Managing Label Uncertainty in the Detection of Chagas Disease from the ECG

Casey Hong[1]*, Chun-Ti Chou[2], Sergio González[2]

[1]School of Medicine, National Taiwan University, Taipei, Taiwan
[2]AI Center, Inventec Corporation, Taipei, Taiwan

Abstract

As part of the PhysioNet/Computing in Cardiology Challenge 2025, our team, AIChagas, trained a ResNet model to detect Chagas disease in the electrocardiogram (ECG) by managing label uncertainty from self-reporting data. We soft-labeled patients with multiple self-claims and negative samples with arrhythmias common in Chagas disease. For cases with multiple reports, we assigned soft probabilities according to their proportion of positive and total reports. In addition, we increased the soft probability of negative samples with arrhythmias commonly found in cases of Chagas cardiomyopathy. The soft probabilities of negative samples were capped to prevent relabeling. To avoid biases when training with multiple data sources, we employed several data processing and augmentation techniques. We introduced positive class weighting to mitigate the imbalance problem. Finally, we trained our model by combining binary cross-entropy (BCE) and margin ranking loss between positive and hard negative samples (20% of negative instances with the highest predicted scores) to enhance the model's ranking capability. In the hidden test set, our model received a challenge score of 0.280 (awarded 3rd out of 41 eligible teams).

1. Introduction

Chagas disease is a highly under-detected disease with no apparent symptoms until chronic phases, such as cardiomyopathy. Cost-effective prescreening, such as automatic early sign detection in electrocardiograms (ECG), is desired, as mass screening is costly and inefficient. However, screening difficulties also hinder data collection, label quality, and, consequently, the training of deep learning models. Furthermore, Chagas serologically positive samples may not have obvious signs in their ECG, potentially confusing the model training. The 2025 George B. Moody PhysioNet Challenge invited teams to develop open-source automatic solutions to this problem [1–3]. We have extended a previous ResNet model for Chagas screening [4] by addressing relevant data problems, such as training biases from multiple datasets [5–9], self-claimed labels, hard negative and positive samples. Our approach involves soft-labeling, comprehensive data augmentation, positive weighting, and a custom ranking loss with hard negative learning.

* Contributed during an internship at Inventec.

2. Methods

2.1. Data Preprocessing and Augmentation

We trained our model on three different datasets with unique characteristics and sources. SaMi-Trop [6] and PTB-XL [7] are small and medium datasets with strong labels, that is, Chagas positive serological data and demographically based negative data, respectively. In contrast, CODE-15% [5] is a significantly larger dataset with self-claimed positive and negative labels. In this context, data preprocessing and augmentation are crucial to avoid deep learning models leveraging source- or device-specific features instead of actual ECG features related to Chagas disease. This behavior is especially concerning in problems where the actual features might be subtle and there are clearly different class distributions across data sources. Our preliminary experiments showed this effect with over-optimistic results but low transferability when training only on SaMi-Trop and PTB-XL.

To address this issue, we first homogenized the data with a series of preprocessing techniques. We reverted the amplitude scale by 2 of SaMi-Trop and CODE-15% [10, 11]. We set the ECG length to 10 seconds and 4096 samples by resampling PTB-XL and padding or cropping shorter or longer signals from SaMi-Trop and CODE-15%. We applied a Butterworth bandpass filter of order 2 between 0.5-60 Hz to remove potential device noise.

To promote model generalization and destroy remaining device-dependent features, we apply additional random augmentations every epoch. Symmetric random masking between 0 and 1.5 seconds was introduced at the beginning and end of the ECG to simulate shorter signals. Random baseline wander was summed into the ECG based on the sine wave $\text{amp} * \sin(2 * \pi * \text{freq} * t)$, where amp is a

constant amplitude of 0.1, freq is the frequency randomly chosen between 0.1 and 0.5, and t is time. We randomly rescaled the signal amplitude by a factor between 0.8 and 1.2. Time warping randomly contracted or stretched the ECG by linear resampling with a random factor between 0.9 and 1.1 of the original length of the ECG signal. We added Gaussian noise with 0 mean and 0.01 standard deviation. Finally, additional masking was applied to a randomly selected segment with a maximum length of 1 second occlusion. Table 1 summarizes these augmentation techniques and their parameters.

2.2. Soft Labeling for Self-Reported Chagas Samples

Chagas disease labels in CODE-15% are from patient reports without serological confirmation. Hence, there is a risk of wrong or noisy labels, especially given the underdiagnosis of Chagas disease and the low prevalence of 1.9% in CODE-15% [2, 3]. Treating the data as high-confident, crisp labels similar to SaMi-Trop and PTB-XL might harm the training of deep learning models.

