# 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, highlights the need for cost-effective and non-invasive methods to identify Chagas disease (CD). CD, caused by the protozoan *Trypanosoma cruzi*, is a neglected tropical disease affecting an estimated six million people worldwide. Efficient use of limited diagnostic resources requires approaches that reduce false positives while increasing the detection of true positive cases. To address this need, the PhysioNet/CinC Challenge 2025 was organized to detect CD using 12-lead ECG signals, leveraging the fact that CD can cause cardiac abnormalities detectable in ECG waveforms. In this study, as part of the PhysioNet 2025 Challenge, we developed a CNN-based lead-wise feature learning model for CD detection. Our team, PhysioWinn, achieved a challenge score of 0.326 on the official leaderboard, ranking 22nd out of 67 teams. In addition, we performed a comprehensive statistical analysis to assess feature- and lead-level importance, revealing that RR Interval RMSSD was significant across all leads and that the most discriminative features were concentrated in the precordial (anterior chest) leads. These findings suggest that targeted feature engineering in precordial leads may further improve CD detection in future work.**

## I. Introduction

Chagas disease (CD), caused by the protozoan *Trypanosoma cruzi*, is a neglected tropical disease that affects an estimated six million people worldwide [1]. In its chronic phase, the disease frequently leads to cardiac complications, including conduction abnormalities and heart failure, many of which are detectable in electrocardiogram (ECG) recordings [1]. The 12-lead ECG is a widely available, non-invasive, and cost-effective diagnostic tool [2]. However, interpreting these recordings manually, especially for early-stage Chagas-related cardiac changes, is time-consuming and demands specialized expertise. This has motivated research into automated, AI-powered ECG analysis methods.

Recent studies have demonstrated the potential of convolutional neural networks (CNNs) to detect CD directly from ECG waveforms [3]. Similarly, AI-driven ECG analysis has shown promising accuracy in identifying left ventricular systolic dysfunction among Chagas patients(AUC $\approx$ 0.86) [4]. More broadly, DNNs trained on large-scale 12-lead ECG datasets have outperformed cardiology residents in identifying diverse cardiac conditions [5].

Motivated by these successes, we propose a CNN-based model that processes lead-wise feature matrices extracted from 12-lead ECGs to detect CD. We evaluate its performance on multiple datasets with diverse clinical and demographic characteristics. Our contributions include:

- Designing an efficient CNN architecture tailored to lead-wise ECG features.
- Extracting clinically meaningful morphological and temporal ECG features for downstream classification.
- Evaluating generalization across datasets including CODE-15%, SaMi-Trop, and PTB-XL.

## II. Methodology

### A. Datasets

To evaluate the proposed methods, we employed 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 [6].
- **SaMi-Trop** – the complete dataset was used in our experiments. All recordings in this dataset are associated with *positive* labels, reflecting the presence of the target condition [7].
- **PTB-XL** – the complete dataset was included. In contrast to SaMi-Trop, all recordings in this dataset are associated with *negative* labels [8].

These datasets exhibit substantial variation in recording duration, sampling frequency, diagnostic distribution, and demographic composition, and were therefore also recommended by the challenge organizing committee.

### B. 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 $i^{\text{th}}$ 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. Finally,

we processed five ECG records in parallel across multiple CPU cores, which significantly reduced preprocessing time and improved workflow efficiency.

### C. Feature Extraction

After the initial preprocessing steps, we extracted ECG features, as summarized in Table I. The NeuroKit2 Python package was used to detect peaks, and the corresponding values for each listed ECG feature were computed.

TABLE I: 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) |

These features can be grouped into four categories:

- **Basic interval features**: mean and standard deviation of the QRS duration and QT interval.
- **Amplitude features**: statistics of R-peak amplitudes and the net electrical deflection during the QRS complex.
- **Heart rate variability (HRV)**: computed using the root mean square of successive differences (RMSSD) between RR intervals.
- **P-wave morphology**: average and variation in both amplitude and duration of the P-wave.

The resulting feature matrix had the shape $(B, L, F)$, where $B$ is the batch size (set to 256), $L$ is the total number of leads (12), and $F$ is the total number of features per lead (12). The dataset was then split into training, validation, and testing sets in an 8:1:1 ratio. Notably, in each split, only 5% of the samples belonged to the positive class, with the remaining 95% belonging to the negative class. Furthermore, we used lead-wise imputation using scikit-learn's IterativeImputer, which replaces missing values (NaN) with estimates derived from the relationships among features within each lead.

