Transfer Learning and Soft Labels Enable Robust ECG-Based Detection of Chagas Disease

Bas BS Schots[1], Bauke KO Arends[1], Dino Ahmetagic[1], Camila S Pizarro[1], Tim Paquaij[1], Pim van der Harst[1], Rutger R van de Leur[1,2], René van Es[1,2]

[1]University Medical Center Utrecht, Utrecht, The Netherlands
[2]Cordys Analytics, Utrecht, The Netherlands

Abstract

Chagas disease (CD) is a tropical parasitic disease that often remains asymptomatic but can lead to serious long-term cardiac complications. This study describes our contribution to the George B. Moody PhysioNet Challenge 2025, which focused on developing deep learning algorithms to detect CD using 12-lead electrocardiogram (ECG) data. We trained a convolutional neural network initialized with weights from ECGFounder, a foundation model pretrained on over 10 million ECGs. Model development used multinational datasets and addressed label noise by applying soft labels to ECGs with features suggestive of asymptomatic CD, such as right bundle branch block and atrial fibrillation. The model was evaluated using the true positive rate among the top 5% of predictions (TPR@5%), reflecting a scenario of resource-constrained deployment. Our team, UMC Utrecht, achieved a TPR@5% challenge score of 0.280, resulting in a 130th overall place. These findings underscore the potential of ECG-based tools for screening CD in settings with limited access to serological testing.

1. Introduction

Chagas disease (CD) is a zoonotic parasitic disease caused by the protozoan *Trypanosoma cruzi* and primarily transmitted by triatomine insects. Although endemic to Latin America, global migration has led to its spread to non-endemic regions, including Europe, North America, and Australia, with an estimated 6 million people affected worldwide [1].

The early acute phase of Chagas disease is curable with timely treatment, but up to 70% of individuals remain asymptomatic for years. A subset of these patients eventually develop chronic Chagas cardiomyopathy (ChCM), which may manifest decades after the initial infection. This condition encompasses a spectrum of cardiac involvement, from isolated conduction disturbances and mild wall motion defects to severe heart failure, thromboembolic complications, and life-threatening ventricular arrhythmias [1,2]. Once established, ChCM is associated with high morbidity and mortality, and treatment options are limited, making early identification of at-risk individuals important.

Electrocardiographic (ECG) abnormalities often represent the transition from asymptomatic CD to clinically apparent ChCM. Although no single electrical pattern is pathognomonic for CD, common electrocardiographic findings include right bundle branch block (RBBB) with or without left anterior hemiblock, and atrial fibrillation (AF) [2]. Detecting such abnormalities, or even more subtle electrical signatures, provides an opportunity to identify infected patients before progression to overt cardiomyopathy, when preventive measures and monitoring may still alter outcomes.

This forms the rationale for the 2025 George B. Moody PhysioNet Challenge, which calls for the development of models to screen for CD using ECG data [3]. As a contribution to this effort, we developed a convolutional neural network to detect CD based solely on ECG input.

2. Methods

2.1. Datasets

Model development was based on the provided challenge datasets, which consist of three sources: the CODE-15% dataset [4], which includes over 300,000 12-lead ECG recordings from Brazil, with self-reported binary CD labels; the SaMi-Trop dataset [5], containing 1,631 12-lead ECGs from serologically confirmed CD patients in Brazil; and the PTB-XL dataset [6], which includes 21,779 12-lead ECGs from Europe, presumed to be CD-negative. A total of 329,280 patients were included in the model derivation dataset, and 36,905 patients were held out for internal benchmarking purposes.

