Optimal Artificial Neural Network for the Diagnosis of Chagas Disease Using Approximate Entropy and Data Augmentation

Maria Fernanda Rodriguez[†], Diego Rodrigo Cornejo[†], Luz Alexandra Díaz[†], Antonio Ravelo-García[§], Esteban Alvarez[‡], Victor Cabrera-Caso[†], Dante Condori-Merma[†], Miguel Vizcardo Cornejo[†]

[†]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
[‡]Escuela de Física, Universidad Central de Venezuela, Venezuela

Abstract

The use of machine learning for disease diagnosis is gaining popularity due to its ability to process data and provide accurate results, but optimazing it remains a challenge. Chagas disease is endemic in Latin America and has emerged as a health problem in more urban areas. Early and accurate diagnosis is essential to prevent cardiac complications, since an estimated 65 million people are at risk of contracting this disease. This study used a database of 292 subjects distributed into three groups: healthy volunteers (Control group), asymptomatic Chagasic patients (CH1 group) and seropositive Chagasic patients with incipient heart disease (CH2 group). A densely connected neural network was used to classify them into their respective groups. The network received as input the Approximate Entropy values of each individual, which were calculated from the 24-hour circadian profiles every 5 minutes (288 RR subsegments). Time series data augmentation algorithms were applied during the training phase to improve the classification results. This approach allowed to achieve 100% accuracy and precision, validated by the ROC curve with AUC values of 1, proving to be a robust approach for early diagnosis and prevention of heart complications in Chagas disease.

1. Introduction

Chagas disease, or American trypanosomiasis, is caused by Trypanosoma cruzi. This vector is present in 21 continental countries in the Region of the Americas. Approximattely 65 million people are at risk of contracting the infection, which causes approximately 12,000 deaths annually [1]. Additionally, in recent decades, cases have been detected in other non-endemic regions of the Americas [2]. The disease presents acutely and, if not diagnosed and treated in time, becomes a chronic disease. The most important consequence is chronic chagasic cardiomyopathy, which occurs in 20-40% of infected persons and can be potentially lethal.

Given that it is considered a neglected tropical disease [2], the use and optimization of non-invasive and low-cost diagnostic tools is paramount. In this context, machine learning has become popular as a promising technique for disease diagnosis, including Chagas disease [9–11]. Despite this, optimization of these tools remains a challenge, due to, among other reasons, the limited amount of available data. This highlights the need to implement data augmentation techniques to improve their efficiency.

Although image analysis techniques are widely used, heart rate variability (HRV) analysis can be very useful due to its prognostic significance. In particular, Approximate Entropy has been shown to be a valuable statistic for the study of congestive heart failure [5,6], one of the main clinical manifestations of Chagas disease [7], and it was also used to identify significant differences at different times of the day between groups of patients with this disease [8]. Therefore, the present work proposes the development and optimization of a densely conected neural network using the HRV based on the Approximate Entropy of a database of patients with Chagas disease, using data augmentation techniques, which, although are more common in image analysis, they are also applicable to time series.

2. Method

2.1. Database

This research employed used the ECG database of the Tropical Medicine Institute of the Universidad Central de Venezuela, which includes information on 292 individuals who underwent various tests with their respective informed consent. These tests included clinical evaluation, Gerreiro Machado-Serology test, chest X-ray, echocardiogram, electrocardiogram and Holter recording (24 hours).

The patients and volunteers were divided into three groups: the Control group, consisting of 83 healthy persons (volunteers), the CH1 group composed of 102 infected patients only with positive Machado-Gerreiro serology test, and the CH2 group composed of 107 seropositive patients with incipient heart disease, involvement of first-degree atrioventricular block, sinus bradycardia or right bundle branch block of His and were not receiving treatment or medication. ECG signals were recorded at a frequency of 500 Hz with a resolution of 12 bits.

2.2. Data preprocessing

QRS complexes were obtained from the ECG using the Pan-Tompkins [12] algorithm, then generating the 288 5-minute RR tachograms for each subject from the database. In addition, a filter used in [8] was implemented to remove noise.

Given that we are working with time series data, Approximate Entropy (ApEn) was applied to each 5-minute RR subsegment of each subject according to the definition provided by Pincus [13]. In this definition, if the time series data consists of N elements:

$$ApEn(m, r, N) = -\frac{1}{N-m} \sum_{i=1}^{N-m} \log\left(\frac{A_i}{B_i}\right) \quad (1)$$

where m is the embedding dimension, r is a threshold and A_i and B_i are the proximity measures between the embedding vectors in m and $m+1$ dimensions respectively.

