.2
SFM	NB	18	74.62 ± 11.33	77.83 ± 8.7	89.58 ± 7.0	62.34 ± 10.25	72.82 ± 7.7
VAR	BRF	56	79.17 ± 5.47	80.28 ± 5.1	84.38 ± 8.0	67.78 ± 7.77	74.78 ± 5.8

Table 3. Result of the best-performing algorithms for each feature selection flow in scenario 2 with 50 attributes (without 7 Cardiac Restitution Metrics features) for machine learning-based SCD prediction in ChHD.

FS Flow	Algorithm	Features	ACC	AUC	Recall	Precision	F1
ANOVA	NB	29	71.64 ± 5.83	75.57 ± 4.6	83.54 ± 9.6	47.66 ± 6.79	60.18 ± 5.4
Pipeline 1	NB	10	73.28 ± 10.7	77.97 ± 7.1	87.50 ± 10.0	51.12 ± 11.0	63.49 ± 8.4
Pipeline 2	NB	4	71.96 ± 10.54	77.70 ± 6.8	89.38 ± 9.0	49.34 ± 9.60	62.75 ± 7.6
Pipeline 3	BRF	5	78.34 ± 5.86	80.63 ± 5.7	89.58 ± 8.3	64.93 ± 7.26	75.02 ± 6.3
RFE	NB	11	74.87 ± 6.17	78.14 ± 5.7	84.79 ± 12.7	51.41 ± 7.78	63.32 ± 7.0
SFM	NB	7	74.23 ± 9.58	79.78 ± 7.0	91.04 ± 8.2	51.53 ± 10.09	65.14 ± 8.4
VAR	BRF	56	79.37 ± 5.85	80.40 ± 5.6	82.50 ± 8.6	57.34 ± 7.86	67.30 ± 7.0

3. Experimental results and discussion

The test base (20% dataset) results of the machine learning models to predict SCD in patients with ChHD are presented. Table 2 shows the mean and standard deviation values of the metrics of the best-performing algorithms for each of the FS methods in scenario 1 (S1), including 7 CRM features extracted from ECG-Holter (57 features in total). Noting that VAR removes a single feature, its performance is comparable to that of the full-feature case, i.e., no-use feature selection. With a combination of feature selection techniques within flows, the correlated features were removed, leaving only those with little or no correlation. Except for using only the variance filter (VAR) with BRF, Naive Bayes (NB) performed the best in almost all flows. The NB algorithm classification model found the best results, which was most suited to scenario 1 [44].

Table 4. The number of features picked by each feature group from feature selection (FS) flows in scenario 1 with patients' clinical data and CRM features (57 attributes in total) for predicting the risk of SCD in ChHD.

FS Flow	Algorithm	Clinica Data	Classification	Treatments	CRM	Heart Tests	**Total**
ANOVA	NB	7	4	2	1	10	**24**
Pipeline 1	NB	3	3	2	2	6	**16**
Pipeline 2	NB	1	2	0	0	3	**6**
Pipeline 3	NB	0	4	1	1	7	**13**
RFE	NB	3	1	1	3	5	**13**
SFM	NB	3	4	2	3	6	**18**
VAR	BRF	19	4	3	7	22	**56**

Table 2 displays that the proposed hierarchical technique Pipeline 3 with NB using 13 attributes achieved the best results for recall, AUC, and F1-Score, with particular attention to the 90.63% recall value. Other proposed hierarchical techniques, Pipeline 1, and SFM, both with NB, had very similar results for recall but with a large number of features. The VAR method with the BRF algorithm obtained 67.78% precision, but the other methods have similar results in this metric.

One can observe that the Pipeline 3 hierarchical technique with NB had the second-best result for precision using only six attributes. The proposed Pipeline 3, which combines the filter variance,

ANOVA, and RFE techniques, improved recall by more than 4% and decreased its standard deviation by more than 28% compared to the result of using RFE alone.