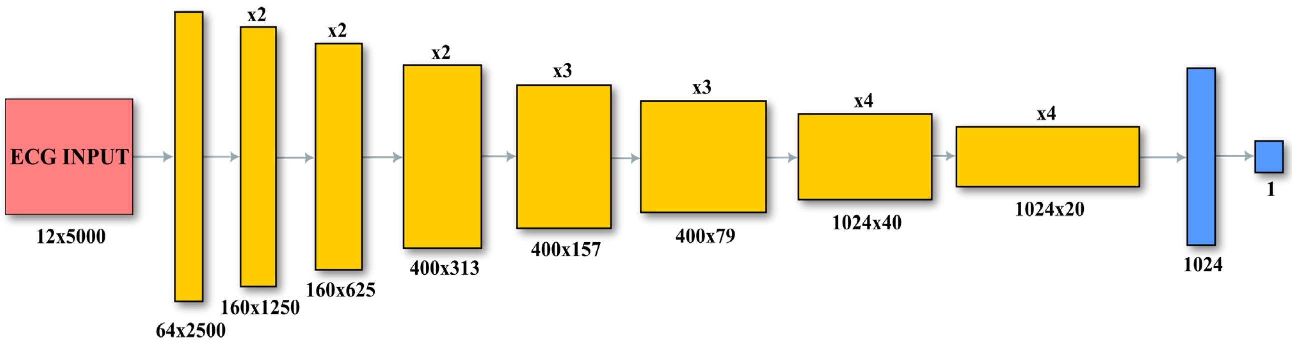

Figure 1. ECGFounder architecture for transfer learning. Input ECG (12×5000) passes through one convolutional (64 filters) and seven residual stages (yellow) with progressive downsampling and channel expansion (64 to 1024; 5000 to 20). Representations are aggregated and passed to a fully connected output layer (blue).

2.2. Preprocessing

Model training was based solely on raw ECG signal data. ECGs were excluded if deemed excessively noisy, defined as having amplitudes exceeding ±10 mV after mean-centering, or a standard deviation greater than 3 mV. This led to the exclusion of 13,137 (3.99%) of ECGs.

Remaining ECGs were resampled to 500 Hz where necessary, followed by application of a bandpass filter (0.5-100 Hz) and notch filters at 50 Hz and 60 Hz to remove powerline interference. Signals were either cropped or zero-padded to a fixed length of 5,000 samples (10 s). Both transformations were centered, meaning that cropping and zero padding both took place at the beginning and end of the waveform in equal parts. The model derivation cohort was split into a training (90%) and internal validation (10%) set.

The classification task used binary labels indicating the presence of CD, derived from either serologically confirmed or self-reported sources. For patients in the CODE-15% dataset with a negative label but ECG abnormalities suggestive of asymptomatic CD, soft labels were applied. Soft probabilities were assigned based on log odds ratios (OR) associated with RBBB (OR 4.6) and AF (OR 1.7), as described in [2]. These probabilities were scaled and capped at 0.49 to prevent reclassification as positive cases.

2.3. Model Architecture

We employed transfer learning using ECGFounder, a pretrained ECG foundation model [7]. This model, trained on over 10 million 12-lead ECGs from the Harvard-Emory database, is designed to extract generalizable representations of cardiac electrical activity. In short, the model was trained using a multilabel ECG classification task, with an adapted loss function to account for unlabelled positive cases.

The input to ECGFounder is a 12×5000 ECG matrix (12 leads, 5000 time steps). The first layer applies a 1D convolution with 64 filters, followed by a sequence of residual convolutional stages with squeeze-and-excitation blocks and skip connections to stabilize training (Figure 1). The network progressively downsamples the temporal dimension while expanding the channel depth, extracting increasingly abstract features. By the final stage, representations are aggregated and passed to a fully connected output layer.

For our task, we initialized the backbone with pretrained ECGFounder weights and replaced the final classification head with a single sigmoid output neuron for CD prediction. We finetuned the network in two parameter groups: the backbone was updated conservatively (learning rate = 1×10^{-5}) to preserve pretrained features, while the classification head was trained with a higher learning rate (1×10^{-4}) to adapt to the new task. Optimization used AdamW with separate learning rate schedules**:** a cosine annealing warm restart scheduler for the backbone ($T_0 = 8$, $\eta_{\min} = 1\times10^{-6}$), allowing gradual and cyclic learning rate reductions to encourage stable convergence without catastrophic forgetting, and a linear warmup scheduler for the head (start factor = 0.1, 3 iterations), ensuring smooth adaptation from randomly initialized weights. This combination balanced conservative fine-tuning of the pretrained representation with more aggressive training of the classification head.

