CNN-Based Chagas Disease Detection with 12-Lead ECG

Shyamal Y Dharia[1], Mahdis Hojjati[1], Saminur Rahman[1],
Mir Md Taosif Nur[1], Camilo E Valderrama[1,2]

[1] Department of Applied Computer Science, University of Winnipeg, Winnipeg, Canada
[2] Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Canada

Abstract

Limited access to blood tests in underrepresented regions, such as parts of South America, underscores the need for cost-effective, non-invasive methods to identify Chagas disease (CD) in clinical practice. Efficient use of scarce diagnostic resources requires approaches with high sensitivity and low false-positive rates to ensure reliability. To address this challenge, the PhysioNet/CinC Challenge 2025 focused on detecting CD from 12-lead ECG signals by leveraging CD-associated cardiac abnormalities and temporal features. In response, we developed a CNN-based, lead-wise feature learning model that achieved a challenge score of 0.22 in the test phase, ranking 16th out of 41 participating teams. Statistical analysis of feature- and lead-level importance identified RR interval RMSSD as significant across all leads and highlighted the precordial (anterior chest) leads as the most discriminative. These results suggest that emphasizing precordial leads in feature engineering could further improve the accuracy and generalizability of ECG-based CD detection systems.

1. Introduction

Chagas disease (CD), caused by the protozoan *Trypanosoma cruzi*, affects about six million people worldwide [1], particularly in low-resource regions of South America. To address this public health burden, the 2025 George B. Moody PhysioNet Challenge [2] aims to develop computational tools that lessen the reliance on limited clinical expertise.

CD often leads to cardiac complications, which can be detectable using electrocardiogram (ECG) recordings [1]. A recent study highlights the potential of convolutional neural networks (CNNs) in detecting CD from ECG recordings [3]. Motivated by these advances, we propose a CNN-based approach that processes lead-wise features derived from 12-lead ECGs to detect CD on multiple datasets.

2. Methodology

2.1. Datasets

To develop the proposed CNN model, we used three publicly available 12-lead ECG datasets with diverse acquisition protocols and patient populations:

- **CODE-15%** – a subset of the CODE dataset, restricted to Part 1 [4].
- **SaMi-Trop** – includes only samples from Chagas-positive subjects [5].
- **PTB-XL** – includes only samples from non-Chagas subjects [6].

2.2. Preprocessing

First, each lead signals were individually normalized to the range $[-1, 1]$ using a min–max scaling function:

$$x'_i = -1 + 2\,\frac{x_i - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}, \tag{1}$$

where x_i is the i^{th} raw signal value from the original signal vector $\mathbf{x}$, x'_i is the normalized value scaled to the range $[-1, 1]$, $\mathbf{x}$ denotes the full vector of raw signal values from a single lead, $\min(\mathbf{x})$ is the minimum value in $\mathbf{x}$, and $\max(\mathbf{x})$ is the maximum value in $\mathbf{x}$.

To ensure consistency across all ECG records, all 12-lead ECG recordings were then downsampled to 100Hz. A 0.5Hz high-pass filter was subsequently applied to suppress baseline wander and other low-frequency artifacts.

2.3. Feature Extraction

After the preprocessing steps, we used the NeuroKit2 Python package [7] to detect the ECG peaks, which later were used to calculate features shown in Table 1.

The resulting feature matrix had the shape (B, L, F), where $B = 256$ (batch size), $L = 12$ (leads), and $F = 12$ (features per lead). The dataset was split into training, validation, and test sets in an 8:1:1 ratio, with

ISSN: 2325-887X DOI: 10.22489/CinC.2025.086

Table 1. Extracted ECG features for each lead after preprocessing.

#	Feature
1	Mean QRS duration (ms)
2	Standard deviation of QRS duration (ms)
3	Mean QT interval (ms)
4	Standard deviation of QT interval (ms)
5	Mean R-wave amplitude (mV)
6	Standard deviation of R-wave amplitude (mV)
7	QRS net deflection (mV)
8	RR interval RMSSD (ms)
9	Mean P-wave amplitude (mV)
10	Standard deviation of P-wave amplitude (mV)
11	Mean P-wave duration (ms)
12	Standard deviation of P-wave duration (ms)

only 5% positive and 95% negative samples per split. Missing values were imputed lead-wise using scikit-learn's `IterativeImputer`. All features were then standardized using z-score normalization.