Table 4 presents the classification models and the number of features per group. The model with the Pipeline 3 that obtained the best performance for recall selected 13 attributes:

- 4 Cardiology Guidelines Classification (CGC): NYHA, Rassi points, Rassi score, 2005 Guideline Classification;
- 1 Treatment: Amiodarone;
- 1 CRM: Tpeak-Tend 5th-Percentile (ms);
- 7 Heart Tests: NSVT, TVE, LVAD, F-Teichholz, Classification, Diastolic Dysfunction, and Segmental dysfunction

The Pipeline 2 model with the NB selected 6 attributes:

- 1 Clinical: Sincope;
- 2 Cardiology Guidelines Classification (CGC): Rassi score, 2005 Guideline;
- 3 Heart Tests: NSVT, TVE, Classification.

One can see that CRM feature are present in the subset of best features of almost all of the SCD classification models. In the RFE and SFM methods, 3 of the seven cardiac restitution metrics were selected among the features of the highest relevance: TendQ 5th Percentile (s), Percentage (%) of beats with Qtend/TendQ > 1, and Tpeak-Tend 5th-Percentile (ms).

The results indicate an optimized set of attributes for using a possible classification model of SCD in patients with ChHD to support the therapeutic decision-making process. The "Tpeak-Tend 5th-Percentile (ms)" was selected among the most relevant features, showing the relevance of CRM calculation from ECG signals. Two reasons can probably explain this. First, it is well established that the electrophysiological heterogeneity resulting from the cardiac remodeling induced by autonomic dysfunction quantitatively favors the presence of fast ventricular arrhythmias and SCD in ChHD. Furthermore, these individuals do not have the parasympathetic protection [6], and the increase in norepinephrine levels [45] might be responsible for fast ventricular arrhythmia. The second possible explanation is that the vulnerability associated with the CRMs is not solely a function of a static set of electrophysiological properties but also depends on changes that occur over multiple time scales ranging from seconds to hours [46].

Table 5 displays studies relating to SCD in the context of ChHD and general cardiology (non-ChHD). Due to the scarcity of studies on SCD in ChHD, works with ML related to SCD in non-ChHD were included. All the studies satisfy the criteria of predicting SCD in patients in a way that goes back more than a day and with time spent watching the patients [8, 13, 26]. Due to a lack of access to a universal health system and underreporting of the disease, the majority of non-ChHD studies have a larger number of participants (N) than ChHD research, demonstrating the importance of N in this search. In the non-ChHD context, Random Forest (RF) is one of the most commonly used ML models. A variation of this model (Balanced RF) also appears in VAR Fs Flow, shown in Table 4. In relation to the number of features used to predict SCD, in the context of ChHD, the model by Souza et al. [8] uses only 4, do not use ML, and has several limitations. Alberto [26] used eight features but had only 82 participants. The no-ChHD research has a variance in the number of features (4, 7, 8, 9, 10, 12, 15,

Table 5. SCD-related works in the contexts of ChHD and non-ChHD. The table displays the study, the context, the number of research participants (N), the ML model, and the number of attributes required to predict SCD.

Study	Participants	ChHD Context	Machine Learning Model	Features
Souza et al. [8]	373	Yes	Classical Linear Regression (Not using ML)	4
Tse et al. [47]	376	No	Non-negative matrix factorization	4
Lyon et al. [48]	123	No	Density-based clustering	7
Nakajima et al. [49]	526	No	Random Forest, KNN, Gradient Boosted	8
Alberto et al. [26],	82	Yes	KNN	8
Shakibfar et al. [50]	19,935	No	Random Forest	9
Atallah et al. [51]	288	No	Random Forest	10
Rodriguez et al. [52]	140	No	SVM	12
This Work (Pipeline 3)	**218**	**Yes**	**Naive Bayes**	**13**
Vergara et al. [53]	502	No	Random Forest	15
Lee et al. [54]	516	No	SVM	26
Zoni-Berisso et al. [55]	404	No	Mandansky artificial neural network	61
Goldstein et al. [56]	1628	No	Random Forest and KNN	72

26, 61, 72). Pipeline 2, our second-best result, used six features and is the third method with a smaller number of features, including ChHD and non-ChHD contexts. A possible hypothesis for some works using fewer features is the amount of data available, which does not happen in the context of SCD in ChHD.

