Deep Learning for Early Chagas Disease Diagnosis: A Comparative Analysis of 12-Lead ECG and Derived VCG

Alejandro Pascual-Mellado[1], Vicent Torres-Sastre[1], Cristina Albert[1], Alejandro Pérez-González[1], Raúl Alós[1], Elisa Ramírez[1], Francisco Castells[1], José Millet[1]

[1] EP Analytics Lab - ITACA Institute, Universitat Politècnica de València

Abstract

Chagas disease (ChD) is a chronic parasitic condition that can lead to severe cardiac complications. The use of ECG analysis has emerged as a promising tool for early, non-invasive detection. This work, developed by the EPBandoleroLab team for the PhysioNet Challenge 2025, presents a deep learning approach for ChD classification using the CODE-15, SaMi-Trop and PTB-XL databases. Our methodology explores the effectiveness of different signal representations, comparing the standard 12-lead ECG with the derived Vectorcardiogram (VCG). Furthermore, we address the significant class imbalance through a controlled sampling strategy. Our findings indicate that the model performs best when trained on the full 12-lead ECG representation with a moderately imbalanced dataset. This configuration achieved a mean Challenge Score of 0.174 in the official phase, positioning us among the top 30 teams selected for publication.

1. Introduction

Chagas disease (ChD), caused by the protozoan *Trypanosoma cruzi*, is a neglected tropical disease affecting millions in Latin America. In its chronic phase, it can lead to severe cardiac complications. Transmission occurs primarily via triatomine insects (”kissing bugs”) and other routes [1].

Although the acute phase is often asymptomatic, a significant proportion of patients develop chronic ChD cardiomyopathy, characterized by ventricular dysfunction, thromboembolism, arrhythmias and dysautonomia [2]. Early diagnosis is crucial but is often hindered by limited access to serological testing in rural or under-resourced settings.

ECG, as a low-cost and widely available diagnostic tool, holds promise for detecting early signs of chronic ChD cardiomyopathy. Certain alterations in ECG, such as right bundle branch block, premature ventricular beats, ST-T changes, abnormal Q waves, various degrees of AV block, sick sinus syndrome and low QRS voltage, may suggest ChD even in asymptomatic individuals [2, 3].

The 2025 PhysioNet Challenge [4–6] focuses precisely on this problem: detecting ChD disease using standard 12-lead ECG recordings via machine learning and deep learning methods.

Recent studies highlight the strong potential of AI in ECG-based disease detection, with deep neural networks surpassing medical residents and accurately classifying multiple arrhythmias [7].

2. Materials

The dataset provided for the PhysioNet Challenge 2025 comprises three distinct ECG databases, each offering unique characteristics relevant to the ChD detection task:

- **CODE-15%:** Subset of the larger CODE cohort, with over 300,000 12-lead ECGs (7.3–10.2 s, 400 Hz) collected in Brazil between 2010–2016, split into 18 partitions. Chagas labels are self-reported (weak) with unknown accuracy and low prevalence, introducing real-world noise and variability [8].

- **SaMi-Trop:** Contains 1,631 12-lead ECGs recorded between 2011–2012 from patients in Brazilian endemic regions. Chagas diagnosis is serologically confirmed, making this the only strongly labeled positive dataset [9]

- **PTB-XL:** Includes 21,799 12-lead ECGs from Germany (10 s, 500 Hz). Due to its European origin, it is assumed to be Chagas-negative, providing a clean, low-noise negative class [10].

To create a robust **held-out test set** for evaluating generalization, we randomly held out two of the eighteen CODE-15% partitions. This held-out set was reserved exclusively for final testing and was not used during training or hyperparameter tuning.

The remaining data, including the other 16 CODE-15% partitions, SaMi-Trop, and PTB-XL constituted our development set. From this set, various experimental subsets were generated using a data sampling strategy, which is further detailed in the Methods section. Each of these subsets was then split at the patient level into training (70%) and validation (30%) partitions.

The final test sets, used for the official evaluation and ranking of competitors, remain hidden and are exclusively accessed by the event organizers.

3. Methods

3.1. Data Selection and Preprocessing

Due to a class imbalance favoring negative cases, a data selection strategy was implemented. While all positive records from the three databases were included, negative samples were filtered based on demographic characteristics (age and gender). This allowed us to control the ratio between negative and positive records using a balancing parameter R to experimentally investigate whether a controlled imbalance is beneficial for the model's generalization.