We applied a static soft labeling strategy to mitigate the impact of possible incorrect labels. Several patients in CODE-15% had multiple self-reported labels that do not necessarily agree. Then, we used these disagreements to signify confidence in their samples. We assigned soft probabilities c_i according to their proportion of Chagas positive and total reports:

$$c_i = 0.51 * y_i + 0.49 * \frac{positive}{total} \tag{1}$$

where y_i is the original label of sample i. This linear combination enables simple soft-labeling for inconsistent reports, but avoiding relabeling. For example, patient 1264 has 3 positive samples out of a total of 5. Their positive and negative samples are assigned with the new probabilities of 0.804 and 0.294, respectively.

This first approach only contributes to a small subset and does not increase the positive class ratio. Thus, we applied a second strategy to negative ECG samples with arrhythmias commonly found in Chagas cardiomyopathy [12], such as right bundle branch block (RBBB), atrioventricular block (AVb), and atrial fibrillation (AF). In CODE-15%, RBBB is 8 times more common in Chagas positive samples (20% among positive vs. 2.4% among negative). AF and AVb are about 2.5 times more common (5.8% vs. 1.9% and 4% vs. 1.6%). Furthermore, the simultaneous presence of RBBB with AF or AVb is 11 and 9 times more frequent. Given these statistics, we increased the soft probability of negative samples with RBBB by 0.15 or 0.25 if AVb or AF are also present. The soft probabilities of negative samples were capped at 0.49 to prevent relabeling. This strategy slightly increases the positive sample ratio from 1.9% to 2.3%.

2.3. 12-Lead ECG Based ResNet Model

The trained deep learning model is based on previous studies for ECG diagnosis [5] and Chagas screening [4]. The network architecture is a ResNet model with 1-dimensional convolutional neural networks (CNN) that consists of an initial downsampling layer and batch normalization, followed by four residual blocks, and ending with flattening and a linear layer with a single output. Each residual block has two blocks of CNN, batch normalization, ReLU activation function, dropout, and a residual connection. This architecture takes ECG as input with the 12 leads as CNN channels and has one output with a sigmoid activation function for binary classification. The model was completely trained from scratch on the challenge datasets. Table 1 shows the detailed settings.

2.4. Combined BCE and Ranking Loss

As a binary classification problem, we could use binary cross-entropy loss (BCE) to train our ResNet model. However, BCE is known to have problems with noisy labels or outliers [13]. In particular, deep learning models tend to overfit on these incorrect labels, resulting in poor generalization. Such issues would probably arise in the detection of Chagas disease, as we are training with self-claimed labels in CODE-15% and positive ECGs in SaMi-Trop with potentially no obvious signs of the disease yet.

Furthermore, this approach aims to prescreen the population to later confirm by serological tests. With the challenge metric, we also maximize the recall or true positive rate within the samples with the highest predicted probability. Thus, we are not looking for a perfect prediction or separation between positive and negative samples, as in BCE, but rather a ranking.

Therefore, we included a margin ranking loss. Although commonly used directly on latent features or with Siamese or triplet nets, this ranking loss in our classification approach aims to directly optimize the relative order of predicted class scores, rather than absolute probabilities, making it more suitable for this task. Compared to BCE, the ranking loss leaves a margin of freedom in the predicted probability, mitigating overfitting. The formulation of the margin ranking loss is the following:

$$L_{\text{rank}} = \max(0, -p_1 + p_2 + \text{margin}_{1,2}) \tag{2}$$

where p_1 is the predicted probability for the actual positive samples, p_2 is the probability for the negative samples, and the margin sets the room of freedom. We also leveraged the previous soft labeling in this loss. Given a positive-negative pair, the difference in confidences of the pair is

Data processing & augmentation	
Signal length	10 seconds
Number of leads	12
Butterworth bandpass filter	Order 2, 0.5-60 Hz
Symmetric random mask	0-1.5 seconds (0-648 samples)
Random baseline wander	amplitude 0.1, frequency 0.1-0.5
Random amplitude scale	0.8-1.2
Random time warping	0.9-1.1 of the original length
Gaussian noise	μ 0, σ 0.01
Random masking	0-1 seconds
ResNet model	
Input size (C, L)	(12, 4096)
Kernel size	17
Number of residual blocks	4
Output channels	[64, 128, 196, 256, 320]
Output length	[4096, 1024, 256, 64, 16]
Dropout rate	0.8
Batch norm. eps. & momentum	1e-05 & 0.1
Activation function	ReLU
Number of classes	1
Optimizer & Loss	
Optimizer	Adam
Learning rate	0.001
Batch size	256
Loss function	Combined BCE & Ranking loss
Positive weight	10
Margin	0.5
Hard negative ratio	0.2
Combined ratio (α)	2.5