Finally, all features were standardized using z-score normalization:

$$z = \frac{x - \mu}{\sigma}, \tag{2}$$

where $x$ is the raw feature value, $\mu$ is the mean, and $\sigma$ is the standard deviation, both computed from the training set.

### D. Proposed Architecture

The proposed model is a convolutional neural network (CNN) designed to process lead-wise feature matrices of size $L \times F$ for binary classification. It consists 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 II summarizes the hyperparameters.

TABLE II: 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) |

This architecture uses convolutional layers to capture spatial dependencies across leads and feature dimensions, while the fully connected layers consolidate these learned representations for the final binary classification.

### E. 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.

*a) Focal loss.:* We used focal loss [9] to address class imbalance by down-weighting non-Chagas (negative) examples and focusing the training on Chagas (positive) samples. For an input logit vector $\mathbf{z} \in \mathbb{R}^C$ and a target class $y \in \{1, \ldots, C\}$, the focal loss is defined as:

$$\mathcal{L}_{\text{focal}} = -\alpha_y (1 - p_y)^\gamma \log(p_y), \tag{3}$$

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 $\alpha_y$ is the class weight for class $y$. The class weights $\alpha_y$ were computed from the training set as:

$$\alpha_y = \begin{cases} 1.0, & \text{if } y = 0, \\ \frac{N_0}{N_1}, & \text{if } y = 1, \end{cases} \tag{4}$$

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

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

$$\mathcal{L}_{\text{rank}} = \frac{1}{|P||N|} \sum_{i \in P} \sum_{j \in N} \max\left(0, m - (s_i - s_j)\right), \tag{5}$$

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.

