Chagas disease: an analysis with temporal features extraction, permutation entropy and a stratification of heart risk by a deep learning model

Zayd Isaac Valdez[†] , Luz Alexandra Díaz[†], Miguel Vizcardo Cornejo[†], Antonio G. Ravelo-García[§‡]
[†]Escuela Profesional de Física, Universidad Nacional de San Agustín de Arequipa, Perú
[§]Institute for Technological Development and Innovation in Communications, Universidad de Las Palmas de Gran Canaria, Spain
[‡] Interactive Technologies Institute (ITI/LARSyS and ARDITI), 9020-105 Funchal, Portugal

Abstract

Chagas disease is an endemic disease that in recent decades has ceased to be a rural disease to become mainly an urban disease. In this way, it currently constitutes a public health problem since 70 million people are at risk of contagion of this potentially fatal disease. This disease has an acute and a chronic phase, where in the latter it usually has cardiac involvement that can often be silent and asymptomatic at the beginning. As a result, the establishment of early markers in this type of patients is of great interest. To achieve this, the present study proposes the analysis of RR data through permutation entropy and feature extraction.
This study analyzes three groups: 83 volunteers (Control), 102 with Chagas but without cardiac involvement (CH1) and 107 with mild to moderate incipient heart failure (CH2). The data used is from the 24-hour ECG recording, RR intervals are shown in 288 5-minute frames.
The analysis performed using permutation entropy and feature extraction shows significant differences between the 3 groups. These data, after a selection of significant segments and dimension reduction by means of PCA, were used in a densely connected neural network that has shown more than satisfactory results, obtaining 98% total accuracy and precision greater than 97% when classifying each group, thus constituting a powerful tool for risk stratification and classification of patients.

1. Introduction

Chagas disease, caused by the parasite Trypanozoma Cruzi [1], is a serious and extremely fatal disease that occurs mainly in Latin America where, according to the Pan American Health Organization [2], there are currently between 6 and 7 million people infected and tens of millions more at risk of infection. This originally endemic disease in Latin America has been spreading in recent years to other countries such as Canada, the United States and some European countries [3], due to the fact that the epidemiological pattern no longer focuses solely on vector transmission, considering now also blood transfusion and organ donation, this, consummated with the mobility of the population due to migrations, has expanded the scope of the disease [4, 5]. This situation creates a imperious necessity for cheaper and faster ways to detect the disease. The disease has an initial acute phase, commonly asymptomatic, with a duration of approximately 2 months, which, if not treated, evolves to a chronic phase in which 40% of patients develop cardiac involvement, frequently congestive heart failure, [6–9].
Previous works have used heart rate variability analysis techniques in patients with congestive heart failure [10], likewise, there are studies that demonstrate the efficacy of the application of information entropies [11] in similar problems. Thus, this work proposes the use of permutation entropy, which has already shown good results to distinguish between patients with Chagas disease from healthy people [12] . This non-linear parameter, together with the statistical characteristics that characterize each patient, were used as input data for a neural network, with the aim of creating a powerful tool that correctly classifies and distinguishes between healthy patients, patients with the disease but without involvement cardiac, and patients with cardiac complications.

2. Database

The database used was provided by the Institute of Tropical Medicine of the Central University of Venezuela [12]. It consists of 24-hour ECG recordings taken from individuals who underwent various evaluations, including Machado-Guerreiro serologic test, electrocardiogram, echocardiogram, and chest X-ray. Based on the results of these assessments, each patient was classified into one of three groups. The first is the control group which is made up of 83 healthy people. The CH1 group is made up of 102 patients with positive serology but normal in the rest. Finally, there are 107 patients with positive serology and, in

addition, evident cardiac involvement in one or more tests (with incipient heart disease, first-degree AV block, sinus bradycardia, or RBBB), who make up the CH2 group.

3. Method

In orden to obtain the QRS complexes from the ECG, the Pan-Tompkins [16] algorithm was applied. The data was then divided into 5-minute segments, covering the entire day in 288 segments. The tachogram (R-R intervals) was generated for each segment, thus representing each patient by a matrix with 288 rows containing the RR intervals. Finally, the data was processed with an adaptive filter [17].

The RR data of each patient were treated using permutation entropy (PE), a nonlinear characteristic that measures the complexity and regularity of a time series, based on the presence of patterns in it. It was defined by Bandt and Pompe as [18]:

$$H(n) = -\sum p(\pi) \log p(\pi) \tag{1}$$

The permutation entropy takes into account all the possible permutation patterns π of consecutive values and accounts for their relative appearance in the time series $p(\pi)$, it is necessary to emphasize that this parameter does take into account the order of appearance of the values. The permutation entropy was calculated for each row of RR intervals, in such a way that each patient has a vector of 288 PE values that represents it. The vectors of each patient were grouped into 3 groups according to their condition: CONTROL, CH1 and CH2, thus forming matrices for each one. An average per column was taken from these matrices to create PE circadian profiles for each group, which were analyzed using the Kruskal Wallis Test to verify that there were significant differences between them. Then, with this technique various PE parameters were tested until choosing to work with dimension 3 and a time delay of 1.

