Feature Extraction Strategies for Predicting Reduced Left Ventricular Ejection Fraction in Chagas Disease Patients

João Paulo do Vale Madeiro[1], Luís Otavio Rigo Jr[2], Roberto C Pedrosa[3]

[1] Federal University of Ceará, Fortaleza, Brazil
[2] University for the International Integration of the Afro-Brazilian Lusophony, Redenção, Brazil
[3] Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Abstract

Cardiovascular problems are the most important manifestation of the Chagas disease (CD), which can cause arrhythmias, heart failure and thromboembolisms. Echocardiography (ECO) provides diagnostic tools, mainly considering left ventricular systolic dysfunction (LVSD), being more expensive and difficult to access than electrocardiography (ECG).

The present work aims to investigate strategies for extracting ECG parameters for predicting LVSD, defined as a left ventricular ejection fraction (LVEF) determined by ECO below a given threshold, which may be 35, 40, 45, 50 or 55%. We process a dataset containing single-lead ECG holter signals from 219 CD patients obtained from University Hospital Clementino Fraga Filho – Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. The approach proposes the segmentation of the original signals in intervals during: (a) 5 minutes; (b) 10 minutes; (c) 15 minutes and (d) 30 minutes. For each scenario, we obtain statiscal measures related to waveform amplitudes and durations, and also statistical measures related to wavelet decomposition coefficients, heart rate variability parameters and non-linear analysis. Then, a set of Machine Learning (ML) algorithms are applied for each scenario to discriminate between LVSD patients and non-LVSD patients.

As results, we obtain the highest performance for 15-minute ECG intervals: recall 72% +- 9% and Area under ROC curve 0,75 +- 0,09 for Gradient Boosting. The results indicate the feasibility of using short-term one-channel ECG signals to predict reduced LVEF.

Keywords—**Chagas disease. Ejection fraction. ECG.**

1. Introduction

Chagas disease (CD), caused by the protozoan *Trypanosoma cruzi*, is one of the 17 neglected tropical diseases, according to the World Health Organization, and continues to be a chronic health problem, despite the control of its transmission. As a frequent and severe manifestation of CD, the chronic Chagas cardiomyopathy (CCC) is a leading cause of morbidity and death in South America, and its symptoms arise from heart failure, cardiac arrhythmias, and thromboembolism [1]. Heart failure is a chronic and progressive disease, affecting around 37.7 million people worldwide, and a similar number of people have undetected or asymptomatic left ventricular systolic dysfunction (LVSD) [2]. The research of algorithms for LVSD detection based on ECG processing is highly useful and attractive because ECGs are accessible, inexpensive, and ubiquitous.

A systematic review in [3] presents a detailed comparison of fifteen approaches concerning artificial intelligence-enabled electrocardiography (AIeECG) to detect LVSD, exploring similarities and differences in relation to thresholds for left ventricular ejection fraction (LVEF), study populations and the use of natriuretic peptides. From the studies considered eligible for that investigation, just one approach has applied the AI algorithm using a single-lead ECG, achieving areas under the curves of 0.874 and 0.929 during the internal and external validation, respectively, considering hospitalized population that does not belong to a specific group and the threshold of LVEF $<$40% [4].

At this context, the present work address the issue of predicting LVSD within patients with Chagas disease. The proposed algorithm aims to provide a tool for screening CD patients for early detection of LVSD.

2. Methodology

2.1. Data description

The Clementino Fraga Filho University Hospital of the Federal University of Rio de Janeiro (HUCFF-UFRJ) is a regional reference center in Brazil concerning Chagas disease. The holter ECG signals, which were sampled at a 128-Hz sampling frequency, were acquired from 219 patients with CCC (chronic chagasic cardiomyopathy), collected between 1992 and 2013. As for gender, the signals

belong to 125 (57.1%) women and 94 (42.9%) men. Regarding age, 1.3% of the patients are between 15 and 20 years old, 7.1% between 20 and 30 years old, 25% between 30 and 40 years, 34.4% between 40 and 50 years, 30.8% between 50 and 60 years, and 3.9% between 60 and 70 years old. Of all the analyzed signals, only one belonged to a patient who had some other type of heart disease. Regarding the target class, 75 (20%) of the signals were obtained from patients with left ventricular ejection fraction lower than 40%, 159 (42,5%) were obtained from patients with LVEF lower than 45%, 180 (48,1%) from patients with LVEF lower than 50%, and 203 (54,28%) from patients with LVEF lower than 55%. The local ethics committee approved the research (number 45360915.1.1001.5262).