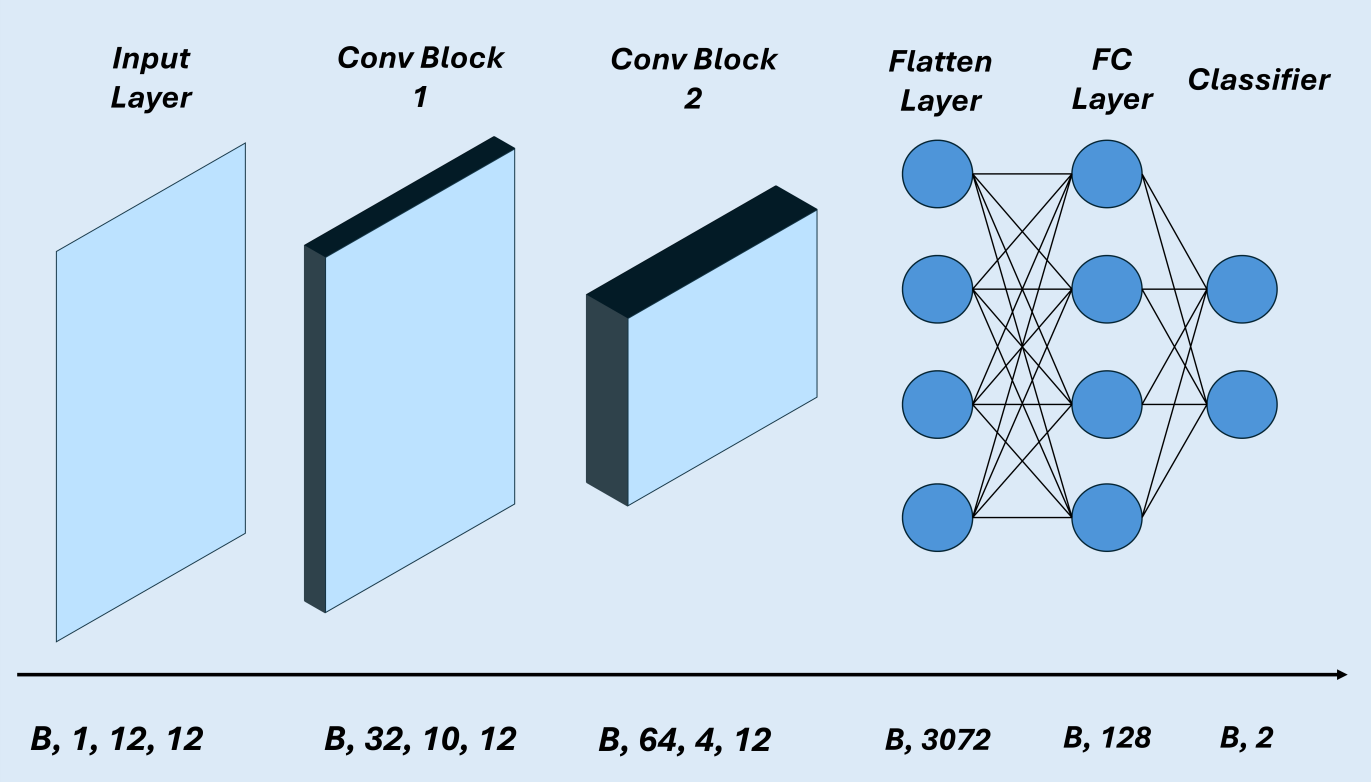


Fig. 1: Overall architecture of the proposed CNN for ECG classification. The network processes a $(1 \times L \times F)$ input, where L is the number of leads (12) and F is the number of features (12) through two convolutional blocks, followed by fully connected layers for binary classification.

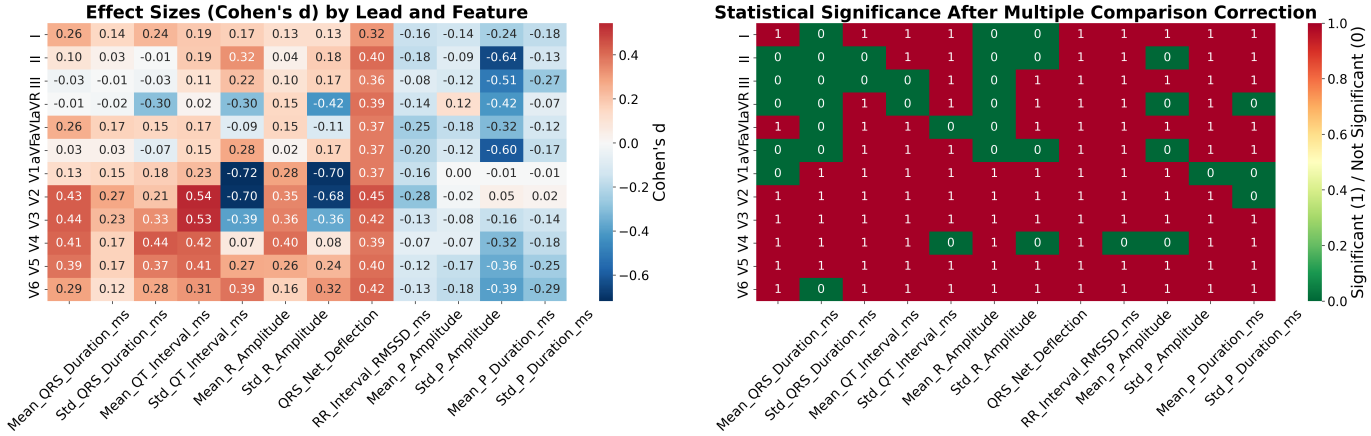


Fig. 2: (Left) Cohen's $d$ effect sizes for each ECG feature across all 12 leads, quantifying the magnitude of difference between Chagas-positive and control groups. (Right) Binary significance map after Bonferroni false discovery rate (FDR) correction, where 1 indicates statistical significance ($p < 0.05$) and 0 indicates non-significance. For each lead–feature pair, normality (Shapiro–Wilk) and variance equality (Levene's test) were assessed; if both assumptions were met, an independent two-sample $t$-test was applied, otherwise a Mann–Whitney U test was used.

*c) Final loss.:* The total loss is a weighted sum of the two components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{rank}}, \tag{6}$$

### F. Training Environment

All internal evaluations were performed on an NVIDIA RTX A6000 GPU. We used the AdamW optimizer with a weight decay of 0.1. A relatively strong decay yielded better performance and reduced overfitting during our intermediate

evaluations. The initial learning rate was set to $1 \times 10^{-4}$, and training was scheduled for 300 epochs with early stopping patience of 50 epochs based on the validation challenge score.

A OneCycleLR [10] scheduler was used to gradually increase the learning rate during the first $10\%$ of training steps (warm-up) from $\eta_{\max}/25$ to the maximum learning rate $\eta_{\max}$, and then decrease it to $\eta_{\max}/10^4$ over the remaining steps. The total number of steps was defined as $S_{\text{total}} = N_{\text{epochs}} \times N_{\text{batches}}$. With these configurations, we performed a 3-fold cross-validation to report our internally evaluated challenge score.

