Early ECG Warning for Chagas Patients: Implementation of TinyML for Low-Resource Areas in Peru

Diego A. Taquiri, Student, Erick A. Valdivia, Student, Ana B. Mantilla, Student and Armando A. Flórez, Student,

[1]

***Abstract*— Cardiovascular diseases (CVDs) are the leading cause of death globally, claiming approximately 17.9 million lives annually. These disorders, including coronary heart diseases, cerebrovascular diseases, and rheumatic heart diseases, predominantly affect low- and middle-income countries, where over 75% of deaths occur. In the Americas, CVDs are responsible for an estimated 2 million deaths per year. In Peru, cardiovascular diseases are the second leading cause of death. This study aims to develop an early warning system based on electrocardiography (ECG) for detecting high-risk Chagas disease patients, with potential applications to other cardiovascular conditions. We employ advanced signal processing and machine learning techniques, specifically wavelet transform feature extraction and Tiny Machine Learning (TinyML), to efficiently and cost-effectively identify specific patterns in the ECG signals of Chagas patients. This system aims to improve timely clinical intervention and reduce mortality rates associated with Chagas disease in low-resource settings.**

***Index Terms*—Cardiovascular diseases, Chagas disease, ECG, TinyML, signal processing, wavelet transform, early warning system.**

I. Introduction

Cardiovascular diseases (CVDs) are the primary cause of death worldwide, accounting for approximately 17.9 million deaths annually [1]. This group of disorders, including coronary heart diseases, cerebrovascular diseases, and rheumatic heart diseases, significantly impacts low- and middle-income countries, where more than 75% of CVD-related deaths occur [1]. In the Americas, the prevention and treatment of these diseases are crucial as they cause an estimated 2 million deaths annually [2]. In Peru, EsSalud highlights that cardiovascular diseases are the second leading cause of death [3]. A study by the INEI revealed that over 40% of individuals over the age of 15 are at very high cardiovascular risk, with 60% of these cases being women aged 60 and above [3, 4]. The vice president of the Peruvian Society of Cardiology, Dr. José Ercilla, emphasizes that these diseases pose a future challenge, especially with the projected significant increase in the population over 50 years old by 2050 [4].

Chagas disease, a parasitic, systemic, chronic condition transmitted by vectors and caused by the protozoan Trypanosoma cruzi, is a major health concern in Latin America. Approximately 30% of infected patients develop Chagas cardiomyopathy, the leading cause of non-ischemic cardiomyopathy in the region, characterized by diffuse myocarditis with focal fibrosis, primarily in the apex and basal segments of the posterior and inferior wall, resulting in a highly arrhythmogenic disease [5, 6]. The Pan American Health Organization estimates that around 8 million people are infected in these regions, causing approximately 12,000 deaths annually [6]. The transmission primarily occurs vectorially but can also occur through blood transfusions, congenital transmission, organ transplants, or orally [7]. Although mortality has significantly decreased, the disease can cause irreversible and chronic consequences in the heart, digestive system, and nervous system [7]. In Peru, the endemic area for Chagas disease is mainly in the southwestern Pacific region, between 13 and 19 degrees south latitude and altitudes of 10 to 3,000 meters above sea level [9, 10]. Poverty in rural areas contributes to the presence of vectors in precarious housing, increasing the risk of infection. Additionally, the migration of rural populations to urban centers has increased the relevance of this disease in areas where the presence of the vector has not been detected, highlighting possible non-vector transmission mechanisms, such as in Lima [9].

To prevent premature deaths from cardiovascular diseases, it is essential to identify high-risk individuals and ensure they receive appropriate treatment, which requires universal access to essential medications and technologies in all primary care centers [1]. Cardiologists emphasize the importance of reducing disability and premature death through timely treatments, reducing waiting lists, and adopting less invasive procedures for faster recovery [4]. Additionally, effective control of Chagas disease requires coordinated efforts among involved entities, drawing from successful strategies implemented in other Latin American countries [9].

[1]Manuscript submitted July 02, 2024. This work was supported in part by Universidad Peruana Cayetano Heredia.