After testing values of m ranging from 1 to 4, and r ranging from 10% to 50% of the standard deviation (SD), the parameters $m = 2$ and $r = 40\%$ of the SD were ultimately selected. These values were chosen due to their ability to effectively discriminate among the three groups, as well as between each group and the others.

Finally, some missing ApEn data (produced by noise filtering and the database itself) were interpolated using the Matlab function fillgaps in order to predict missing data in a series. Thus, each subject was characterized by a complete record of 288 ApEn values.

2.3. Network architecture and data augmentation

A Densely Connected Neural Network was implemented in Pyhton, using Keras and Scikit-learn, with a sequential model and dense layers. 288 ApEn values were the input layer nodes, which were previously standardized. The outputs corresponded to the three groups: Control, CH1 and CH2.

The Adam optimizer was used with a small learning rate. The loss function was categorical cross entropy, and the activation function is chosen according to the training, except in the last layer, where it was softmax. All other hyperparameters were adjusted based on the network training.

For the purpose of this study, data were divided randomly as follows: 70% of 292 subjects constituted the training set, and the other 30% the test set. Additonally, a validation set was considered and involved 30% of the training set during the network training phase.

To enhance the performance of the model, data augmentation techniques were introduced. These techniques are known for their ability to increase the generalization capacity of machine learning models by increasin the sample size (subjects) in the training set. Most of these are inspired by image recognition. Thus, scaling and jittering algorithms were selected as augmentation algorithms because of their great ability to preserve the temporal pattern of the data [14].

3. Results

An optimal 3-hidden layer architecture was found with 15, 10 and 8 neurons respectively. The activation function was sigmoid in all layers except the output layer. The Adam optimizer was used with a learning rate of 0.002, and the training was limited to 200 epochs with a batch size of 10. To mitigate overfitting, an earlystopping function was implemented, which stopped the training when the validation loss did not decrease for 5 consecutive epochs.

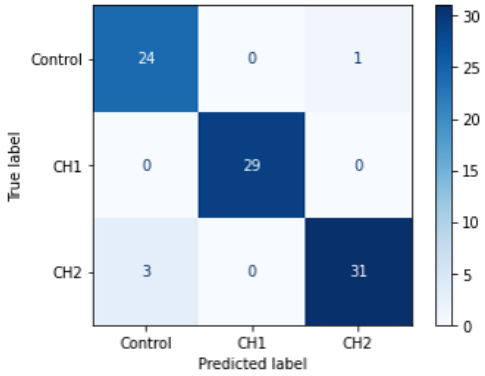

Figure 1. Confusion matrix without data augmentation

To asses the performance of the model, we initially examinated the results without implemented data augmentation. Figure 1 displays the confusion matrix, and with it, the classification results of our model were as follows: for the Control group we obtained a precision of 0.889, recall of 0.960 and F1-score of 0.923. For the CH1 group, the precision was 1.000, recall 1.000 and F1-score 1.000. And for the CH2 group the results were 0.969 for precision, 0.912 for recall and 0.939 for F1-score. The accuracy of the model reached 95.5%, and the overall weighted preci-

sion was 95.6%.

To observe the success rate, the receiver operating characteristic curve (ROC curve) was plotted. Being a multiclass classification, an extended version of the ROC curve had to be applied with the micro and macro averaging algorithm in the scikit-learn library.

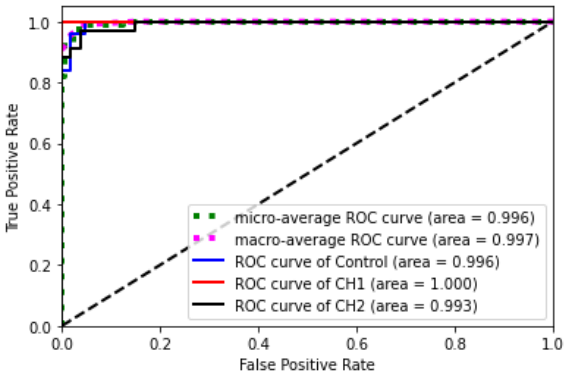

Figure 2. ROC curve without data augmentation

Thus, Figure 2 shows one curve for each group (one versus all) and two general curves for the entire classification. Since all AUC values are very close to 1, this confirms a good performance of the model, even in the absence of regularization or data augmentation techniques. However, as these are neural networks, the percentage can be improved.