The works shown in Table 5 have their results presented by several metrics (accuracy, AUC, sensitivity, specificity, or F1-score) [13]. For comparative purposes, Table 6 shows only the perfomance of these works with the same parameter (sensitivity) as the current study. It is observed that the results obtained in this present research are similar to those obtained in non-ChHD context studies, presenting the second highest sensitivity. Despite the differences in terms of number of features, number of participants, and ML method applied, our results are relevant to the scenario of SCD in ChHD.

In scenario 2, excluded 7 CRM features extracted from ECG-Holter (remaining 51 features in total), the experiments were repeated, running the eight algorithms on each FS technique. Table 3 shows the results for the test base (20% dataset). The NB algorithm, similar to scenario 1, performed better than the other ML algorithms in most FS flows. BRF performed better in the models using only the variance and the Pipeline 3 hierarchical techniques. Both algorithms ran with the same hyperparameters as in scenario 1. The results in S1 and S2 scenarios of all algorithms in each FS technique are available in this online document.

Table 6. Performance of the machine learning model of studies on predicting SCD risk, considering sensitivity as a reference parameter.

Study	ChHD Context	Recall (Sensitivity)
Tse *et al.* [47]	No	70.00%
Zoni-Berisso et al. [55]	No	96.00%
Lee et al. [54]	No	89.00%
Atallah et al. [51]	No	88.40%
This Work (Pipeline 3)	**Yes**	**90.63%**

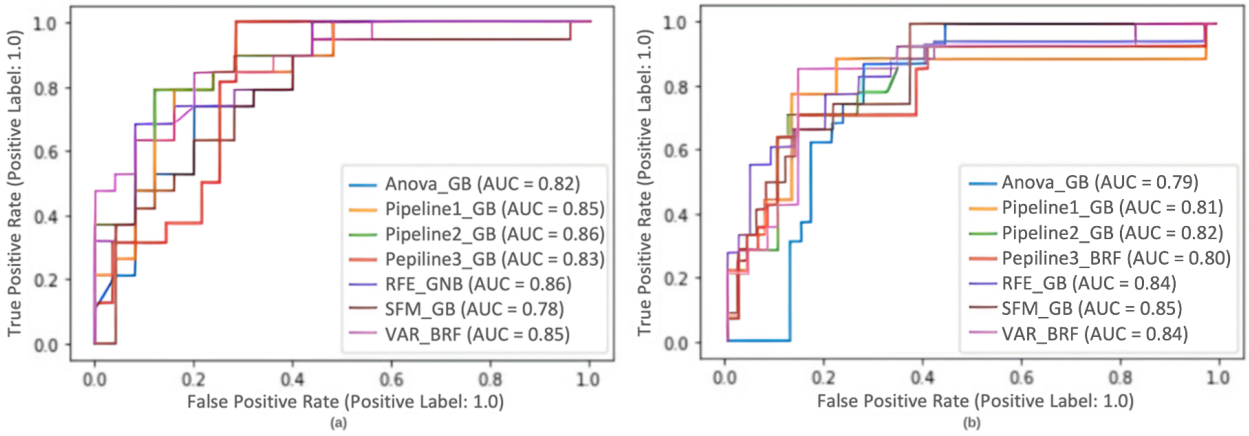

Figure 3. ROC curves of (a) scenario 1 (218 patients and 57 features, including CRM features) and (b) scenario 2 (315 patients and 50 features, excluding CRM features) from the classification algorithms with the best performance in each FS flow.

