Detecting Chagas Disease from the ECG with Sharpness Aware Minimization and Domain Adversarial Learning

Jad Haidamous[1], Philip Hempel[2,3], Maurice Rohr[1], Tizian C Dege[1], Marcus Vollmer[3,4], Nicolai Spicher[3,5], Christoph Hoog Antink[1,6]

[1] KIS*MED - AI Systems in Medicine, Technical University of Darmstadt, Darmstadt, Germany
[2] Department of Medical Informatics, University Medical Center, Göttingen, Germany
[3] German Centre for Cardiovascular Research (DZHK), Berlin, Germany
[4] Institute of Bioinformatics, University Medicine Greifswald, Greifswald, Germany
[5] Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark
[6] Hessian Center for AI (hessian.AI), Darmstadt, Germany

Abstract

The detection of Chagas disease through serological testing is time-consuming and only available in limited quantities. To ensure adequate prioritization of patients for serological testing, this work aimed to develop an algorithm to detect signs of Chagas disease in electrocardiograms (ECG) as part of the 2025 George B. Moody PhysioNet Challenge. Due to the scarcity of serologically validated training data, using self-reported data with potentially incorrect labels is necessary.

We trained a multi-layer perceptron classifier on top of the features from a frozen ECGFounder model. Our training algorithm tackled three main challenges: Class imbalance, dataset bias, and noisy labels. First, we alleviated the class imbalance problem by using specialized ECG data augmentation methods and sharpness aware minimization. Then, we mitigated dataset bias by using domain adversarial learning to ensure our model learns dataset-invariant features. Subsequently, we recast the noisy labels problem within the framework of semi-supervised learning and addressed it using the FixMatch algorithm. Finally, we improved the performance of our algorithm on unseen datasets by employing test-time adaptation.

The proposed approach achieved a score of of 0.585 in five-fold cross-validation and 0.144 *(rank* $35/40$*, team DEbuggers) on the hidden test set.*

1. Introduction

The gold standard for the diagnosis of Chagas disease is a serological blood test, that must be specifically ordered and is often missed due to unspecific or absent symptoms. In contrast, ECGs are widely available, routinely recorded and can prompt this targeted serology and enable population-scale screening [1, 2]. Robust Chagas detection is challenging due to: **(i) class imbalance** (positives are scarce), **(ii) dataset bias** from regional demographics and device differences that can mislead models away from disease physiology, and **(iii) label noise**, since self-reported status and proxy rules can be incorrect and some seropositive patients may have ECGs labeled "normal", since they can possibly manage the infection without developing structural heart diseases.

For the 2025 George B. Moody PhysioNet Challenge [1–3], we built on the ECG foundation model *ECGFounder* [4] with three prespecified aims: **(i) mitigate class imbalance** using specialized ECG augmentations and sharpness-aware minimization (SAM) [5]; **(ii) reduce dataset bias** via domain-adversarial learning [6] to learn dataset-invariant features; and **(iii) leverage unlabeled/weakly labeled data** while being robust to label noise using FixMatch [7]; additionally, we applied test-time adaptation to address residual distribution shift [8].

2. Methods

Our approach combined a frozen pretrained large neural network, *ECGFounder* [4], with three lightweight fully connected neural networks – a feature extractor, dataset classifier, and Chagas classifier. The feature extractor was efficiently trained through the dataset and Chagas classifiers to map *ECGFounder*'s embeddings to dataset invariant features, specialized for Chagas disease detection. An overview of the complete model architecture is provided in Figure 1. In the following sections, we present a comprehensive description of each system component, the optimization procedure, and implementation details.

2.1. Dataset & Preprocessing

This challenge's training set consisted of three datasets: SaMi-Trop [9], CODE-15% [10], and PTB-XL [11]. The hidden test set contained samples from REDS-II [12], ELSA-Brasil [13], and a private dataset. Chagas disease may in some cases not result in abnormal ECG waveforms. Therefore, 286 patients' ECGs in the SaMi-Trop dataset were labeled as normal despite testing positive for Chagas disease. We labeled these patients as negatives and the rest as positives. We labeled all normal ECGs in CODE-15% as negatives and opted to ignore the self-reported Chagas labels, therefore treating the abnormal ECGs as unlabeled. Finally, we labeled every sample in PTB-XL as negative.