The signals were preprocessed by resampling to 400 Hz, followed by a two-stage filtering process (median and wavelet) to remove noise. Each lead was then standardized using Z-score normalization. To create fixed-length inputs, signals were segmented into 1024-sample windows. Patient-level data splits were enforced to prevent data leakage, and the final inference for a record is the average of its segment probabilities.

3.2. Signal Representation: ECG vs. VCG

To determine the most effective input representation, an experimental comparison was conducted between two modalities. The first is the standard 12-lead ECG, which provides detailed temporal information of the cardiac vector voltage from multiple anatomical perspectives.

As an alternative, the VCG was evaluated, a three-dimensional representation of the heart's electrical activity mathematically derived from the 12 leads using the inverse Dower transformation (DT) matrix. The VCG projects the information onto three orthogonal axes (X, Y, Z), offering a spatial view of the cardiac vector.

The underlying hypothesis is that the VCG, by eliminating the inherent redundancy among ECG leads, could allow the model to learn global diagnostic features more efficiently [11]. However, it is important to acknowledge that the DT is an estimate of the true cardiac vector, not a direct measurement, which entails the risk of introducing slight signal distortion. This experiment therefore investigates whether the benefit of a compact and non-redundant representation outweighs the potential loss of fidelity, building on previous studies indicating that key diagnostic information is largely preserved.

3.3. Hybrid Model Architecture

A hybrid architecture was designed to use a CNN for feature extraction and a Transformer for contextual modeling. The Transformer's output is fed into a Multilayer Perceptron (MLP) for the final classification. Figure 1 shows a diagram of this architecture.

The stack of 1D convolutional layers processes the input signal. This part acts as a local feature extractor, learning to identify morphological patterns in the signal. Through layers of convolution, normalization, and pooling, the CNN transforms the signal into a shorter, denser sequence of feature vectors.

The feature sequence generated by the CNN is fed into a Transformer encoder. This component, through its multi-head self-attention mechanism, models long-term temporal dependencies in the signal. A special classification token (`[CLS]`) is prepended to the sequence to aggregate the contextual information of the entire segment into a single representation vector.

The feature vector corresponding to the `[CLS]` token is then passed to a final MLP. After an initial normalization, the MLP projects the input through dense layers with Dropout to produce a single logit, which is mapped to a probability using a sigmoid function.

3.4. Training and Evaluation Strategy

For the experimentation, multiple development sets were generated by varying the balance ratio and the signal representation (ECG vs. VCG). Each configuration was trained and its hyperparameters optimized using its own training and validation subsets. The final performance of each optimized configuration was evaluated on the held-out test set to ensure a fair and rigorous comparison.

Each experimental configuration was trained using the Adam optimizer, a Focal Loss function to address class imbalance and regularization techniques such as Lead Dropout and a learning rate scheduler. To optimize the model, an empirical hyperparameter search was conducted (including learning rate, weight decay, alpha and gamma), selecting the combination that maximized the AUPRC on the validation set. Early stopping was employed to prevent overfitting during this process.

4. Results and Discussions

4.1. Data & Sampling Strategy

Table 1 summarizes the composition of the datasets used. It details the three training set configurations generated by varying the balance ratio, along with the class distribution of the held-out test set, which has a 2% prevalence of positives. This structure allows us to evaluate how