It can be seen from Table 3 that pipelines Pipeline 2 with NB, Pipeline 3 with BRF, and SFM with BRF had similar recall results in this second scenario (S2). The model with Pipeline 3 had the best accuracy result, with a value of 64.93%, improving the second-best result by 7%. It still had the best F1-score (75.02%), the best AUC (78.34%), and selected five features (1 Clinical, 2 Heart Tests, and 2 Classification), being the second smallest subset of attributes. Compared to S1, there was a significant decrease in precision and F1-score in all approaches run in S2, except Pipeline 3 with BRF. One hypothesis may be the absence of the use of CRM features.

The ROC curves were plotted in scenario 1 (Figure 3a) and scenario 2 (Figure 3b). The graphs show the ROC curve of one of the 30 repetitions of the best-performed ML algorithm in each FS technique, illustrated in Tables 2 and 3. The values obtained are within the statistical range of the metric.

The strengths of our study include the vital status of each participant determined by SIM and the use of artificial intelligence. Our work is perhaps best understood in the context of its limitations. Lacks external validity. Further studies are desired to evaluate its incorporation in the clinical. Another possible drawback is the question of how far our findings could be applied to the general CD population. Although our patients constituted an urban cohort from a CD reference center, their baseline characteristics were similar to those reported from rural [57] and urban [8] cohorts of Chagas' endemic areas. Thus, we believe that our cohort represents the factual scenario in a population with CD, where risk scores and prognostic evaluations are often studied.

4. Conclusions

This research presents an SCD Multiparametric Classification System for Chagas Heart Disease'sre Patients based on clinical data and 24-hour ECG Monitoring, which achieved 90.63% recall (sensitivity) and 80.55% AUC with the proposed approach called Pipeline 3 using Naive Bayes (NB) machine learning algorithm. This proposed approach combines three different feature selection techniques (VAR, ANOVA, and RFE) to find an optimized relevant subset of 13 features (among a total of 57 attributes) as input to a classification algorithm. This result is a significant achievement in addressing a

requirement from the team of medical specialists, who considered the number of original attributes too large and hard to interpret. The Tpeak-Tend 5th-Percentile (ms) feature was selected among the most relevant, showing the relevance of the CRM feature from ECG signals.

Our findings suggest that using the ML technique in ChHD may be an additional helpful tool to identify individuals with increased risk for SCD who may benefit from ICD implantation, possibly improving the selection of individuals for ICD. Our results have potential socio-economic implications when limited resources have to be allocated to the appropriate patients and may help guide ICD implantation in the Brazilian Unified Health System (SUS/Brazil) and among doctors in other countries highly affected by Chagas disease. Future research may conduct experiments with the entire 24-hour ECG-holter signal. Additionally, we may use alternative methods to address the problem, such as extracting other features from ECG signals (e.g., heart rate variability metrics) and various types of deep learning, such as convolutional neural networks (CNN), recurrent neural networks (RNN), or long short-term memory (LSTM).

Acknowledgments

The authors thank the support of the Federal Institute of Education and Technology of Ceará, Ceará Foundation for Scientific and Technological Development Support - Funcap (PS1-0186-00439.01.00/21, BP4-0172-00075.01.00/20), and the Brazilian National Council for Scientific and Technological Development (CNPQ).

Conflict of interest

The authors declare no conflict of interest.

References

1. A. F. Members, S. G. Priori, C. Blomström-Lundqvist, A. Mazzanti, N. Blom, M. Borggrefe, et al., 2015 esc guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The task force for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death of the european society of cardiology (esc) endorsed by: Association for european paediatric and congenital cardiology (aepc), *EP Europace*, **17** (2015), 1601–1687. https://doi.org/10.1093/europace/euv319

2. A. S. Adabag, R. V. Luepker, V. L. Roger, B. J. Gersh, Sudden cardiac death: epidemiology and risk factors, *Nat. Rev. Cardiol.*, **7** (2010), 216–225. https://doi.org/10.1038/nrcardio.2010.3

3. World Health Organization, Fourth who report on neglected tropical diseases: Integrating neglected tropical diseases into global health and development, *IV WHO Report on Neglected Tropical Diseases*, 2017. Available from: https://apps.who.int/iris/handle/10665/255011.