(Corresponding author: Lewis de la Cruz Rodríguez).

Ana B. Mantilla is with the Department of Biomedical Engineering, Universidad Peruana Cayetano Heredia, Lima, Peru (e-mail: ana.mantilla@upch.pe).

Armando A. Flórez is with the Department of Biomedical Engineering, Universidad Peruana Cayetano Heredia, Lima, Peru (e-mail: armando.florez@upch.pe).

Diego A. Taquiri is with the Department of Biology, Universidad Peruana Cayetano Heredia, Lima, Peru (e-mail: diego.taquiri@upch.pe).

Erick A. Valdivia is with the Department of Biomedical Engineering, Universidad Peruana Cayetano Heredia, Lima, Peru (e-mail: erick.valdivia@upch.pe).

The authors contributed equally to this work.

II. Materials

The primary resource for this project is the Sami-Trop dataset, which includes ECG signals and mortality data from 2000 patients with Chagas cardiomyopathy. Additionally, we will record signals from healthy patients using Arduino Nano BLE 33. We also utilize methodologies and approaches described in key literature, such as the articles "AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial" [14] and "Deep neural network-estimated electrocardiographic age as a mortality predictor" [15]. These references provide a robust methodological framework for mortality prediction. For ECG signal processing, we employ wavelet filtering techniques using software tools like MATLAB and Python with libraries such as PyWavelets and SciPy. Wavelet transforms, such as the discrete and continuous wavelet transform, are used to filter and denoise the signals, preserving important features like the QRS complexes.

In terms of hardware and development platforms, we use Edge Impulse for the implementation of machine learning models on embedded devices, such as the Arduino Nano, which integrates with ECG sensors for new signal acquisition. These devices allow real-time analysis and can be used to validate the models developed with the Sami-Trop dataset.

III. Methods

A. Signal Acquisition Protocol

The Sami-Trop dataset includes ECG data and mortality information from 2000 patients diagnosed with Chagas cardiomyopathy. These patients were selected following specific clinical criteria for Chagas disease, based on serological confirmation of Trypanosoma cruzi infection and the presence of clinical signs of cardiomyopathy, ensuring that all patients in the study meet the necessary diagnostic criteria for obtaining representative and relevant data for analysis.

The ECGs were recorded using standard 12-lead ECG machines, common in clinical settings. Each ECG recording has a typical duration of 10 seconds per lead, and the digital signals were sampled at a frequency of 400 Hz, ensuring detailed capture of the heart's electrical activity. The recordings were performed in a clinical or hospital setting with patients at rest in the supine position to reduce variability due to physical activity or posture [13].

For recording signals from healthy patients, we use Arduino Nano BLE 33 with the AD8232 ECG sensor. Given that right bundle branch block is the most frequent alteration when evaluating the presence of Chagas disease, this symptom will be observed. It is characterized by a widening of the QRS complex and a change in the T-wave morphology in the right-sided leads (D3 and AVR). These leads will be analyzed by Arduino Nano BLE 33, and the results will be recorded for future comparisons.

The ECG data were stored in digital format, in files containing voltage values for each lead over time. Preprocessing of these data includes baseline correction to eliminate drifts and filtering to remove noise. Additionally, the signals were normalized to ensure uniformity in subsequent analyses. These steps are crucial to improve signal quality and facilitate precise analysis.

B. Determination of Chagas Disease Characteristics

Chagas disease can cause various types of ECG alterations, evolving from a normal ECG to defined but nonspecific alterations of the disease, highlighting ventricular extrasystoles and ventricular repolarization abnormalities [18].

Marques et al.'s study [19] reports ECG abnormalities in populations with both vector-borne and orally transmitted trypanosomiasis. Among the changes found in Chagas patients are initially ST segment and T wave abnormalities (37.86%), QT segment prolongation (2.91%), and right (1.94%) and left (2.91%) bundle branch blocks. Additionally, arrhythmias were reported in 32% of cases, diagnosed through Holter monitoring. Features for subsequent signal filtering can be obtained from these characteristics. Those dependent on signal morphology need to determine the morphology of the P, T waves and QRS complex in terms of amplitude and duration. Although arrhythmias are common in Chagas patients, parameters derived from heart rate variability (TO, TS, SDNN, RMSSD, NN50, and pNN50) [20] would allow a deeper analysis of the arrhythmias that occur. These are the result of continuous monitoring for the detection of heart rate accelerations or decelerations after ectopic beats and are specific to studies with longer samples than we can access.