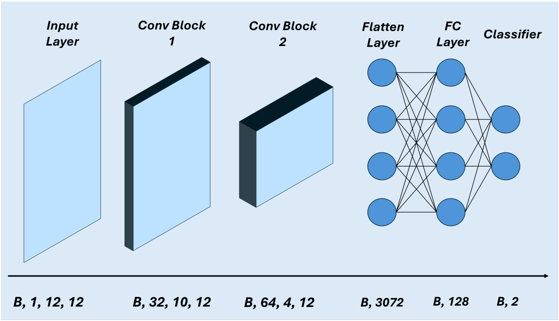

Figure 1. Architecture of the proposed CNN for ECG classification. The $(1 \times L \times F)$ input ($L = 12$, $F = 12$) passes through two convolutional blocks and fully connected layers for binary classification.

2.4. Proposed Architecture

The proposed model was a CNN designed to process lead-wise feature matrices of size $L \times F$ for binary classification. It consisted of two convolutional blocks for hierarchical feature extraction, followed by fully connected layers for classification. The overall proposed model architecture is illustrated in Fig. 1, and Table 2 summarizes the model architecture.

2.5. Loss Functions

Two loss functions were jointly optimized during training: a class-balanced focal loss for classification, and a ranking hinge loss to encourage separation between positive and negative samples.

Table 2. Summary of the proposed CNN architecture.

Stage	Configuration
Input	$(\mathrm{B}, 1, L, F)$
Convolutional Block 1	2D Conv, kernel $(3, 1)$, 32 filters ReLU activation Batch Normalization Dropout $(p = 0.2)$
Convolutional Block 2	2D Conv, kernel $(3, 1)$, 64 filters ReLU activation Batch Normalization Average Pooling $(2, 1)$ Dropout $(p = 0.2)$
Fully Connected Layers	Flatten to 3072-dimensional vector Layer Normalization FC: 3072 $\rightarrow$ 128 ReLU activation Dropout $(p = 0.7)$ FC: 128 $\rightarrow$ 2 (output logits)

Focal loss: We used focal loss to address class imbalance by down-weighting non-Chagas (negative) examples and focusing the training on Chagas (positive) samples. For an input logit vector $\mathbf{z} \in \mathbb{R}^{\mathbb{C}}$ and a target class $y \in \{1, \ldots, C\}$, the focal loss is defined as:

$$L_{\text{focal}} = -\alpha(1 - p_t)^{\gamma} \log(p_t), \tag{2}$$

where $p_y = \frac{\exp(z_y)}{\sum_{c=1}^{C} \exp(z_c)}$ is the predicted probability for the target class, $\gamma > 0$ (Was set it to 2) is the focusing parameter, and α_y is the class weight for class y. The class weights α_y were computed from the training set as:

$$\alpha_y = \begin{cases} 1.0, & \text{if } y = 0, \\ \frac{N_0}{N_1}, & \text{if } y = 1, \end{cases} \tag{3}$$

where N_0 and N_1 are the number of samples in classes chagas negative 0 and positive 1, respectively.

Ranking hinge loss: To encourage a margin between positive and negative predictions, we incorporated a pairwise ranking hinge loss:

$$L_{\text{rank}} = \frac{1}{|P||N|} \sum_{i \in P} \sum_{j \in N} \max\left(0, m - (s_i - s_j)\right), \tag{4}$$

where P and N denote the sets of positive and negative samples, s_i and s_j are the predicted scores, and $m > 0$ is the margin hyperparameter.

Final loss: The total loss is a weighted sum of the two components:

$$L_{\text{total}} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{rank}}, \tag{5}$$

2.6. Training Environment

All experiments were conducted on an NVIDIA RTX A6000 GPU using the AdamW optimizer with a weight decay of 0.1, which helped reduce overfitting. The initial learning rate was 1×10^{-4}, and training ran for 300 epochs with early stopping (patience $= 50$) based on the validation score. A OneCycleLR scheduler increased the learning rate linearly during the first 10% of steps (warm-up) from $\eta_{\max}/25$ to $\eta_{\max}$, then decayed it to $\eta_{\max}/10^4$ over the remaining steps, where $S_{\text{total}} = N_{\text{epochs}} \times N_{\text{batches}}$.

2.7. Evaluation

We submitted our best internal model to the PhysioNet Challenge competition. The official validation was performed using the REDS-II dataset [8], while the test phase included three datasets: REDS-II [8], SaMi-Trop 3 [9], and ELSA-Brasil [10]. Model performance was assessed using multiple metrics, including the Challenge score, area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, and F1-score.

2.8. Statistical Analysis

We computed Cohen's d effect sizes for all 12-lead ECG features to quantify differences between positive and negative CD cases. Statistical significance was assessed using independent t-tests or Mann–Whitney U tests, as appropriate, with Bonferroni FDR correction for multiple comparisons.