## III. Results

### A. Machine Learning

As shown in Table III, our 3-fold cross-validation achieved a mean accuracy of $0.88 \pm 0.01$, an F1 score for positive samples of $0.31 \pm 0.02$, an AUC-ROC of $0.81 \pm 0.01$, and a challenge score of $0.32 \pm 0.02$. This performance is consistent with the results reported on the official leaderboard, indicating no evidence of overfitting.

TABLE III: Cross-validation results for each fold. Mean row shows mean ± SD. CS = Challenge Score.

| Fold | Acc | F1 | AUC-ROC | CS |
|---|---|---|---|---|
| 1 | 0.89 | 0.32 | 0.82 | 0.35 |
| 2 | 0.86 | 0.27 | 0.79 | 0.30 |
| 3 | 0.90 | 0.33 | 0.81 | 0.31 |
| Mean ± SD | $0.88 \pm 0.01$ | $0.31 \pm 0.02$ | $0.81 \pm 0.01$ | $0.32 \pm 0.02$ |
| Leaderboard | – | – | – | 0.32 |

### B. Statistical Analysis

Our statistical analysis, shown in Figure 2, revealed that the RR Interval RMSSD feature was significant across all 12 leads (mean $|d| = 0.387$), with the largest effect observed in V2 ($d = 0.455$; higher in positive cases). Mean P Duration and Mean P Amplitude were significant on 11 leads, with the strongest effects in lead II ($d = -0.644$) and V2 ($d = -0.284$), respectively, indicating lower values in positive cases. Mean R Amplitude and Std QT Interval also showed broad discriminative power (10 significant leads each), with peak effects at V1 ($d = -0.718$) and V2 ($d = 0.544$).

Lead-level importance analysis indicated that precordial leads V3 and V5 exhibited significant differences for all 12 features considered, while V2 and V6 were significant for 11 of the 12 features. Although less consistent overall, V1 achieved the single largest absolute effect size ($|d| = 0.718$). Overall, the precordial leads (V1–V6) concentrated the strongest and most consistent effects, suggesting that discriminative information for CD detection is primarily localized to anterior chest leads.

## IV. Discussion

Our CNN model achieved a challenge score (CS) of 0.326, demonstrating the potential of our lead-wise feature learning approach. The CNN architecture, with kernel sizes set to process one feature across all leads, successfully identified patterns yielding a CS comparable to our cross-validation performance (0.32). We believe this approach can be further scaled using more advanced architectures such as graph neural networks or transformer-based models.

From the statistical analysis, we demonstrated that several features exhibit strong and consistent discriminative power, particularly in precordial leads V1–V6. This suggests that CD detection efforts could focus more on anterior chest leads, potentially enabling simplified lead configurations for screening. Future work should explore targeted feature engineering and model architectures optimized for these leads, aiming to uncover additional unique and discriminative patterns in Chagas-positive cases.

## V. Conclusion

This study presented a CNN-based lead-wise feature learning approach for CD detection using 12-lead ECG signals. Our method achieved an official PhysioNet Challenge score of 0.326, ranking **PhysioWinn** at 22nd out of 67 teams. Statistical analysis revealed that precordial leads (V1–V6) consistently produced the most discriminative features, suggesting that future work should focus on extracting advanced, domain-specific features from these leads. Such targeted feature engineering, combined with more advanced deep learning architectures, has the potential to further improve 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] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine 2019;25(1):65–69.

[3] Jidling C. Screening for chagas disease from the electrocardiogram using deep neural networks. PLOS Neglected Tropical Diseases 2023;.

[4] Brito BOF, Attia ZI, et al. Left ventricular systolic dysfunction predicted by artificial intelligence using the electrocardiogram in chagas disease patients. PLoS Neglected Tropical Diseases 2021;15(12):e0009974.

[5] Ribeiro ALP, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. arXiv preprint arXiv190401949 2019;.

[6] 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.

[7] 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.

[8] 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.

[9] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision. 2017; 2980–2988.

[10] Smith LN, Topin N. Super-convergence: Very fast training of neural networks using large learning rates. In Artificial intelligence and machine learning for multi-domain operations applications, volume 11006. SPIE, 2019; 369–386.