C. Signal pre-processing

The signals obtained from the dataset tensors were filtered using a Buttersworth bandpass filter of order 4 with cutoff frequencies of 1 Hz and 50 Hz to effectively remove power line noise and to remove baseline noise before introducing the data in the machine learning model platform. The data was subjected to further filtering with the aid of the Edge Impulse platform, where a Wavelet transform daubechies 4 (db4) with 4 decomposition levels was used to further filter the ECG signals from the DI, DII and DIII leads chosen for training and testing.

D. Feature Extraction

Feature extraction is performed through temporal search windows in a multi-step process:

- R-R intervals: The R-peak is identified by detecting the local maxima in the ECG wave filtered by the algorithm described by Masomenos et al [22]. First and second derivatives were taken in order to pinpoint the zones of greatest change in the signal, corresponding to the R peaks. R-R peak distances would be obtained from the result afterwards.
- Skewness: Skewness measures the asymmetry of the ECG signal distribution. It is calculated by evaluating the third standardized moment of the signal, which provides insights into the data distribution's tilt.
- Kurtosis: Kurtosis quantifies the ECG signal's

peakiness or flatness compared to a normal distribution. This is determined by the fourth standardized moment, indicating the presence of outliers or the sharpness of the signal peaks.
- Entropy: Entropy assesses the complexity and irregularity of the ECG signal. It is computed using the Shannon entropy formula, which evaluates the signal's unpredictability and information content.
- RMSSD (Root Mean Square of Successive Differences): RMSSD is a time-domain measure of heart rate variability. It is calculated by taking the square root of the mean of the squares of the successive differences between adjacent R-R intervals, reflecting the variability in heart rate.

E. ML model training

For this study, we selected the XGBoost machine learning algorithm to build a regression model aimed at predicting the time to death ("timey") of each patient. XGBoost, or Extreme Gradient Boosting, is an efficient implementation of gradient-boosted decision trees designed for speed and performance. It was chosen for its ability to handle various data types and its robustness against overfitting. The XGBoost regression model was trained on the prepared dataset, with the extracted features serving as input variables and "timey" (time to death) as the target variable. The data used for training the model consisted of three leads: DI, DII, or DIII.

To evaluate the performance of the regression model, we employed the following metrics:

- R-squared (R^2): This metric indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. An R^2 value of 1 implies perfect prediction, while 0 indicates no predictive power. It helps assess the model's goodness-of-fit.
- Mean Squared Error (MSE): Measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. Lower MSE values indicate better model performance, as they reflect smaller prediction errors.
- Mean Absolute Error (MAE): Represents the average absolute difference between predicted and actual values. MAE is less sensitive to outliers compared to MSE and provides a straightforward interpretation of prediction accuracy in the original units of the target variable.
- Loss: the loss function typically refers to the objective function being minimized during training. It measures how well the model's predictions match the actual data, guiding the optimization process.

This dataset was divided into 80% for the training set and 20% for the test set. This division allows the model to be trained with the majority of the available data while reserving a portion to validate and test the model independently, to evaluate its generalization capability [16].

For the model training, we will use the Edge Impulse platform [17], a tool designed for the development of embedded systems that facilitates the training of artificial intelligence models and their deployment on hardware, such as the Arduino Nano. Edge Impulse provides a comprehensive environment for data collection, feature extraction, and model training. For deployment, we used EdgeImpulse with the custom Arduino library option. We used the TensorFlow Lite optimization module, Quantized, for INT8.