Table 1. Hyperparameters for the data augmentation, model and loss.

used as a factor of a constant margin, that is, $\text{margin}_{1,2} = \text{margin} * \max(0.1, c_1 - c_2)$, where the margin is 0.5 in our final model. For example, with $p_1 = 0.5$ and $p_1 = 0.2$, the loss results in $-0.5 + 0.2 + 0.5 = 0.2$. However, if their soft labels are $c_1 = 0.8$ and $c_1 = 0.4$, then the loss is $\max(0, -0.5 + 0.2 + 0.5 * (0.8 - 0.4)) = 0$. That is, the loss function is less strict on low-confidence samples.

This margin ranking loss is a pairwise comparison function applied between positive and negative samples. However, using all negative samples could reduce its impact, especially when there are many "easy-to-learn" negative samples, such as normal sinus rhythm ECGs ($\sim$39% in CODE-15% dataset). Thus, we applied the ranking loss only to hard negative samples, that is, the 20% of negative instances with the highest predicted scores within the given batch.

Using only the ranking loss might cause poor calibration of the predicted probabilities. In addition, batches with only negative samples are reasonably common, given the high imbalance ratio of the problem. Therefore, we combined the two loss functions with the parameter α to weigh the contribution of the ranking loss:

$$L = L_{\text{BCE}} + \alpha * L_{\text{rank}} \tag{3}$$

	All	**PTBvsSaMi**	**Chagas 5%**	**CODE-15%**
Baseline (All) [4]	0.306	0.324	0.326	0.273
Baseline (CCC) [4]	0.295	0.242	0.283	0.290
Data augmentation	0.375	0.341	0.320	0.411
Soft-labeling	0.379	0.352	0.356	0.385
Ranking loss	0.381	0.363	0.354	0.403
Hard negative	**0.402**	**0.389**	**0.388**	**0.416**

Table 2. Scores using challenge metric for our incremental contributions on our internal evaluation sets.

	AUC	**AP**	**Challenge**
Baseline (All) [4]	0.819	0.333	0.326
Baseline (CCC) [4]	0.785	0.234	0.283
Data augmentation	0.830	0.290	0.320
Soft-labeling	0.831	0.314	0.356
Ranking loss	**0.844**	0.310	0.354
Hard negative	0.824	**0.369**	**0.388**

Table 3. Scores using different classification metrics for our incremental contributions on our internal evaluation set with 5% of positive Chagas samples.

3. Results

In order to internally evaluate our models, we randomly reserved part of the training datasets, splitting them by patient identifiers: 50% of SaMi-Trop, 35% of PTB-XL, and 30% of CODE-15%. We combined these subsets to form 4 different evaluation sets. The first set included all subsets. Another consisted only of strong labels from PTB-XL and SaMi-Trop. The third set (Chagas 5%) had all PTB-XL and SaMi-Trop, and randomly selected negative samples from CODE-15% until reaching a prevalence of 5%. The last set included only self-reported samples of CODE-15%. Although we scored our approaches on all four sets, we prioritized our model selection according to the PTBvs-SaMi and Chagas 5% evaluation sets. For evaluation metrics, we used the area under the ROC curve (AUC), the average precision score (AP), and the challenge metric.

Table 2 shows the results using the challenge metric in our four evaluation sets. Table 3 collects the performance of our model in terms of the different metrics with the Chagas 5% subset. We included the results of running the ResNet model with the available online weights [4]. Baseline (All) and baseline (CCC) were trained to detect Cha-

Training	Validation	Test	Ranking
0.402	0.360	0.280	3/40

Table 4. Challenge scores for our selected entry (team AIChagas), including the ranking of our team on the hidden test set among all teams. We evaluated our model on a held-out set on the public training data, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

gas disease and Chagas cardiomyopathy, respectively. Our approaches are depicted in an incremental manner with the name of the added techniques. That is, a given row comprises its named technique and those of the preceding rows. The last row is our final approach that includes augmentation, soft-labeling, and ranking loss with hard negative learning.