We applied the aforementioned data augmentation techniques while keeping the same network architecture. Rows (patients) were added to the standardized ApEn matrix of the original training set. Consequently, the network was trained with 3 times the size of the original training set. Data augmentation was not applied to the test set to evaluate the performance of the network with original data.

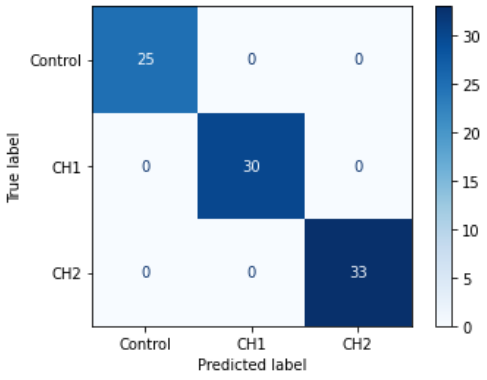

Figure 3. Confusion matrix with data augmentation

The confusion matrix, using the same data division (70% training and 30% test) is plotted in Figure 3. It is

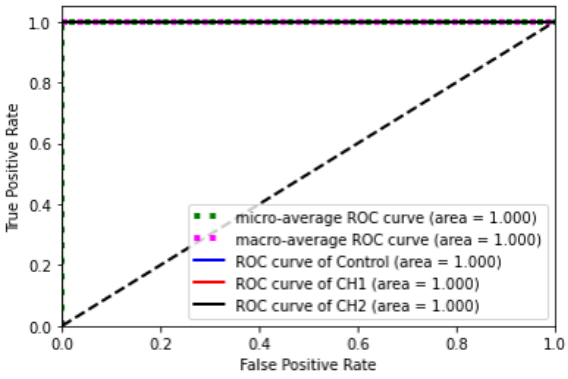

Figure 4. ROC curve with data augmentation

evident that all evaluation metrics (precision, recall and F1-score) for each group as well as the overall metrics achieved a perfect score of 100% This is supported by the multiclass ROC curve (Figure 4), whose AUC values were exactly 1.

Furthermore, taking into account the larger dataset resulting from data augmentation, we explored variations in the division of the original dataset (training and test). The overall classification metrics are summarized in the table 1, revealing that the accuracy of the model is higher than 90% even when only 30% of the original patients are used for training.

Table 1. Overall evaluation metrics, using data augmentation for different original patient divisions

Test set	**Accuracy**	**Overall precision**	**Weighted overall precision**
30% (88 subjects)	100.0%	100.0%	100.0%
40% (117 subjects)	98.3%	98.1%	98.3%
50% (146 subjects)	97.3%	97.1%	97.3%
60% (176 subjects)	94.9%	94.8%	95.0%
70% (205 subjects)	90.7%	90.5%	90.7%

4. Discussion and conclusions

Approximate Entropy proved to be a powerful statistical tool to characterize and discriminate time series data for Chagas Disease. We achieved strong performance with our neural network model even without the use of regularization techniques or data augmentation algorithms. However, in pursuit of a reliable diagnosis, we decided to implement data augmentation algorithms, despite already obtaining highly acceptable results

Data augmentation tripled the number of training samples and an excellent classification capacity was achieved: 100% accuracy and precision for the same division of the original data. Likewise, by training with a higher number of samples, it was possible to decrease the number of

original patients used for training, obtaining a classification accuracy of more than 90% even when only 30% of original patients were utilized.

Previous works have already documented high classification accuracy when employing machine learning techniques for diagnosing Chagas Disease. Cornejo et al. [15] and Rodriguez et al. [16] achieved accuracies of 91% and 98% respectively for the same database, using a deep neural network. Furthermore, 100% accuracy was already reported in the study by Hevia et al.[17] for the control versus acute infection and control versus chronic infection groups, using temporal data from four modalities in mice.

These exceptional prior studies share the same obstacle: a limited number of patients or samples. In constrast, our approach aimed to overcome this limitation by implementing training data augmentation, achieving excellent precision. Therefore, the proposed approach is presented as a highly reliable diagnostic tool that only requires ECG recordings.

Finally, it is worth highlighting that a larger, non-synthetic, and updated database would be a good way to improve the scope of this diagnostic method and contribute to timely treatment.