4. A. Rassi Jr, A. Rassi, W. C. Little, S. S. Xavier, S. G. Rassi, A. G. Rassi, et al., Development and validation of a risk score for predicting death in chagas' heart disease, *N. Engl. J. Med.*, **355** (2006), 799–808. https://doi.org/10.1056/NEJMoa053241

5. World Health Organization, *Chagas Disease (also known as American Trypanosomiasis)*, 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis), Accesse date: 10 January 2022.
6. R. C. Pedrosa, Dysautonomic arrhythmogenesis: A working hypothesis in chronic chagas cardiomyopathy, *Int. J. Cardiovasc. Sci.*, **33** (2020), 713–720. https://doi.org/10.36660/ijcs.20200169
7. J. A. Marin-Neto, E. Cunha-Neto, B. C. Maciel, M. V. Simões, Pathogenesis of chronic chagas heart disease, *Circulation*, **115** (2007), 1109–1123. https://doi.org/10.1161/CIRCULATIONAHA.106.624296
8. A. C. J. de Souza, G. Salles, A. M. Hasslocher-Moreno, A. S. de Sousa, P. E. A. A. do Brasil, R. M. Saraiva, et al., Development of a risk score to predict sudden death in patients with chaga's heart disease, *Int. J. Cardiol.*, **187** (2015), 700–704. https://doi.org/10.1016/j.ijcard.2015.03.372
9. J. A. Pérez-Molina, I. Molina, Chagas disease cardiomyopathy treatment remains a challenge–authors' reply, *Lancet*, **391** (2018), 2209–2210. https://doi.org/10.1016/S0140-6736(18)30776-1
10. F. M. Rassi, L. Minohara, A. Rassi, L. C. L. Correia, J. A. Marin-Neto, A. Rassi, et al., Systematic review and meta-analysis of clinical outcome after implantable cardioverter-defibrillator therapy in patients with chagas heart disease, *JACC: Clin. Electrophysiol.*, **5** (2019), 1213–1223. https://doi.org/10.1016/j.jacep.2019.07.003
11. F. Lopez-Jimenez, Z. Attia, A. M. Arruda-Olson, R. Carter, P. Chareonthaitawee, H. Jouni, et al., Artificial intelligence in cardiology: present and future, *Mayo Clin. Proc.*, **95** (2020), 1015–1039. https://doi.org/10.1016/j.mayocp.2020.01.038
12. F. Ali, S. El-Sappagh, S. R. Islam, D. Kwak, A. Ali, M. Imran, et al., A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion, *Inf. Fusion*, **63** (2020), 208–222. https://doi.org/10.1016/j.inffus.2020.06.008
13. J. Barker, X. Li, S. Khavandi, D. Koeckerling, A. Mavilakandy, C. Pepper, et al., Machine learning in sudden cardiac death risk prediction: a systematic review, *Europace*, **24** (2022), 1777–1787. https://doi.org/10.1093/europace/euac135
14. P. Pace, G. Aloi, R. Gravina, G. Caliciuri, G. Fortino, A. Liotta, An edge-based architecture to support efficient applications for healthcare industry 4.0, *IEEE Trans. Ind. Inf.*, **15** (2018), 481–489. https://doi.org/10.1109/TII.2018.2843169
15. G. Alfian, M. Syafrudin, M. F. Ijaz, M. A. Syaekhoni, N. L. Fitriyani, J. Rhee, A personalized healthcare monitoring system for diabetic patients by utilizing ble-based sensors and real-time data processing, *Sensors*, **18** (2018), 2183. https://doi.org/10.3390/s18072183
16. J. A. L. Marques, T. Han, W. Wu, J. P. do Vale Madeiro, A. V. L. Neto, R. Gravina, et al., Iot-based smart health system for ambulatory maternal and fetal monitoring, *IEEE Internet Things J.*, **8** (2020), 16814–16824. https://doi.org/10.1109/JIOT.2020.3037759
17. D. L. T. Wong, J. Yu, Y. Li, C. J. Deepu, D. H. Ngo, C. Zhou, et al., An integrated wearable wireless vital signs biosensor for continuous inpatient monitoring, *IEEE Sens. J.*, **20** (2020), 448–462. https://doi.org/10.1109/JSEN.2019.2942099