The results show a gradual increase in performance with a few exceptions. This incremental improvement is clear with the challenge metric. Our final approach has a significantly better challenge score than the baseline for Chagas detection and is overwhelmingly better than the Chagas cardiomyopathy baseline. Interestingly, the performance of the baseline models is considerably worse in CODE-15% compared to our approach, even though they were trained with such data. In Table 3, the AUC results are slightly inconsistent and overoptimistic compared to the AP scores, which translate better to the challenge metric.

Table 4 presents the challenge scores in the hidden datasets. There is a drop in model performance in validation and test compared to the complete training evaluation set. However, the performance is closer compared to a more realistic subset with a higher prevalence of 5%. Thus, this could indicate decent robustness and transferability.

4. Discussion and Conclusions

In this study, we focus our efforts on managing the label uncertainty of self-reported data and Chagas-positive samples that might lack an obvious sign in their ECG. We trained a ResNet model with a soft-labeling strategy to deal with self-claimed labels and a combined loss of BCE and margin ranking loss to mitigate overfitting on potential outliers. Our experimental results show improvements compared to previous models. Our approach and other challenge proposals could be highly valuable in prescreening a large population with ECG before proceeding with serological confirmations for Chagas disease.

The main limitation of our approach is the static computation of the soft labels. Our procedure would be a bit arbitrary, although it is based on some clinical evidence [12] and data statistics in CODE-15%. During the course of the challenge, we tried several more complex models for classification with noisy labels, such as DivideMix [14], or alternatively, to dynamically assign label confidence during training. However, these approaches did not converge or brought minor performance improvements that we did not submit to the challenge.

References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[2] Reyna MA, Koscova Z, Pavlus J, Weigle J, Saghafi S, Gomes P, et al. Detection of Chagas disease from the ECG: the George B. Moody PhysioNet challenge 2025. In Computing in Cardiology 2025, volume 52. 2025; 1–4.

[3] Reyna MA, Koscova Z, Pavlus J, Saghafi S, Weigle J, Elola A, et al. Detection of Chagas disease from the ECG: the George B. Moody PhysioNet challenge 2025, 2025. URL https://arxiv.org/abs/2510.02202.

[4] Jidling C, Gedon D, Schön TB, Oliveira CDL, Cardoso CS, Ferreira AM, et al. Screening for Chagas disease from the electrocardiogram using a deep neural network. PLoS Neglected Tropical Diseases 2023;17(7):e0011118.

[5] Ribeiro A, Ribeiro M, Paixão G, Oliveira D, Gomes P, Canazart J, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications 2020;11(1):1760.

[6] Cardoso C, Sabino E, Oliveira C, de Oliveira L, Ferreira A, Cunha-Neto E, et al. Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. BMJ Open 2016;6(5):e0011181.

[7] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 2020;7:154.

[8] Nunes MCP, Buss LF, Silva JLP, Martins LNA, Oliveira CDL, Cardoso CS, et al. Incidence and predictors of progression to Chagas cardiomyopathy: long-term follow-up of Trypanosoma cruzi-seropositive individuals. Circulation 2021;144(19):1553–1566.

[9] Pinto-Filho M, Brant L, Dos Reis R, Giatti L, Duncan B, Lotufo P, et al. Prognostic value of electrocardiographic abnormalities in adults from the Brazilian longitudinal study of adults' health. Heart 2021;107(19):1560–1566.

[10] Ribeiro AH, Paixao G, Lima EM, Ribeiro MH, Pinto Filho MM, Gomes PR, et al. CODE-15%: a large scale annotated dataset of 12-lead ECGs (1.0.0), 2021. Zenodo.

[11] Lima EM, Ribeiro AH, Paixão GM, Ribeiro MH, Pinto-Filho MM, Gomes PR, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. Nature Communications 2021;12(1):5117.

[12] Nunes MCP, Bern C, Clark EH, Teixeira AL, Molina I. Clinical features of Chagas disease progression and severity. The Lancet Regional Health Americas 2024;37.

[13] Feng L, Shu S, Lin Z, Lv F, Li L, An B. Can cross entropy loss be robust to label noise? In International Joint Conferences on Artificial Intelligence. 2021; 2206–2212.

[14] Li J, Socher R, Hoi SC. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In International Conference on Learning Representations. 2020; .

Address for correspondence:

Sergio González
111 No. 166, Sec. 4, Chengde Rd., Shilin Dist., Taipei, Taiwan
gonzalez-vazquez.sergio@inventec.com