The matrices were filtered using matrices elaborated by Aproximate Entropy, in such a way that those segments corresponding to noise (a high value of ApEn) were eliminated in the PE matrices to later interpolate and extrapolate the missing values. Then the Kruskal Wallis test was performed between columns of each group to keep only significant segments, reducing the segments from 288 to 244. The Time Series Feature Extraction Library (TSFEL) was also used to extract 390 features from the RR data corresponding to all days recorded for each patient. Therefore, a vector of 390 characteristics was obtained for each patient, which were grouped according to their group and compared between them using the Kruskal-Wallis test, selecting only the features with a p value less than 0.05, in the end only 210 features remained and were added to the 244 PE values. Finally, a data augmentation process was carried out using Gaussian noise by a factor of 0.02 into the data to triple it.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

Specialized machine learning libraries and modules such as Tensorflow (and the GUI implemented on it, Keras), as well as Scikit-learn were used. In addition, basic modules such as numpy, pandas, itertools and matplotlib were used. As a first step, the previously generated data was imported along with the corresponding tags for each patient (Control, CH1, or CH2) that were coded for future use in a neural network. Then, a dimensionality reduction process was carried out using the principal component analysis (PCA), which reduced the input dimension to 150. In this way, it was possible to condense the information of each patient into shorter vectors with the advantage of preserving most of the information contained in the original vectors. Next, the data was divided into two groups for training and testing, made up of 75% and 25% of the total data respectively. Additionally, 25% of the training data was used to validate the model.

The model has 1 input layer with 20 neurons and dimension 150, which uses the activation function *sigmoid*, 3 hidden layers of 15, 9 and 7 neurons, all three with activation function *gelu* and finally an output layer with 3 neurons and *softmax* as activation function.

Likewise, an optimizer *Adam* with a learning rate of 0.001 was implemented , the number of epochs was set at 250 but the callback *EarlyStopping* was used with patience of 3 epochs to avoid overfitting. Finally, we obtain the training and validation graphs with respect to the epochs. From the results of the test set we also obtain the confusion matrix that shows the correct and incorrect classification for each group and the ROC curve for the classifier from which we can obtain the respective area under the curve.

4. Results

The loss function in Figure 1 shows a smoothly decreasing behavior both in the training set and in the validation set, reaching values quite close to 0, reaching approximately 0 in the training and up to approximately 0.2 in the validation. Since the loss function is used to indicate the error made by the neural network, its tendency to decrease means that it is improving over the epochs and this improvement is fast as it only needs 30 epochs for its convergence.

Regarding the categorical accuracy, showed in 2, both the curve corresponding to the training and the validation had a smoothly increasing trend throughout the epochs, both reaching values that tend to 1.0, the training curve fully reached this value, but the corresponding at validation it stagnated at a value slightly above 0.9. However, despite

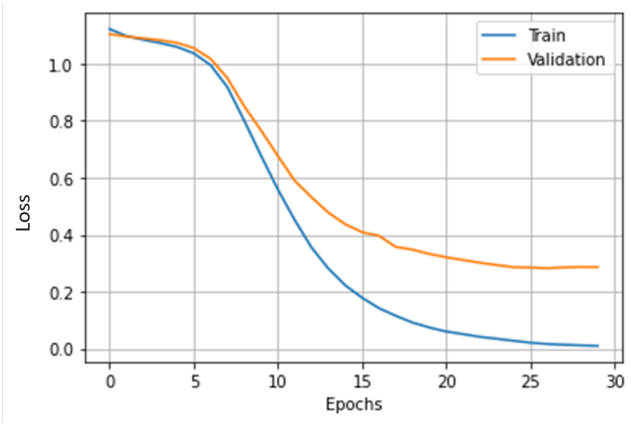

Figure 1. Evolution of the Loss function trough the epochs

this, its increasing trend and smooth curves symbolize a correct performance of the network when classifying the cases between Control, CH1 and CH2. It is possible to