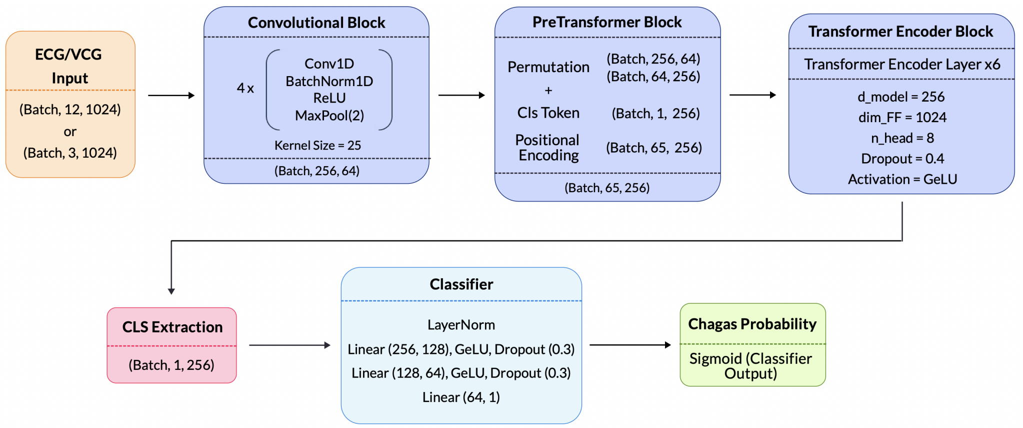

Figure 1. Model Architecture Scheme.

different training data compositions affect performance in a realistic and consistent testing scenario.

Table 1. Distribution of classes in the datasets used for training and evaluation.

Dataset	N Positives	N Negatives	Prevalence
Ratio 1:1	7392	7392	50.0%
Ratio 3:1	7392	22176	25.0%
Ratio 5:1	7392	36960	16.7%
Held-out test set	798	39003	2%

4.2. Model Performance Analysis

Table 2 summarizes two main findings regarding model performance: First, increasing the balance ratio during training consistently enhances performance. For instance, in the ECG-12 representation, the Challenge Score increases from 0.363 (ratio 1:1) to 0.405 (ratio 5:1), indicating that greater exposure to a broader range of negative cases improves model generalization.

Table 2. Performance of best hyperparameters configuration for each combination of signal representation and balance ratio on the held-out test set.

ID	Repr.	Ratio	AUROC	Challenge Score	AUPRC
M1	ECG-12	1:1	0.805	0.363	0.139
M2	ECG-12	3:1	**0.819**	0.388	0.153
M3	ECG-12	5:1	0.811 (0.652)	**0.405 (0.174)**	**0.160 (0.059)**
M4	VCG-3	1:1	0.791	0.362	0.130
M5	VCG-3	3:1	0.805	0.377	0.146
M6	VCG-3	5:1	0.810	0.402	0.159

The M3 model was our final submission. The average metrics achieved on the official hidden test sets are in parentheses.

Second, the 12-lead ECG representation slightly outperforms the VCG. Although the VCG provides a more compact representation, its performance is consistently lower, as observed in the Challenge Score (0.405 for ECG-12 vs. 0.402 for VCG-3 with a 5:1 ratio). A plausible explanation is that the DT, being an estimate, may introduce subtle distortions that degrade diagnostic information.

Consequently, the M3 model (ECG-12, ratio 5:1) emerges as the optimal configuration, underscoring the importance of maximizing the volume of training data while preserving the full richness of the original input signal.

Table 3. Final comparison of best performance by signal representation in the held-out test set (2% prevalence).

Metrics	ECG-12 (M3)	VCG-3 (M6)
Challenge Score	**0.405 (0.174)**	0.402
AUPRC	**0.160 (0.059)**	0.159
AUROC	**0.811 (0.652)**	0.810
F1 Score	**0.152 (0.092)**	0.146
Precision	**0.088**	0.084
Recall	**0.564**	0.551

The average metrics achieved on the official hidden test sets are in parentheses. VCG model was not submitted

Table 3 presents a direct performance comparison between our two best final configurations, one based on 12-lead ECG (M3) and the other on 3-lead VCG (M6), both evaluated on the challenging held-out test set with a 2% positive prevalence.

The analysis reveals that the model using the 12-lead ECG representation demonstrates a consistent, albeit slight, superiority across most key metrics. This reinforces that while the VCG representation is more compact, the unaltered signal information present in the full 12 leads is beneficial for the model's discriminative capacity in this task.

On the held-out test set, the performance of the M3 model (ECG-12, 5:1 ratio) demonstrated its effectiveness in a realistic screening scenario. It achieved a high recall of 0.564, meaning it identified over half of the affected individuals. In this low-prevalence environment, the consequence was a precision of 0.088, a value over four times the 2% baseline, indicating highly informative alerts.

In the official evaluation phase, the model achieved a mean Challenge Score of 0.174. This performance drop from our development set is likely due to domain generalization challenges, such as different disease prevalence and patient demographics in the hidden data.