3. Results

3.1. Challenge Evaluation

Tables 3 and 4 show the performance of the approach on test datasets, and the comparison of our model with the top five teams. Our approach ranked 16th of 41 with a challenge score of 0.22 (top score: 0.32). Our model performed best on REDS-II (accuracy: 0.88, CS: 0.11) but showed lower performance on SaMi-Trop 3 and ELSA-Brasil, particularly in F1 (0.13, 0.05) and AUCPRC (0.08, 0.04), reflecting reduced sensitivity. AUC-ROC values were moderate across datasets (0.56–0.71), indicating that the model captured a predictive signal but generalized inconsistently across cohorts.

3.2. Statistical Analysis

Our statistical analysis (Figure 2) showed that the RR Interval RMSSD feature was significant across all 12 leads (mean $|d| = 0.38$), with the largest effect in V2 ($d = 0.45$; higher in positive cases). Mean P Duration and Mean P Amplitude were significant on 11 leads, with strongest effects in lead II ($d = -0.64$) and V2 ($d = -0.28$), respectively, indicating lower values in positive cases. Mean R Amplitude and Std QT Interval also showed broad discriminative power (10 significant leads each), peaking at V1 ($d = -0.72$) and V2 ($d = 0.54$).

Table 3. Performance of the model on the official validation (REDS-II) and test datasets (REDS-II, SaMi-Trop 3, ELSA-Brasil) using the following metrics: acc (accuracy), F1 (F1 score), AR (AUCROC), AP (AUCPRC), and CS (Challenge Score).

Dataset	Acc	F1	AR	AP	CS
Validation					
REDS-II	0.91	0.15	0.68	0.13	0.33
Test					
REDS-II	0.88	0.13	0.71	0.21	0.31
SaMi-Trop 3	0.75	0.13	0.71	0.08	0.25
ELSA-Brasil	0.78	0.05	0.57	0.04	0.10
Mean	0.80	0.10	0.66	0.11	0.22
SD	0.07	0.04	0.08	0.09	0.11

Table 4. Comparison of our approach with the top five teams based on the final Challenge Score.

Rank	Challenge Score	Team Name
1	0.323	Biomed-Cardio
2	0.283	DlaskaLabMUI
3	0.280	AIChagas
4	0.271	ISIBrno-AIMT
5	0.269	Ahus AIM
16	**0.220**	**PhysioWinn**

Lead-level analysis revealed that precordial leads V3 and V5 showed significant differences across all 12 features, while V2 and V6 were significant for 11 features. The consistent relevance of the precordial leads (V1–V6) suggests that the anterior chest region is particularly informative for CD detection.

4. Discussion

Our CNN-based approach achieved an average test challenge score of 0.22, demonstrating the effectiveness of lead-wise feature learning. With kernels spanning all leads per feature, the model captured patterns relevant to CD. The potential of the features suggests that using graph- or transformer-based architectures that process features spatially could further improve performance.

Statistical analysis revealed that features extracted from the precordial leads V1–V6 exhibited the highest discriminative power for CD detection. This suggests that CD detection could prioritize anterior chest leads, thus simplifying lead configurations for screening. Future work should focus on developing models that leverage these leads to boost CD detection.

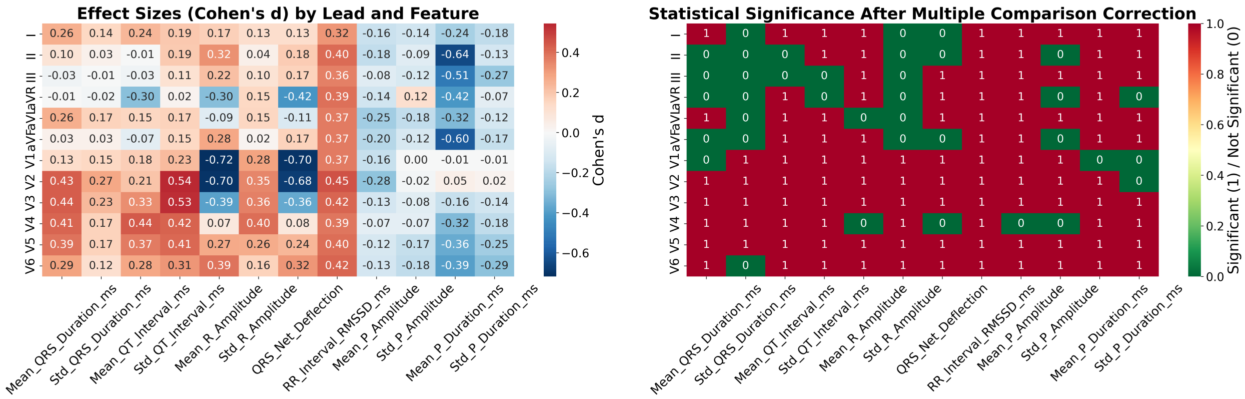

Figure 2. (Left) Cohen's d effect sizes for ECG features across 12 leads comparing Chagas-positive and controls. (Right) Binary significance map after FDR correction ($p < 0.05$; 1 = significant, 0 = not). Tests: t-test or Mann–Whitney U, depending on normality (Shapiro–Wilk) and variance equality (Levene).