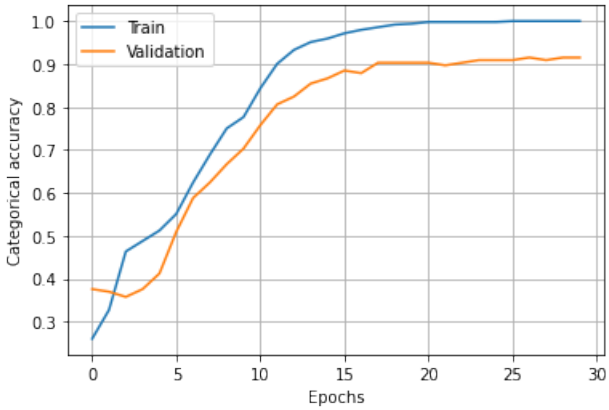

Figure 2. Evolution of the Categorical Accuracy trough the epochs

concentrate the information of the classification of the patients in a confusion matrix, which is shown in Figure 3. This evidenced an excellent classification work since all the elements outside the diagonal are minimum values, with only 5 cases being erroneously classified. This predominant density of cases on the diagonal indicates a more than satisfactory classification by the neural network. Quantitatively, from this matrix it is possible to find the total accuracy of the model and the precision, recall and F1 score for each group. Thus, there is a total accuracy of 98%. The CONTROL group obtained 99% precision, recall and F1-score. Regarding the CH1 group, the precision is 97%, the recall is 96% and the F1-score is 97%. Finally, for the CH2 group, the precision was 97%, the recall 99% and the F1-score 98%. Figure 4 shows the different ROC curves where the individual ROC curves for each group

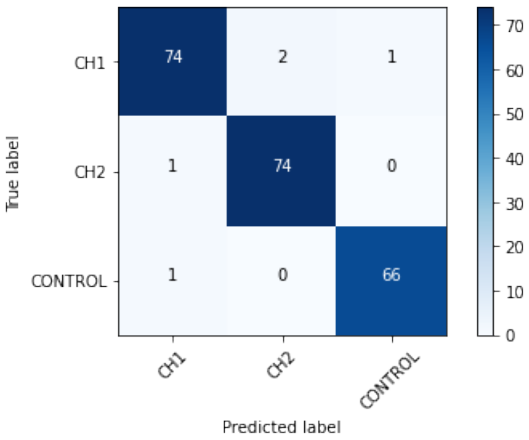

Figure 3. Confusion matrix

use a one-versus-all approach, taking that group against the rest. In a ROC graph, the rate of false positives or 1-specificity is located on the x-axis and on the y-axis the true positives or sensitivity, which are sought to be minimized and maximized respectively. Therefore, the closer a point is to the left corner of the graph, the better the performance in the classification. The observed diagonal line that is drawn starting from the origin indicates the randomness in the classification work, it is called the discrimination line, curves obtained above this indicate a satisfactory classification. From Figure 4 it is evident that all the plotted curves are quite close to the upper left corner, indicating a more than satisfactory classification.

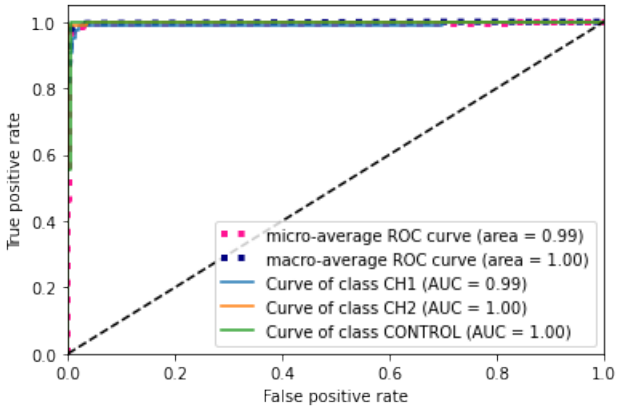

Figure 4. ROC curve

Another parameter to take into account is the area under the curves (AUC), where the closer to 1 they are, the classifier shows better performance. It is possible to see that the AUC of curves corresponding to curve of class CONTROL and CH2 have a value of 1, and the AUC of class CH1 a value of 0.99, very near to 1. About the macro average

ROC curve, it has an AUC value of 1 and the micro average a value of 0.99, all of these results are excellent and indicates a good classification work made by the network.

5. Discussion and conclusions

The analysis of RR data by permutation entropy and feature extraction to be used as input data for a densely connected neural network whose specific architecture was previously described constitutes an effective method for the diagnosis and classification of patients with Chagas disease. This method is validated by having a satisfactory perforance, reaching a total accuracy of 98%, group precision greater than 97%, and AUC greater than 0.99 for all plotted ROC curves. These high values demonstrate the great classification work carried out by the network.
Likewise, the graphs that show the evolution of loss and categorical accuracy through the epochs show a good evolution both in the training and validation stages, obtaining values as low as approximately 0 in the case of loss and as high as approximately 1 in the case of categorical accuracy in the validation stage. The only observation that could be made regarding these is their slight stagnation at slightly higher values in the validation curve, however, since they are not drastically high values, they do not mean a problem In this way, the results found in the present work validate the use of permutation entropy in RR data for the differentiation between patients with Chagas disease [12]. Likewise, feature extraction and data augmentation improve network performance by overcoming data limitations, improving the accuracy result from 91% to 98% in the classification work compared to previous works [20].