18. J. C. P. Dias, A. N. Ramos, E. D. Gontijo, A. Luquetti, M. A. Shikanai-Yasuda, J. R. Coura, et al., 2nd brazilian consensus on chagas disease, 2015, *Rev. Soc. Bras. Med. Trop.*, **49** (2016), 03–60. https://doi.org/10.1590/0037-8682-0505-2016

19. M. C. P. Nunes, A. Z. Beaton, H. Acquatella, C. Bern, A. F. Bolger, L. E. Echeverría, et al., *Circulation*, **138** (2018), e169–e209. https://doi.org/10.1161/CIR.0000000000000599

20. R. J. Moll-Bernardes, P. H. Rosado-de Castro, G. C. Camargo, F. S. N. S. Mendes, A. S. Brito, A. S. Sousa, New imaging parameters to predict sudden cardiac death in chagas disease, *Trop. Med. Infect. Dis.*, **5** (2020), 74. https://doi.org/10.3390/tropicalmed5020074

21. N. Sharma, K. Saroha, Study of dimension reduction methodologies in data mining, in *International Conference on Computing, Communication & Automation*, IEEE, (2015), 133–137. https://doi.org/10.1109/CCAA.2015.7148359

22. S. Velliangiri, S. Alagumuthukrishnan, I. T. J. Swamidason, A review of dimensionality reduction techniques for efficient computation, *Procedia Comput. Sci.*, **165** (2019), 104–111. https://doi.org/10.1016/j.procs.2020.01.079

23. C. Antzelevitch, S. Sicouri, J. M. Di Diego, A. Burashnikov, S. Viskin, W. Shimizu, et al., Does tpeak–tend provide an index of transmural dispersion of repolarization, *Heart Rhythm*, **4** (2007), 1114–1116. https://doi.org/10.1016/j.hrthm.2007.05.028

24. A. A. Fossa, M. Zhou, Assessing QT prolongation and electrocardiography restitution using a beat-to-beat method, *Cardiol. J.*, **17** (2010), 230–243.

25. W. B. Nicolson, G. P. McCann, M. I. Smith, A. J. Sandilands, P. J. Stafford, F. S. Schlindwein, et al., Prospective evaluation of two novel ecg-based restitution biomarkers for prediction of sudden cardiac death risk in ischaemic cardiomyopathy, *Heart*, **100** (2014), 1878–1885. http://dx.doi.org/10.1136/heartjnl-2014-305672

26. A. C. Alberto, R. C. Pedrosa, V. Zarzoso, J. Nadal, Association between circadian holter ecg changes and sudden cardiac death in patients with chagas heart disease, *Physiol. Meas.*, **41** (2020), 025006. https://doi.org/10.1088/1361-6579/ab6ebc

27. L. E. Hinkle Jr, H. T. Thaler, Clinical classification of cardiac deaths, *Circulation*, **65** (1982), 457–464. https://doi.org/10.1161/01.CIR.65.3.457

28. DATASUS (Departamento de Informática do Sistema Único de Saúde), TabNet Win32 3.0: Morbidade Hospitalar do SUS por Causas Externas por local de internação - Brasil, 2022. Available from: http://tabnet.datasus.gov.br, Accesse date: 10 July 2022.