2.2. Feature Extraction

At this stage, different scenarios are proposed to capture parameters that are potentially sensitive to the presence of left ventricular systolic dysfunction. These scenarios concern the duration of the analyzed ECG signal interval, which can be: (a) 5 minutes; (b) 10 minutes; (c) 15 minutes; and (d) 30 minutes. Thus, initially applying the discrete wavelet transform (DWT), with Daubechies-4 as mother function, we obtain eight levels of decomposition. For each level of decomposition, we obtain for each group of detail coefficients and for the last group of approximation coefficients the following statistical metrics: mean, variance, maximum value, minimum value, skewness, and the 1st, 2nd, 5th, 10th, 25th, 50th, 75th, 90th, 95th, 98th and 99th percentiles. Furthermore, we also obtain as features the sum of the amplitudes of the coefficients of the fast Fourier transform of the ECG interval in the following ranges: 0-1.5Hz; 1.5 - 4Hz; 4 - 8Hz; 8 - 20Hz; 20 - 50Hz; 50 - 64Hz.

Next, we move on to obtaining the second group of parameters. First, we perform the detection and segmentation of the QRS complex. Concerning QRS detection, we have adopted our already validated approach, which is based on Hilbert and Wavelet transforms, first-derivative and adaptive threshold technique [5]. For delineation, that is, the precise detection of the onset and offset of each QRS complex, we adopted the phasor transform technique [6], which achieves to convert each instantaneous ECG sample into a phasor. Thus, considering the ECG signal of N samples as $x[n]$, the phasorial transform is defined from the following equations

$$y[n] = R_v + j.x[n]. \tag{1}$$

$$\phi[n] = tan^{-1}(\frac{|x[n]|}{Rv}), \tag{2}$$

$$M[n] = \sqrt{R_v^2 + x[n]^2}, \tag{3}$$

where the value of R_v determines the degree with which waveforms are enhanced in the phasorial signal $\phi[n]$. Due to the sensitivity of the non-linear transformation given by the equation 3, we can associate the QRS start and end points with local minima or edges of subtle variations for $\phi[n]$.

Next, we detect T-wave peak and T-wave end, also based on the phasorial transform applied over each interval starting at each QRS offset and during half of the current R-R interval. Taking the R_v parameter as 0.1, the phasor transform of the aforementioned searching interval is obtained, and then we identify among the samples with the highest amplitudes within $\phi[n]$ signal a local maximum of $M(n)$ for T-wave peak. For T-wave end detection, we compute de first derivative of the signal $\phi[n]$ within a searching window beginning at the T-wave peak location and ending 90 ms after. Then, starting at the sample corresponding to the minimum of the derivative of $\phi[n]$ and ending at the sample corresponding to the maximum of the derivative of $\phi[n]$, we associate the T-wave end location to the sample corresponding to the first detected zero-crossing.

Once we have delineated each QRS complex and each T-wave, we compute the following measures: each beat-to-beat interval (R-R interval), duration of each QRS complex, amplitude of each QRS complex, amplitude of each T-wave, duration of each interval from a given QRS-onset to the subsequent T-wave end (named QTend), duration of each interval from a given QRS-onset to the subsequent T-wave peak (named QTpeak), duration of each interval from a given T-wave end to the subsequent QRS-onset (named TendQ), duration of each interval from a given T-wave peak to the subsequent QRS-onset (named TpeakQ), the ratio between each QTend interval and the subsequent TendQ interval (named QT/TQ), the ratio between each QTpico interval and the subsequent TpicoQ interval (named QTp/TpQ), the percentage from all the measures of QT/TQ for which $QT/TQ > 1$ (named $QT/TQ-r1$), the percentage from all the measures of QTp/TpQ for which $QTp/TpQ > 1$ (named $QTp/TpQ - r1$), the duration concerning the distance beween each T-wave peak and each T-wave end (named TpTe), and unevenness of the ST segment, computed as the difference between the amplitude of the QRS-onset and the amplitude of the QRS-offset. As an illustrative example of the extracted parameters pertaining the the fisrt group of features, we present in Figure 1 a seguence of QTend intervals.

Taking into consideration the extracted features of all segments within the different scenarios for durations of analysis intervals, we compute statistical functions which work as data compressors. Therefore, we define as input parameters for the machine learning models, regarding

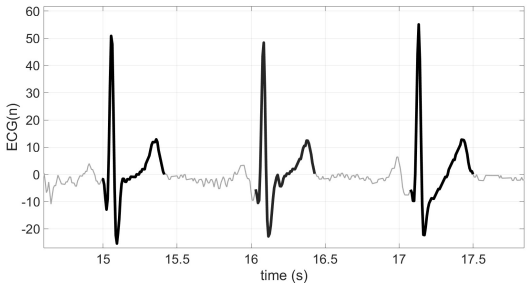

Figure 1: Examples of QTend interval durations extracted from an ECG excerpt.