Acknowledgements

Universidad Nacional de San Agustin de Arequipa

References

[1] Briceño-León R.(2009). La enfermedad de Chagas en las Américas: una perspectiva de ecosalud. *Cadernos de saúde pública, 25*,S71-S82.

[2] Pan American Health Organization(2017). *Enfermedad de Chagas en las Américas. Hoja informativa para los trabajadores de salud*.

[3] World Health Organization (2010). First WHO report on neglected tropical diseases: working to overcome the global impact of neglected tropical diseases. In *First WHO report on neglected tropical diseases: Working to overcome the global impact of neglected tropical diseases* (p.172)

[4] Schmunis, G. A. (2007). Epidemiology of Chagas disease in non endemic countries: the role of international migration. *Memorias do Instituto Oswaldo Cruz*, 102 ´ , 75-86

[5] *Chagas disease control and prevention in Europe. Report of a WHO Informal Consultation (jointly organized by WHO headquarters and the WHO Regional Office for Europe)*. Geneva, Switzerland, 17–18 December 2009. Geneva, World

[6] Mendoza, I., Moleiro, F., Marques, J., & Britto, I. M. (2008). Arritmias y muerte súbita en la enfermedad de Chagas. *Publ. del Instituto de Medic. Tropical.*

[7] Moleiro, F. (1980). Miocardiopatía crónica chagásica: Un estudio epidemiológico utilizando métodos electrofiosiológicos de exploración clínica

[8] Hagar, J. M., & Rahimtoola, S. H. (1991). Chagas' heart disease in the United States. *New England Journal of Medicine*, 325(11), 763-768.

[9] Di Lorenzo Oliveira, C., Nunes, M. C. P., Colosimo, E. A., de Lima, E. M., Cardoso, C. S., Ferreira, A. M., ... & Ribeiro, A. L. P. (2020). Risk Score for Predicting 2-Year Mortality in Patients With Chagas Cardiomyopathy From Endemic Areas: SaMi-Trop Cohort Study. *Journal of the American Heart Association, 9*(6), e014176.

[10] Chen, W., Liu, G., Su, S., Jiang, Q., Nguyen, H. (2017). A CHF detection method based on deep learning with RR intervals. 3369-3372.

[11] Vizcardo, M., Manrique, M., Ravelo Garcia, A., Gomis, P. (2019). Application of the Entropy of Approximation for the Nonlinear Characterization in Patients with Chagas Disease. 2019 Computing in Cardiology Conference. https://doi.org/10.22489/CinC.2019.112

[12] Cornejo, D. & Rodríguez, M. & Díaz, L. & Alvarez, E. & Vizcardo, M. (2020,September). Application of Permutation Entropy in the stratification of patients with chagas disease. In *Computing in Cardiology* (pp. 1-4). IEEE.

[13] E. Haselsteiner & G. Pfurtscheller, Using time-dependent neural networks for EEG classification. *IEEE Transactions on Rehabilitation Engineering, 8*(4), pp. 457-463. doi: 10.1109/86.895948.

[14] Chatterjee, R., & Bandyopadhyay, T. (2016, January). EEG based Motor Imagery Classification using SVM and MLP. In *2016 2nd International Conference on Computational Intelligence and Networks (CINE)* (pp. 84-89). IEEE.

[15] Raad, A., Kalakech, A., & Ayache, M. (2012). Breast cancer classification using neural network approach: MLP and RBF. *networks, 7*(8), 9.

[16] Pan, J., & Tompkins, W. J. (1985). A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering, BME-32(3), 230–236.

[17] Wessel, N., Voss, A., Kurths, J., Saparin, P., Witt, A., Kleiner, H. J., & Dietz, R. (1994, September). Renormalised entropy: a new method of non-linear dynamics for the analysis of heart rate variability. In *Computers in Cardiology 1994* (pp. 137-140). IEEE.

[18] Bandt, Christoph & Pompe, Bernd. (2002). Permutation Entropy: A natural complexity measure for time series. *Physical review letters, 88*(17), 174102.

[19] Akaike, H. (1969). Fitting Autoregressive for Prediction Models. *Statist Math, 21*, 243-247.

[20] Cornejo, D. R., Ravelo-García, A., Alvarez, E., Rodríguez, M. F., Díaz, L. A., Cabrera-Caso, V., ... Cornejo, M. V. (2022, September). Deep Learning and Permutation Entropy in the Stratification of Patients with Chagas Disease. In *2022 Computing in Cardiology (CinC)* (Vol. 498, pp. 1-4). IEEE.