Acknowledgements

Universidad Nacional de San Agustin de Arequipa

References

[1] Pan American Health Organization (2021). *Enfermedad de Chagas e Inmunosupresión. Decálogo para la prevención, el diagnóstico y el tratamiento*. https://iris.paho.org/handle/10665.2/54561

[2] 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* (pp. 172-172).

[3] Marin-Neto, J., Cunha-Neto, E., Maciel, B. & Simões, M. (2007). Pathogenesis of chronic Chagas heart disease. *Circulation, 115*(9),1109-1123. doi:10.1161/CIRCULATIONAHA.106.624296.

[4] Di Lorenzo Oliveira, C., Nunes, M. C., Colosimo, E., ... & 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. doi:10.1161/JAHA.119.014176.

[5] Beckers, F., Ramaekers, D. & Auber, E. (2001). Approximate Entropy of Heart Rate Variability: Validation of Methods and Application in Heart Failure. *Cardiovascular Engineering: An International Journal, 1*, 177–182. https://doi.org/10.1023/A:1015212328405

[6] Namazi, H., Baleanu, D. & Krejcar, O. (2021). Age-Based Analysis of Heart Rate Variability (hrv) for Patients with Congestive Heart Failure. *Fractals, 29* (3), 2150135-1073. doi:10.1142/S0218348X21501358

[7] Rassi, AJr., Rassi, A. & Marin-Neto, JA. (2009). Chagas heart disease: pathophysiologic mechanisms, prognostic factors and risk stratification. *Mem Inst Oswaldo Cruz, 1*, 152-158. doi:10.1590/s0074-02762009000900021

[8] Vizcardo, M., & Ravelo, A. (2018). Use of Approximation Entropy for Stratification of Risk in Patients With Chagas Disease. In *2018 Computing in Cardiology Conference (CinC)*, 45, 1-4. doi:10.22489/CinC.2018.234

[9] Cochero, J., Pattori, L., Balsalobre, A., Ceccarelli, S. & Marti, G. (2022). A convolutional neural network to recognize Chagas disease vectors using mobile phone images. *Ecological Informatics, 68*, 101587. doi.org/10.1016/j.ecoinf.2022.101587.

[10] Sanchez-Patiño, N., Toriz-Vazquez, A., Hevia-Montiel, N. & Perez-Gonzalez, J. (2021). Convolutional Neural Networks for Chagas Parasite Detection in Histopathological Images. *International Conference of the IEEE Engineering in Medicine & Biology Society*, 2732-2735. doi:10.1109/EMBC46164.2021.9629563.

[11] Pereira, A., Mazza, L., Pinto, ... & Soares, G. (2022). Deep convolutional neural network applied to Trypanosoma cruzi detection in blood samples. *International Journal of Bio-Inspired Computation, 19*(1), 1-17. doi:10.1504/IJBIC.2022.10044882

[12] Pan, J., & Tompkins, W. J. (1985). A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering, 32*(3), 230–236. doi:10.1109/TBME.1985.325532

[13] Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences, 88*(6), 2297–2301. doi:10.1073/pnas.88.6.2297

[14] Iwana, B. & Uchida, S. (2020). Time Series Data Augmentation for Neural Networks by Time Warping with a Discriminative Teacher. *International Conference on Pattern Recognition*, 3558-3565. 10.1109/ICPR48806.2021.9412812

[15] Cornejo, D., Ravelo, A., ... & Vizcardo, M. (2022). Deep Learning and Permutation Entropy in the Stratification of Patients with Chagas Disease. In *2022 Computing in Cardiology Conference (CinC)*, 49, 1-4. doi:10.22489/CinC.2022.311

[16] Rodriguez, M., Ravelo, A., ..., & Vizcardo, M. (2022). Approximate Entropy and Densely Connected Neural Network in the Early Diagnostic of Patients with Chagas Disease. In *2022 Computing in Cardiology Conference (CinC)*, 49, 1-4. doi:10.22489/CinC.2022.313

[17] Hevia-Montiel, N., Perez-Gonzalez, J., Neme, A., & Haro, P. (2022). Machine Learning-Based Feature Selection and Classification for the Experimental Diagnosis of Trypanosoma cruzi. *Electronics, 11*(5), 785. doi.org/10.3390/electronics11050785

Correspondence:
Miguel Vizcardo Cornejo, Av. Independencia s/n Ciudad Universitaria, Edificio de Fìsica, Laboratorio Nro. 305, Arequipa 04001, Perú. Email; *mvizcardoc@unsa.edu.pe*