V. Results and discussion

A. Exploratory analysis of the data set.

The dataset metadata contains the following information: "exam_id", which is the id used for internal uses; "age", which indicates the patient's age in years at the time of the examination; "is_male", which is true if the patient is male and false if the patient is female; "normal_ecg", which is true if the patient has a normal ECG; “death,” which is true if the patient fails during the follow-up time; and "timey", which indicates the time until the patient dies if it fails, and if not, the follow-up time.

exam_id	age	is_male	normal_ecg	death	timey
278507	68	True	False	True	1.52486019352465
253942	69	True	False	True	0.350828341626142
278635	87	True	False	True	1.18508156344579
222356	78	False	False	True	0.0773479808309604
248627	87	False	False	True	0.298342211776562
294669	67	True	False	False	2.11601976130413
291318	65	True	False	False	3.0773446659175
247007	67	False	False	False	2.37845041055203
181629	34	False	False	False	2.67679262232859
250434	52	False	False	False	2.11878218919095
253698	67	False	False	False	2.07182091511501

Fig. 1. Metadata of the dataset.

We plotted the ECG signals to examine the wave's morphological characteristics. A pattern of abnormal ECGs was observed in Chagas patients who died during the trial, compared to those who survived (Figure 2). Notably, 104 patients died, while there were 1528 ECGs from surviving patients, evidencing an unbalanced dataset.

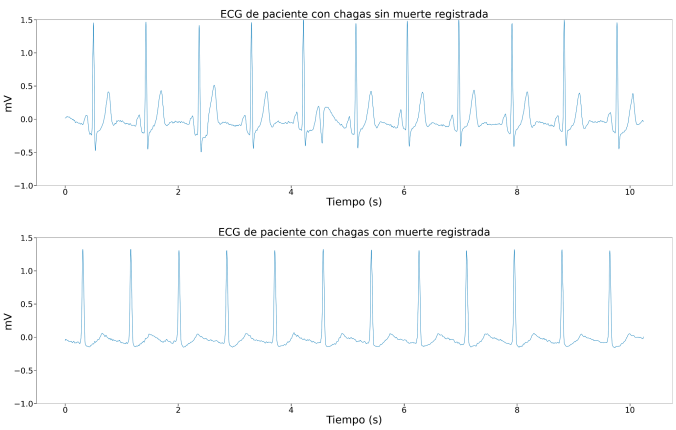


Fig. 2. Comparison of ECG signals from a Chagas patient who did not die (top) vs a Chagas patient who died (bottom).

D.Performance of the models

Our objective was to develop a low-cost implementation of this algorithm suitable for deployment on an Arduino platform, thereby making it accessible in resource-limited settings. To achieve this, we simplified several components of the original model [12]. We transitioned from a large convolutional neural network to a more compact machine learning model, XGBoost, enabling the Arduino Nano to handle inference tasks within its RAM constraints. Additionally, we reduced the number of ECG leads from 12 to one, as the use of multiple leads is expensive and typically only feasible in high-resource hospital settings.

We also replaced the Cox loss function with a simpler linear regression loss. To compensate for these simplifications, we integrated a feature extraction step, incorporating features such as skewness and entropy, to enhance the model's ability to extract and utilize signal information for making predictions. Despite these reductions, the model demonstrated learning capability and produced satisfactory metric scores, such as R-squared, indicating its potential viability in low-resource, portable, and wearable applications.

TABLE I
PERFORMANCE RESULTS OF THREE XGBOOST MODELS WITH THREE DIFFERENT ECG LEADS OF INPUT.

Lead	R^2	MSE	MAE	Loss
DI	0.31	0.57	0.65	0.57
DII	0.48	0.42	0.55	0.42
DIII	**0.61**	**0.38**	**0.51**	**0.38**

The performance evaluation of the XGBoost model for prognosticating time until death in Chagas disease patients reveals a nuanced picture of its effectiveness and areas for improvement (Table I). With a Mean Squared Error (MSE) of 0.38 and a Mean Absolute Error (MAE) of 0.51, the model demonstrates a reasonably accurate prediction capability, though there is potential for enhancement. The MSE value indicates that while the model achieves a relatively low average squared error in predictions, it still experiences some discrepancies in estimating the time to death. The MAE further corroborates this observation by showing an average absolute difference of 0.51 units between predicted and actual outcomes, which reflects the model's overall predictive accuracy but also highlights that occasional larger errors occur. These metrics suggest that while the model is performing reasonably well, there is a significant opportunity for refining its predictions, particularly in minimizing larger errors that are penalized more by MSE.