2.4. Model Training and Evaluation

Training was performed using an NVIDIA RTX A6000 GPU. We used a batch size of 100, for up to 10 epochs using binary cross-entropy loss with logits. Early stopping (patience = 3 epochs) was applied, and the best checkpoint (lowest validation loss) was retained. Additional metrics (AUROC, accuracy, precision, recall, F1) were monitored for completeness but not used for checkpointing.

Model performance was assessed using the challenge score, defined as the true positive rate for the binary prediction among the top 5% of model predictions

(TPR@5%). Model's area under the receiver operating curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, and F-measure were also calculated.

3. Results

On the official hidden data, our model achieved a challenge score of 0.280, placing it rank 130/367. Internal 5-fold cross-validation on the training dataset yielded a challenge score of 0.402 (range 0.388-0.415) and an AUROC of 0.842 (range 0.841-0.843) (Table 1). This should be viewed relative to the maximum attainable challenge score of 0.455, determined by the 2.3% prevalence in the 316,143 total cases in the training set.

Table 1. Challenge score on the hidden test set and classification metrics from 5-fold cross-validation on the training data.

	Mean (range)
Hidden test set	
Challenge score	0.280
5-fold cross-validation training data	
Challenge score	0.402 (0.388-0.415)
AUROC	0.842 (0.841-0.843)
AUPRC	0.190 (0.184-0.197)
Accuracy	0.976 (0.975-0.977)
F-measure	0.183 (0.164-0.197)

Abbreviations. AUPRC, area under the precision recall curve; AUROC, area under the receiver operating curve.

Table 2. Baseline characteristics stratified by top and bottom 5% of predicted probabilities.

	Top 5%	**Bottom 5%**
CD, n (%)	2,885 (18.3%)	10 (0.1%)
Age, mean (SD)	67.3 (14.6)	42.4 (24.8)
Male sex, n (%)	7,835 (49.6)	7,331 (46.4)
SaMi-Trop, n (%)	570 (3.6)	2 (0.0)
CODE-15%, n (%)	15,151 (95.8)	9,180 (58.1)
PTB-XL, n (%)	87 (0.6)	6,626 (41.9)
LBBB, n (%)*	403 (2.6)	86 (0.5)
RBBB, n (%)*	7,606 (49.9)	246 (1.6)
AF, n (%)*	2,301 (15.1)	85 (0.5)
1dAVb, n (%)*	816 (5.4)	171 (1.1)
Normal ECG, n (%)	294 (1.9)	9,800 (62.0)

*Not available for SaMi-Trop cases.

Abbreviations. AF, atrial fibrillation; 1dAVb, first degree atrioventricular block; CD, Chagas disease; LBBB, left bundle branch block; RBBB, right bundle branch block.

When comparing patients in the top versus bottom 5% of predicted probabilities, individuals in the highest probability group were older (mean age 67.3±14.6 vs 42.4±24.8 years), more frequently male (49.6% vs 46.4%), and more commonly from the SaMi-Trop and CODE-15% dataset (95.8% vs 58.1% and 3.6% vs 0.0%) (Table 2). This group also showed a higher prevalence of electrocardiographic abnormalities, including left bundle branch block (2.6% vs 0.5%), right bundle branch block (49.9% vs 1.6%), atrial fibrillation (15.1% vs 0.5%), and first-degree atrioventricular block (5.4% vs 1.1%).

4. Discussion

We applied transfer learning with ECGFounder, a large-scale foundation model pretrained on more than 10 million ECGs, to the task of CD detection from 12-lead ECGs. Our approach achieved a challenge score of 0.280 on the hidden test set data, ranking 130th overall. Internal 5-fold cross-validation yielded a higher score of 0.402. The gap between internal and external performance likely reflects population and label differences, as well as some overfitting to the training distribution. Importantly, the prevalence of CD was similar across datasets and is consistent with real-world settings (~2%). In this context, our findings should be interpreted as proof of concept: ECG-based AI can detect patterns relevant to CD, but our performance was likely constrained by reliance on weakly labeled data, pretraining primarily on non-endemic populations, and potential signal loss from fixed-length signal standardization.