29. L. Capuani, A. L. Bierrenbach, F. Abreu, P. L. Takecian, J. E. Ferreira, E. C. Sabino, Accuracy of a probabilistic record-linkage methodology used to track blood donors in the mortality information system database, *Cad. Saúde Pública*, **30** (2014), 1623–1632. https://doi.org/10.1590/0102-311X00024914

30. R. Laurenti, M. de Mello Jorge, S. Gotlieb, Underlying cause-of-death mortality statistics: considering the reliability of data, *Pan Amer. J. Public Health*, **23** (2008), 349–356. https://doi.org/10.1590/s1020-49892008000500007

31. New York Heart Association, The Criteria Committee of the New York Heart Association, *Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels*, 9th edition, Boston, Massachusetts, Little, Brown & Co, 1994.

32. J. P. de Andrade, J. A. Marin Neto, A. A. V. de Paola, F. Vilas-Boas, G. M. M. Oliveira, F. Bacal, et al., I latin american guidelines for the diagnosis and treatment of chagas' heart disease: executive summary, *Arq. Bras. Cardiol.*, **96** (2011), 434–442. https://doi.org/10.1590/S0066-782X2011000600002

33. H. Liu, H. Motoda, *Computational Methods of Feature Selection*, CRC Press, 2007.

34. J. C. Ang, A. Mirzal, H. Haron, H. N. A. Hamed, Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **13** (2015), 971–989. https://doi.org/10.1109/TCBB.2015.2478454

35. R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, J. Saeed, A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction, *J. Appl. Sci. Technol. Trends*, **1** (2020), 56–70. https://doi.org/10.38094/jastt1224

36. J. Richter, H. Kotthaus, B. Bischl, P. Marwedel, J. Rahnenführer, M. Lang, Faster model-based optimization through resource-aware scheduling strategies, in *International Conference on Learning and Intelligent Optimization*, Springer, (2016), 267–273. https://doi.org/10.1007/978-3-319-50349-3_22

37. Q. Chen, Z. Meng, X. Liu, Q. Jin, R. Su, Decision variants for the automatic determination of optimal feature subset in RF-RFE, *Genes*, **9** (2018), 301. https://doi.org/10.3390/genes9060301

38. P. Schwartz, M. La Rovere, E. Vanoli, Autonomic nervous system and sudden cardiac death. experimental basis and clinical observations for post-myocardial infarction risk stratification, *Circulation*, Suppl., **85** (1992), I77–91.

39. J. P. V. Madeiro, E. Santos, P. C. Cortez, J. H. S. Felix, F. S. Schlindwein, Evaluating gaussian and rayleigh-based mathematical models for t and p-waves in ECG, *IEEE Lat. Am. Trans.*, **15** (2017), 843–853. https://doi.org/10.1109/TLA.2017.7910197

40. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in python, *J. Mach. Learn. Res.*, **12** (2011), 2825–2830.

41. B. Lantz, *Machine Learning with R: Expert Techniques for Predictive Modeling*, Packt publishing ltd, 2019.

42. J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and, Techniques*, 3rd edition, 2011.

43. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority oversampling technique, *J. Artif. Intell. Res.*, **16** (2002), 321–357. https://doi.org/10.1613/jair.953

44. I. Rish, An empirical study of the naive bayes classifier, in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, **3** (2001), 41–46.

45. A. B. Cunha, D. M. Cunha, R. C. Pedrosa, F. Flammini, A. Silva, E. A. Saad, et al., Norepinephrine and heart rate variability: a marker of dysautonomia in chronic chagas cardiopathy, *Port. J. Cardiol.: Off. J. Port. Soc. Cardiol.*, **22** (2003), 29–52.