The Explained Variance Score of 0.61 indicates that the model is able to explain 61% of the variance in the time to death, leaving 39% of the variance unexplained. This moderate level of explained variance suggests that while the model captures a substantial portion of the variability in patient outcomes, there are underlying factors or complexities that remain beyond its current scope. This result implies that the model's ability to account for variations in time to death is solid but not exhaustive, and there are likely additional variables or interactions that could improve predictive accuracy if incorporated into the model.

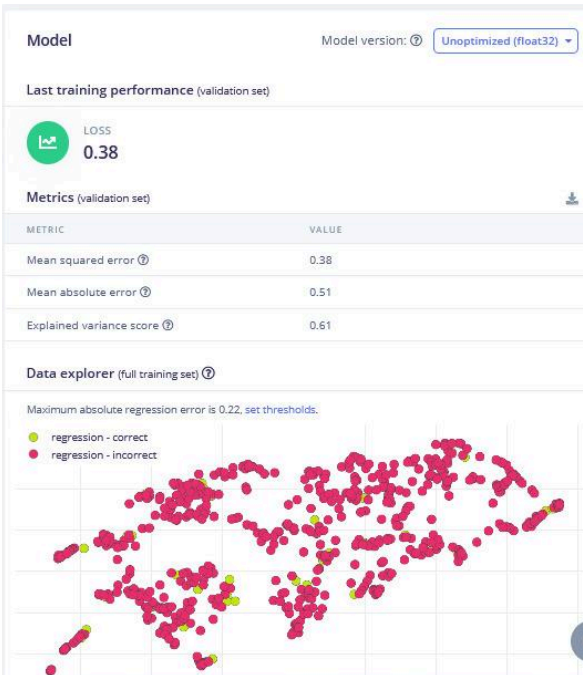


Fig. 3. Evaluation metrics of the machine learning model trained on the dataset.

To test the model, we deployed it on an Arduino BLE 33 and acquired an ECG signal from a healthy subject (Fig. 4). We then ran the inference, and the model predicted a value of 3.2. This result is similar to the ECGs in the dataset of patients who survived.

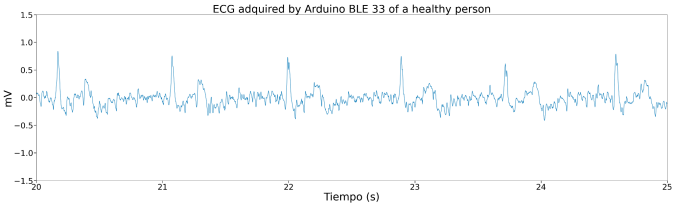


Fig. 4. ECG signal acquired by Arduino BLE 33 of a healthy subject.

The test shows the limitations of this study, which are centered on the skewed data distribution used in the model. While the dataset in reference [12] was trained on raw ECG

data from 450,000 ECGs, our model utilized a smaller mortality dataset of 1,600 samples. Additionally, to predict mortality, only a small subset of samples from patients who died during the follow-up period and had corresponding records could be used. The follow-up period for surviving patients in the dataset was limited to a maximum of three years, making outcomes for patients with longer-term survival less reliable, and unreliable for patients without the disease, for which a classification model or survival data for healthy patients would be needed for a reliable prediction.