5. Conclusion

This study introduces a CNN-based, lead-wise feature learning approach for CD detection. The proposed model, *PhysioWinn*, achieved an official PhysioNet Challenge score of 0.22, ranking 16th out of 41 teams. Statistical analysis showed that the precordial leads (V1–V6) yielded the most discriminative features, suggesting that future work should focus on extracting domain-specific features from these leads to further enhance detection performance.

References

[1] Haro P, Hevia-Montiel N, Perez-Gonzalez J. ECG marker evaluation for the machine-learning-based classification of acute and chronic phases of trypanosoma cruzi infection in a murine model. Tropical Medicine and Infectious Disease 2023;8(3):157.

[2] Reyna MA, Koscova Z, Pavlus J, Weigle J, Saghafi S, Gomes P, Elola A, Hassannia MS, Campbell K, Bahrami Rad A, Ribeiro AH, Ribeiro ALP, Sameni R, Clifford GD. Detection of chagas disease from the ECG: The george b. moody physionet challenge 2025. In Proceedings of the 52nd Computing in Cardiology Conference (CinC), volume 52. 2025; 1–4.

[3] Jidling C, Gedon D, Schön TB, Oliveira CDL, Cardoso CS, Ferreira AM, Giatti L, Barreto SM, Sabino EC, Ribeiro AL. Screening for chagas disease from the electrocardiogram using a deep neural network. PLoS Neglected Tropical Diseases 2023;17(7):e0011118.

[4] Ribeiro AH, Paixao GM, Lima EM, Horta Ribeiro M, Pinto Filho MM, Gomes PR, Oliveira DM, Meira Jr W, Schon TB, Ribeiro ALP. Code-15%: a large scale annotated dataset of 12-lead ECGs, June 2021. URL `https://doi.org/10.5281/zenodo.4916206`.

[5] Ribeiro ALP, Ribeiro AH, Paixao GM, Lima EM, Horta Ribeiro M, Pinto Filho MM, Gomes PR, Oliveira DM, Meira Jr W, Schon TB, Sabino EC. Sami-trop: 12-lead ECG traces with age and mortality annotations, June 2021. URL `https://doi.org/10.5281/zenodo.4905618`.

[6] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. Ptb-xl, a large publicly available electrocardiography dataset. Scientific data 2020;7(1):1–15. URL `https://doi.org/10.1038/s41597-020-0495-6`.

[7] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SHA. NeuroKit2: A python toolbox for neurophysiological signal processing. Behavior Research Methods feb 2021;53(4):1689–1696. URL `https://doi.org/10.3758%2Fs13428-020-01516-y`.

[8] Buss LF, Bes TM, Pereira A, Natany L, Oliveira CDL, Ribeiro ALP, et al. Deriving a parsimonious cardiac endpoint for use in epidemiologic studies of chagas disease: Results from the REDS-II cohort. Revista do Instituto de Medicina Tropical de São Paulo 2021;63:e31.

[9] Ribeiro ALP, Ribeiro AH, Paixão GMM, Lima EM, Horta Ribeiro M, Pinto Filho MM, Gomes PR, Oliveira DM, Meira Jr W, Schön TB, Sabino EC. SaMi-Trop: 12-lead ECG traces with age and mortality annotations, June 2021. URL `https://zenodo.org/record/4905618`. Version 1.0.0.

[10] Aquino EML, Barreto SM, Bensenor IM, Carvalho MS, Chor D, Duncan BB, Lotufo PA, Mill JG, Molina MdC, Mota ELA, Passos VMA, Schmidt MI, Szklo M. Brazilian longitudinal study of adult health (ELSA-Brasil): Objectives and design. American Journal of Epidemiology 2012; 175(4):315–324.

Address for correspondence:

Camilo E. Valderrama
University of Winnipeg, 515 Portage Ave, Winnipeg, MB, Canada R3B 2E9
c.valderrama@uwinnipeg.ca