the first group of characteristics, the following measures: mean of R-R intervals, standard-deviation of R-R intervals, median of the R-R intervals, median of the QTend intervals, median of the QTpeak intervals, 5th-percentile of the TendQ intervals, 5-th percentile of the TpeakQ intervals, the ratio $QT/TQ - r1$, the ratio $QTp/TpQ - r1$, the 98-th percentile of the ratios QT/TQ, the 98-th percentile of the ratios QTp/TpQ, the median of the QRS durations, the 5-th and the 98-th percentile of the QRS durations, the median of the QRS amplitudes, the 5-th and the 98-th percentiles of the QRS amplitudes, the median of the T-wave amplitudes, the 5-th and the 98-th percentiles of the T-wave amplitudes, the median of the measures of unevenness of the ST segment, the 5-th and the 98-th percentiles of the unevenness of the ST segment.

Regarding the second group of characteristics, which refers to heart rate variability parameters, we estimate power spectral density (PSD) [7] and compute the following input parameters: locations of peaks for VLF (Very Low Frequency), LF (Low Frequency) and HF (High Frequency) components, the absolute power for each component VLF, LF and HF, the power in normalizes units for LF and HF, the relative (percentual) powers for VLF, LF and HF components, and the ratio between LF power and HF power. Afterwards, we compute the following time-domain metrics as input parameters: mean normal-to-normal (NN) intervals, SDNN, mean and standard-deviations of the intantaneous heart-rate, SDANN,SDNN index, RMSSD, NN50, pNN50, and HRV index [8]. Finally, we considerate as input parameters the components corresponding to short-term (SD1) and long-term (SD2) variations from Poincaré plots [9].

Finally, regarding the fourth group of parameters (features related to theory of dynamic systems), we apply recurrence quantification analysis (RQA) [10]. Therefore, considering the time-series of beat-to-beat intervals, obtained from a given ECG time duration, we compute the following metrics as input parameters for the models: Recurrence rate (RR); Determinism (DET); Entropy (ENT); and Maximal diagonal line length (LMax) [11].

2.3. Applying Machine Learning Models

After obtaining all the set of features, we derive thirty different training and test subsets for each scenario (ECG intervals during 5, 10, 15 and 30 minutes), considering 75% of samples for training and 25% samples for test. We chose to investigate four possible thresholds concerning the definition of reduced LVEF, e.g. $< 40\%$, $< 45\%$, $< 50\%$ and $< 55\%$. In this work, we applied a binary classification, where the class 0 represents a non-reduced LVEF and the class 1, a reduced LVEF (positive class).

The computing experiments were performed using Python 3.0 and Jupyter Notebook platform ($sklearn$ library). Below, we detail the list of applied classifiers and the intervals where we preset the models to seek optimal values thorugh a grid search:

- KNN:
  - Number of neighbors: 1, 3, 5, 7, 10, 11, 13, 15, 17, 19.
- Random Forest:
  - Number of estimators: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100;
  - Maximum depth: 3, 6, 10, 15, 20.
- Multi-Layer Perceptron:
  - Activation function: hiperbolic tangent and ReLU;
  - Learning rate: constant and adaptive;
  - Hidden layer sizes: [10,10], [30,10], [20,10,5].
- Gradient Boosting:
  - Loss function: log-loss, exponential;
  - Number of estimators: [10, 30, 50, 70];
  - number of features to consider when looking for the best split: 'sqrt', 'log2'.
- Decision Tree:
  - Criterion: ['gini', 'entropy', 'log-loss'];
  - Maximum depth: [80, 100, 120].
- Extreme Gradient Boosting:
  - Number os estimators: [10, 30, 50, 70, 100];
  - Learning rate: [0.1, 0.3]
  - Maximum depth: [3, 6].

2.4. Results

The results for the applied Machine Learning models were obtained considering two possible scenarios concerning the class imbalance: (1) using all the available samples from both classes (reduced and non-reduced LVEF); (2) applying undersampling over the majority class and considering a perfect balance between the two classes for both training and testing stages.