5. Conclusion

This study demonstrates, beyond the challenge, the potential of deep learning as a tool for the early diagnosis of ChD. Our model is proposed as an effective initial screening system, designed to identify high-risk patients who would benefit most from a confirmatory serological test. This approach could help prioritize clinical resources and streamline the diagnostic process for at-risk populations.

Although overall performance remains moderate, this analysis provides two key methodological insights. First, the exploration of VCG-derived representations highlights their feasibility as a compact alternative to conventional 12-lead ECGs, even if their current performance is slightly lower. Second, the systematic evaluation of data balancing strategies demonstrates their clear effect on model generalization, emphasizing the importance of dataset composition in arrhythmia classification. These findings contribute to a better understanding of model behavior under realistic and imbalanced conditions.

The main limitation is its low precision, which leads to a high rate of false positives, a common challenge in low-prevalence problems. However, we propose its use as a clinical decision support system, where the model's alerts act as a "second reader" to motivate a more thorough review of the case by a specialist. This synergy between AI and clinical expertise represents a promising avenue for improving early detection and preventing irreversible cardiac damage.

The competitiveness of this approach was validated in the official phase of the Physionet Challenge, where our model achieved a mean Challenge Score of 0.174, positioning us among the top 30 teams selected for publication.

Code Availability

The source code for the models and experiments presented in this paper is publicly available on GitHub at: EP-BandoleroLab Team Code

Acknowledgments

This work was supported by PID2022-142514OB-I00 (National Research Program, Ministerio de Ciencia e Innovación, Spanish Government) and CIBERCV CB16/11/00486 (Instituto de Salud Carlos III).

References

[1] Prata A. Clinical and epidemiological aspects of chagas disease. The Lancet Infectious Diseases 2001;1(2):92–100.

[2] Nunes M, Dones W, Morillo C, Encina J, Ribeiro A. Chagas disease: an overview of clinical and epidemiological aspects. Journal of the American College of Cardiology 2013;62(9):767–776.

[3] Rojas L, Glisic M, Pletsch-Borba L, Echeverría L, et al. Electrocardiographic abnormalities in chagas disease in the general population: A systematic review and meta-analysis. PLoS Neglected Tropical Diseases 2018;12(6):e0006567.

[4] Reyna MA, Koscova Z, Pavlus J, Weigle J, Saghafi S, Gomes P, Elola A, Hassannia MS, Campbell K, Bahrami Rad A, Ribeiro AH, Ribeiro AL, Sameni R, Clifford GD. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025. In Computing in Cardiology 2025, volume 52. 2025; 1–4.

[5] Reyna MA, Koscova Z, Pavlus J, Saghafi S, Weigle J, Elola A, Seyedi S, Campbell K, Li Q, Bahrami Rad A, Ribeiro A, Ribeiro ALP, Sameni R, Clifford GD. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025, 2025. URL https://arxiv.org/abs/2510.02202.

[6] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet. Circulation 6 2000;101(23). URL https://pubmed.ncbi.nlm.nih.gov/10851218/.

[7] Kim H, Sunwoo M. An automated cardiac arrhythmia classification network for 45 arrhythmia classes using 12-lead electrocardiogram. IEEE Access 2024;12:44527–44538.

[8] Ribeiro A, Paixao G, Lima E, Horta Ribeiro M, Pinto Filho M, Gomes P, et al. Code-15 https://doi.org/10.5281/zenodo.4916206, 2021.

[9] Ribeiro A, Ribeiro A, Paixao G, Lima E, Horta Ribeiro M, Pinto Filho M, et al. Sami-trop: 12-lead ecg traces with age and mortality annotations (1.0.0). https://doi.org/10.5281/zenodo.4905618, 2021.

[10] Wagner P, Strodthoff N, Bousseljot R, Samek W, Schaeffter T. Ptb-xl, a large publicly available electrocardiography dataset (version 1.0.3). https://doi.org/10.13026/kfzx-aw45, 2022.

[11] Ramirez E, Ruiperez-Campillo S, Casado-Arroyo R, Merino JL, Vogt JE, Castells F, Millet J. The art of selecting the ecg input in neural networks to classify heart diseases: a dual focus on maximizing information and reducing redundancy. Frontiers in Physiology 2024;Volume 15 - 2024. ISSN 1664-042X. URL https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2024.1452829.

Address for correspondence:

Alejandro Pascual-Mellado
ale.pas.mel@gmail.com