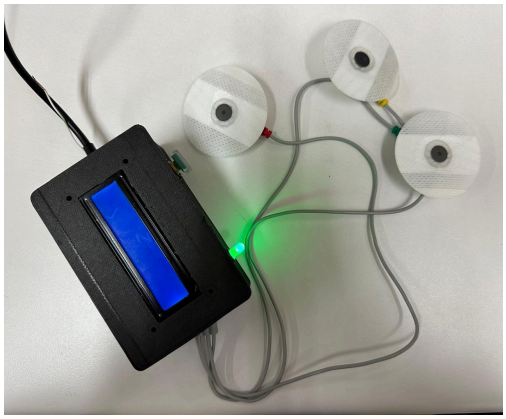


Fig 5. Hardware prototype to visualize the result

VII. Conclusion

These findings underscore that the XGBoost model is a good starting point for predicting time to death in Chagas disease patients, offering a foundation for future enhancements. To improve performance, future work could focus on several strategies: refining the hyperparameters of the XGBoost algorithm for better optimization, exploring additional features derived from the ECG signals or other relevant clinical data, and expanding the dataset to include a wider range of patient conditions and outcomes. Additionally, exploring alternative machine learning models or advanced techniques could further enhance the model's effectiveness.

References

[1] "Cardiovascular Diseases," Who.int. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases.

[2] "Heart diseases remain the leading cause of death in the Americas," Paho.org. [Online]. Available: https://www.paho.org/en/news/29-9-2021-heart-diseases-remain-leading-cause-death-americas.

[3] "Heart diseases are the second leading cause of death in Peru," Essalud, March 12, 2022. [Online]. Available: http://noticias.essalud.gob.pe/?inno-noticia=heart-diseases-are-the-second-leading-cause-of-death-in-peru.

[4] INEI, "National Institute of Statistics and Informatics," Gob.pe. [Online]. Available: https://m.inei.gob.pe/prensa/noticias/39.9-of-peruvians-15-and-older-have-at-least-one-comorbidity-12903/.

[5] N. G. Echavarría, L. E. Echeverría, M. Stewart, C. Gallego, and C. Saldarriaga, "Chagas disease: Chronic Chagas cardiomyopathy," Curr. Probl. Cardiol., vol. 46, no. 3, p. 100507, 2021.

[6] "Chagas disease (American trypanosomiasis)," Who.int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis).

[7] "Chagas disease," Paho.org. [Online]. Available: https://www.paho.org/en/topics/chagas-disease.

[8] M. Martinez Palomino, "Chronic Chagasic cardiomyopathy," 2022.

[9] C. N. Velarde, P. Huapaya, Y. E. Blanco, and S. V. Chirinos, "Chagas disease," Gob.pe, 2001. [Online]. Available: https://bvs.minsa.gob.pe/local/OGEI/796_MS-OGE109.pdf.

[10] "Health situation room for Chagas disease. SE 32," Gob.pe, 2022. [Online]. Available: https://www.dge.gob.pe/portal/docs/tools/teleconferencia/2022/SE332022/03.pdf.

[11] "Chagas disease (American trypanosomiasis)," Who.int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis).

[12] C.-S. Lin et al., "AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial," Nat. Med., pp. 1–10, 2024.

[13] A. L. P. Ribeiro, A. H. Ribeiro, G. M. M. Paixao, E. M. Lima, M. Horta Ribeiro, M. M. Pinto Filho, P. R. Gomes, D. M. Oliveira, W. Meira Jr., T. B. Schon, and E. C. Sabino, "Sami-Trop: 12-lead ECG traces with age and mortality annotations (1.0.0) [Data set]," Zenodo, 2021. https://doi.org/10.5281/zenodo.4905618

[14] C.-S. Lin, W.-T. Liu, D.-J. Tsai, et al. "AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial," Nat Med (2024). https://doi.org/10.1038/s41591-024-02961-4 (free access link: https://rdcu.be/dIkU2)

[15] E. M. Lima, A. H. Ribeiro, G. M. M. Paixão, et al. "Deep neural network-estimated electrocardiographic age as a mortality predictor," Nat Commun 12, 5117 (2021). https://doi.org/10.1038/s41467-021-25351-7

[16] H. V. Denysyuk, R. J. Pinto, P. M. Silva, R. P. Duarte, F. A. Marinho, L. Pimenta, and I. M. Pires, "Algorithms for automated diagnosis of cardiovascular diseases based on ECG data: A comprehensive systematic review," Heliyon, vol. 9, no. 2, 2023. https://doi.org/10.1016/j.heliyon.2023.e13601

[17] "Adding Custom Learning Blocks," Edge Impulse Documentation, Edge Impulse, [Online]. Available: https://edge-impulse.gitbook.io/docs/edge-impulse-studio/learning-blocks/adding-custom-learning-blocks.