46. G. A. Ng, A. Mistry, X. Li, F. S. Schlindwein, W. B. Nicolson, Lifemap: towards the development of a new technology in sudden cardiac death risk stratification for clinical use, *EP Europace*, **20** (2018), f162–f170. https://doi.org/10.1093/europace/euy080
47. G. Tse, J. Zhou, S. Lee, T. Liu, G. Bazoukis, P. Mililis, et al., Incorporating latent variables using nonnegative matrix factorization improves risk stratification in brugada syndrome, *J. Am. Heart Assoc.*, **9** (2020), e012714. https://doi.org/10.1161/JAHA.119.012714
48. A. Lyon, R. Ariga, A. Mincholé, M. Mahmod, E. Ormondroyd, P. Laguna, et al., Distinct ECG phenotypes identified in hypertrophic cardiomyopathy using machine learning associate with arrhythmic risk markers, *Front. Physiol.*, **9** (2018), 213. https://doi.org/10.3389/fphys.2018.00213
49. K. Nakajima, T. Nakata, T. Doi, H. Tada, K. Maruyama, Machine learning-based risk model using 123 i-metaiodobenzylguanidine to differentially predict modes of cardiac death in heart failure, *J. Nucl. Cardiol.*, **29** (2020), 1–12. https://doi.org/10.1007/s12350-020-02173-6
50. S. Shakibfar, O. Krause, C. Lund-Andersen, A. Aranda, J. Moll, T. O. Andersen, et al., Predicting electrical storms by remote monitoring of implantable cardioverter-defibrillator patients using machine learning, *EP Europace*, **21** (2019), 268–274. https://doi.org/10.1093/europace/euy257
51. J. Atallah, M. C. G. Corcia, E. P. Walsh, Participating Members of the Pediatric and Congenital Electrophysiology Society, Ventricular arrhythmia and life-threatening events in patients with repaired tetralogy of fallot, *Am. J. Cardiol.*, **132** (2020), 126–132. https://doi.org/10.1016/j.amjcard.2020.07.012
52. J. Rodriguez, S. Schulz, B. F. Giraldo, A. Voss, Risk stratification in idiopathic dilated cardiomyopathy patients using cardiovascular coupling analysis, *Front. Physiol.*, **10** (2019), 841. https://doi.org/10.3389/fphys.2019.00841
53. P. Vergara, W. S. Tzou, R. Tung, C. Brombin, A. Nonis, M. Vaseghi, et al., Predictive score for identifying survival and recurrence risk profiles in patients undergoing ventricular tachycardia ablation: the I-VT score, *Circ.: Arrhythmia Electrophysiol.*, **11** (2018), e006730. https://doi.org/10.1161/CIRCEP.118.006730
54. S. Lee, J. Zhou, K. H. C. Li, K. S. K. Leung, I. Lakhani, T. Liu, et al., Territory-wide cohort study of brugada syndrome in hong kong: predictors of long-term outcomes using random survival forests and non-negative matrix factorisation, *Open Heart*, **8** (2021), e001505. http://dx.doi.org/10.1136/openhrt-2020-001505
55. M. Zoni-Berisso, D. Molini, S. Viani, G. S. Mela, L. Delfino, Noninvasive prediction of sudden death and sustained ventricular tachycardia after acute myocardial infarction using a neural network algorithm, *Ital. Heart J.: Off. J. Ital. Fed. Cardiol.*, **2** (2001), 612–620.
56. B. A. Goldstein, T. I. Chang, A. A. Mitani, T. L. Assimes, W. C. Winkelmayer, Near-term prediction of sudden cardiac death in older hemodialysis patients using electronic health records, *Clin. J. Amer. Soc. Nephrol.*, **9** (2014), 82–91. http://dx.doi.org/10.2215/CJN.03050313

57. M. S. Marcolino, D. M. Palhares, L. R. Ferreira, A. L. Ribeiro, Electrocardiogram and chagas disease: a large population database of primary care patients, *Global Heart*, **10** (2015), 167–172. https://doi.org/10.1016/j.gheart.2015.07.001

AIMS Press