Considering the first scenario, the threshold related to reduced LVEF which yields the best results was $LVEF < 55\%$. In table 1, we illustrate the classification results, metrics area under the curve (AUC), recall and F1-score,

ECG TD	AUC	Recall	F_1-score
5-min	0.66 $\pm$0.04	0.73 $\pm$0.09	0.66 $\pm$0.03
10-min	0.66 $\pm$0.05	0.72 $\pm$0.09	0.66 $\pm$0.04
15-min	0.66 $\pm$0.05	0.71 $\pm$0.09	0.66 $\pm$0.06
30-min	0.65 $\pm$0.04	0.72 $\pm$0.09	0.65 $\pm$0.04

Table 1: Results for reduced LVEF detection ($< 55\%$) considering unbalanced classes for Gradient Boosting

ECG TD	AUC	Recall	F_1-score
5-min	0.73 $\pm$0.09	0.71 $\pm$0.10	0.70 $\pm$0.07
10-min	0.74 $\pm$0.09	0.71 $\pm$0.10	0.69 $\pm$0.08
15-min	0.75 $\pm$0.09	0.72 $\pm$0.09	0.70 $\pm$0.07
30-min	0.74 $\pm$0.09	0.72 $\pm$0.09	0.70 $\pm$0.07

Table 2: Results for reduced LVEF detection ($< 40\%$) considering balanced classes for Gradient Boosting

for the ECG time durations (TD) corresponding to 5 minutes, 10 minutes, 15 minutes and 30 minutes for Gradient-Boosting classifier, which obtained the highest values for AUC. Considering the second scenario, the threshold related to reduced LVEF which yields the best results was $LVEF < 40\%$. In table 2, we illustrate the corresponding results using Gradient-Boosting classifier.

3. Conclusion

This work has proposed a methodology for ECG signal feature extraction which aims to detect left ventricular systolic dysfunction within patients with chronic chagasic cardiomyopathy. The computing experiments run over ECG holter signals, and a diversified group of features were obtained, considering samples of ECG intervals during 5, 10, 15 and 30 minutes. Different thresholds concerning left ventricular ejection fraction were analysed, and the overall best results were obtained considering the threshold ($< 40\%$) for characterizing a reduced LVEF. The obtained values for AUC, Recall and F1-score suggest the capability of the models to recognize patterns related to heart failure concerning the ECG signals from Chagasic patients. Future work will consider the application of feature selection approaches as well as the combination of features derived from convolutional neural networks with the features proposed here.

Acknowledgments

We are thankful for the financial support of the Ceará State Foundation for the Support of Scientific and Technological Development - Funcap, Process PS1-0186-00439.01.00/21.

References

[1] Attia ZI, Ribeiro A, Friedman P, Nunes MC, Gomes P, Ferreira A, Figueiredo B, Sabino E, Noseworthy P, Kapa S, et al. Validation of an artificial intelligence electrocardiogram based algorithm for the detection of left ventricular systolic dysfunction in subjects with chagas disease. Journal of the American College of Cardiology 2021; 77(18_Supplement_1):3254–3254.

[2] Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, Bernard ME, Rosas SL, Akfaly A, Misra A, et al. Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nature Medicine 2021;27(5):815–819.

[3] Bjerkén LV, Rønborg SN, Jensen MT, Ørting SN, Nielsen OW. Artificial intelligence enabled ecg screening for left ventricular systolic dysfunction: a systematic review. Heart Failure Reviews 2023;28(2):419–430.

[4] Cho J, Lee B, Kwon JM, Lee Y, Park H, Oh BH, Jeon KH, Park J, Kim KH. Artificial intelligence algorithm for screening heart failure with reduced ejection fraction using electrocardiography. ASAIO Journal 2021;67(3):314–321.

[5] Madeiro JP, Cortez PC, Marques JA, Seisdedos CR, Sobrinho CR. An innovative approach of qrs segmentation based on first-derivative, hilbert and wavelet transforms. Medical engineering physics 2012;34(9):1236–1246.

[6] Martínez A, Alcaraz R, Rieta JJ. Application of the phasor transform for automatic delineation of single-lead ecg fiducial points. Physiological measurement 2010;31(11):1467.

[7] Kim KK, Kim JS, Lim YG, Park KS. The effect of missing rr-interval data on heart rate variability analysis in the frequency domain. Physiological measurement 2009; 30(10):1039.

[8] Electrophysiology TFotESoCtNASoP. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. Circulation 1996;93(5):1043–1065.

[9] Brennan M, Palaniswami M, Kamen P. Do existing measures of poincare plot geometry reflect nonlinear features of heart rate variability? IEEE transactions on biomedical engineering 2001;48(11):1342–1347.

[10] Webber CL, Marwan N. Recurrence quantification analysis. Theory and Best Practices 2015;426.

[11] Araújo NS, Reyes-Garcia SZ, Brogin JA, Bueno DD, Cavalheiro EA, Scorza CA, Faber J. Chaotic and stochastic dynamics of epileptiform-like activities in sclerotic hippocampus resected from patients with pharmacoresistant epilepsy. Plos Computational Biology 2022;18(4):e1010027.

Address for correspondence:

Dr. João Paulo do Vale Madeiro.
Department of Computing Science, Federal University of Ceará, Campus do Pici, 60440-900, Fortaleza, Ceará, Brazil.
E-mail: jpaulo.vale@dc.ufc.br