[18] B. Oliveira and A. Luiz, "Electrocardiogram in Chagas disease," Revista da Sociedade Brasileira de Medicina Tropical, vol. 51, no. 5, pp. 570–577, Oct. 2018, doi: https://doi.org/10.1590/0037-8682-0184-2018.

[19] J. Marques, I. Mendoza, B. Noya, H. Acquatella, I. Palacios, and M. Marques-Mejias, "ECG Manifestations of the Biggest Outbreak of Chagas Disease due to Oral Infection in Latin-America," Arquivos Brasileiros de Cardiologia, Jan. 2013, doi: https://doi.org/10.5935/abc.20130144.

[20] A. C. Alberto, G. A. Limeira, R. C. Pedrosa, V. Zarzoso, and J. Nadal, "ECG-Based Predictors of Sudden Cardiac Death in Chagas' Disease," in Computing in Cardiology, Rennes, France, Sep. 2017, doi: 10.22489/cinc.2017.087-324

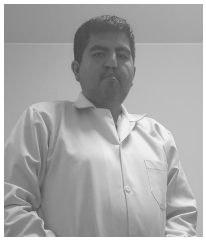
[21] M. Alfaouri and K. Daqrouq, "ECG signal denoising by wavelet transform thresholding," American Journal of Applied Sciences, vol. 5, no. 3, pp. 276-281, 2008. doi:10.3844/ajassp.2008.276.281.

[22] E. B. Mazomenos, T. Chen, A. Acharyya, A. Bhattacharya, J. Rosengarten and K. Maharatna, "A Time-Domain Morphology and Gradient based algorithm for ECG feature extraction," 2012 IEEE International Conference on Industrial Technology, Athens, Greece, 2012, pp. 117-122, doi: 10.1109/ICIT.2012.6209924.

Author Biography

Diego A. Taquiri (Student) was born in Lima, Peru, on October 10, 1998. He is currently pursuing a degree in Biology in Universidad Peruana Cayetano Heredia (UPCH), Lima, Peru.

He has experience in the area of bioinformatics, having been part of the Bioinformatics and Molecular Biology Laboratory, in UPCH, as a research assistant. He is interested in genomics and machine learning applications in the area of healthcare.



Erick A. Valdivia (Student) was born in Arequipa, Peru, on December 21, 1989. Graduated from Universidad Católica de Santa Maria in 2018 with a degree in Medicine. Currently pursuing a degree in Biomedical Engineering at Universidad Peruana Cayetano Heredia, with a strong interest in Clinical Engineering within tertiary healthcare environments.

Ana B. Mantilla Mantilla (Student) was born in Trujillo, Peru, on July 3, 2002. She is currently pursuing a degree in biomedical engineering at Pontificia Universidad Católica del Perú (PUCP) and Universidad Peruana Cayetano Heredia (UPCH), Lima, Peru. She is in her fifth year of study.

She has experience as a teaching assistant in the Academic Department of Engineering in the Section of Electricity and Electronics at PUCP and is currently doing pre-professional internships in the Laboratory of Biomechanical Engineering and Applied Robotics. She is interested in expanding her knowledge in the field of electronics and its applications in biomedicine. However, she remains open to any enriching experiences that allow her to navigate the world of biomedical engineering.

Armando A. Flórez (Student) was born in Lima, Peru, on April 1, 2002. He is currently pursuing a degree in biomedical engineering at Pontificia Universidad Católica del Perú (PUCP) and Universidad Peruana Cayetano Heredia (UPCH), Lima, Peru.

He has experience in the area of biomaterials, having been part of the Materials Characterization Center (CAM) and the Polymers and Nanomaterials Laboratory, in PUCP, as a volunteer and research assistant, respectively. He is interested in this area, alongside tissue engineering and their possible applications in the medical field.