A key strength of our strategy was the use of generalizable ECG representations learned from very large, diverse datasets. Rather than training a network from scratch, we transferred broad ECG knowledge to the CD setting. By finetuning the backbone conservatively while adapting the classification head more aggressively, we preserved pretrained features while tailoring the model to the challenge goal. In addition, the use of soft labels for ECG patterns strongly associated with CD (e.g., right bundle branch block, atrial fibrillation) allowed better exploitation of weakly labeled data and reduced the influence of mislabeled negatives. Together, these design choices likely enhanced robustness to class imbalance and label noise and facilitated detection of subtle abnormalities consistent with CD.

Several limitations should be acknowledged. First, ECGFounder was primarily trained on non-endemic populations, which may limit its ability to fully capture disease-specific features in CD. Second, signal standardization to a fixed 10-second window was suboptimal, as cropping discards useful information while padding includes a large section of flat segments in the training data that may have distorted model learning by making it adapt to artificial patterns rather than physiological signals. To address this, we also trained models using a fixed 6-second input to avoid zero padding; however, performance was inferior, suggesting that the additional signal content outweighed the drawbacks of padding. Third, the CODE-15% dataset relied on self-reported labels, making mislabeling likely. Given the low disease prevalence, even a small proportion of mislabeled cases can have a disproportionate influence on model training. Although the use of soft labels partly mitigated this, label noise remained a challenge.

Looking ahead, the major challenge for AI-based ECG screening of CD remains the combination of low disease prevalence and imperfect labels. Strategies such as oversampling, synthetic data generation, or multitask learning on CD-associated abnormalities may help improve sensitivity to rare cases, but ultimately progress depends on access to larger, serologically confirmed datasets. Moreover, models that can process variable-length inputs without padding or cropping may make better use of the available ECG signals. For deployment in practice, systematic storing of predictions and outcomes would enabling continuous validation and iterative refinement of screening strategies.

5. Conclusion

We developed a deep learning model to detect CD from 12-lead ECGs as part of the 2025 George B. Moody PhysioNet Challenge. By combining large multinational datasets with strategies to address label noise and class imbalance, our model achieved reasonable performance and highlighted characteristic ECG features of high-risk patients. While further improvements will depend on larger serologically confirmed cohorts and methods robust to low prevalence, our findings support the potential of ECG-based AI as a scalable screening tool for CD.

Acknowledgments

None.

References

[1] M. C. P. Nunes *et al.*, “Chagas Cardiomyopathy: An Update of Current Clinical Knowledge and Management: A Scientific Statement From the American Heart Association,” *Circulation*, vol. 138, no. 12, Sept. 2018,

[2] L. Z. Rojas *et al.*, “Electrocardiographic abnormalities in Chagas disease in the general population: A systematic review and meta-analysis,” *PLoS Negl. Trop. Dis.*, vol. 12, no. 6, p. e0006567, June 2018,

[3] M. A. Reyna *et al.*, “Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025,” *Comput. Cardiol.*, vol. 52, pp. 1–4, 2025.

[4] A. H. Ribeiro *et al.*, “Automatic diagnosis of the 12-lead ECG using a deep neural network,” *Nat. Commun.*, vol. 11, no. 1, p. 1760, Apr. 2020,

[5] C. S. Cardoso *et al.*, “Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile,” *BMJ Open*, vol. 6, no. 5, p. e011181, May 2016,

[6] P. Wagner *et al.*, “PTB-XL, a large publicly available electrocardiography dataset,” *Sci. Data*, vol. 7, no. 1, p. 154, May 2020,

[7] J. Li *et al.*, “An Electrocardiogram Foundation Model Built on over 10 Million Recordings,” *NEJM AI*, vol. 2, no. 7, June 2025,

Address for correspondence:

René van Es
Department of Cardiology, University Medical Center Utrecht,
Heidelberglaan 100, 3584 CX Utrecht, The Netherlands
r.vanes-2@umcutrecht.nl