We pre-processed the raw ECG waveforms by first replacing NaN or infinite values with zero and resampling to 500 Hz through linear interpolation. We then applied a sixth order digital Butterworth bandpass filter with cut-off frequencies of 1 Hz and 30 Hz. We subsequently segmented the signal into non-overlapping segments of 5000 datapoints (10 seconds), with the exception of the last segment, which always contained the last 5000 samples of the signal. Segments shorter than 10 seconds were extended through symmetric zero-padding. The segmentation was not necessary for this challenge's training set, but was nevertheless included to deal with possibly longer signals in the hidden test datasets. Finally, we independently applied z-score standardization to each segment.

2.2. Augmentations

We applied data augmentations to the pre-processed ECG waveforms during training to deal with class imbalance [14]. Each waveform was sequentially augmented k-times with $k \sim \mathrm{U}\{1, 3\}$ and augmentations sampled from the set of all augmentations $\mathcal{A}$ without replacement. The set $\mathcal{A}$ contains the following augmentations: Adding zero-mean, scaled gaussian noise with an SNR $s \sim \mathrm{U}(10\mathrm{dB}, 40\mathrm{dB})$, randomly setting n leads to zero with $n \sim \mathrm{U}\{1, 3\}$, quantizing the signal to n bins with $n \sim \mathrm{U}\{256, 1024\}$, adding zero-mean, scaled baseline wander from the MIT-BIH noise stress test database [15] with an SNR $s \sim \mathrm{U}(10\mathrm{dB}, 40\mathrm{dB})$, and finally leaving the signal unchanged. The baseline wander noise had two channels, which were both independently included.

2.3. Feature Extractor

The feature extractor combined *ECGFounder*'s embeddings with the patients' age and sex. Sex was mapped to a 16 dimensional vector through a dictionary with three entries (male, female, other/unknown). Invalid age values were first replaced with -1 and age was then divided by 100 and mapped to a 16 dimensional vector through a fully connected layer. *ECGFounder*'s embeddings and the vectors were then concatenated to form a 1056 dimensional input vector, which was passed through two 1056 dimensional fully connected layers, each with an additional layer norm layer and GELU activation function. We refer to this combination of a fully connected layer, layer norm, and GELU as "fully connected block" in the rest of this work.

2.4. Domain Adversarial Learning

The feature extractor's output was fed into a dataset classifier, consisting of two fully connected blocks (512 and 128 dimensional respectively) with a final 2 dimensional fully connected layer. The dataset classifier was trained to estimate whether a sample originated from CODE-15% or PTB-XL. The dataset classifier was trained exclusively on negative samples, as the model should easily be able to classify positive Chagas cases as originating from CODE-15% due to the lack of Chagas cases in the PTB-XL dataset. The dataset classifier's gradient was reversed and multiplied with $\lambda = \frac{2}{1+e^{-10t}} - 1$ before back-propagating to the feature extractor [6], where t is the current training iteration divided by the maximum number of iterations.

2.5. Chagas Classifier

The feature extractor's output was also fed into a Chagas classifier, consisting of three fully connected blocks (1024, 512, and 128 dimensional respectively) with a final 6 dimensional fully connected layer. The Chagas classifier was a multi-input multi-output (MIMO) network with 3 heads, i.e. the network took in 3 different inputs concatenated into a 3168 dimensional vector during training and outputted a 6 dimensional vector containing their respective predictions [16]. At test-time, the input was repeated 3 times and the predictions were averaged. This resulted in an efficient implicit ensemble network [16].

2.6. FixMatch

We employed the FixMatch algorithm [7] to exploit the unlabeled CODE-15% samples. We denote the previous augmentation pipeline as a "weak augmentation" and define the following pipeline as a "strong augmentation" in accordance with the FixMatch algorithm: Each waveform is sequentially augmented k-times with $k \sim \mathrm{U}\{5, 7\}$ and augmentations sampled from the set of all strong augmentations $\tilde{\mathcal{A}}$ without replacement. The set $\tilde{\mathcal{A}}$ contains the following augmentations: Adding zero-mean, scaled gaussian noise with an SNR $s \sim \mathrm{U}(10\mathrm{dB}, 40\mathrm{dB})$, randomly setting n leads to zero with $n \sim \mathrm{U}\{3, 6\}$, quantizing the signal to n bins with $n \sim \mathrm{U}\{128, 256\}$, adding zero-mean, scaled baseline wander, muscle artifact, or electrode movement from the MIT-BIH noise stress test database with an

SNR $s \sim \mathrm{U}(5\mathrm{dB}, 10\mathrm{dB})$ to simulate strong and realistic signal corruptions [15]. The MIT-BIH noise signals had two channels each, which were independently included, resulting in the set $\tilde{\mathcal{A}}$ containing 9 augmentations. We used a confidence threshold of 0.95 and an unsupervised batch size of 32, equal to the supervised batch size.

2.7. Optimization

We optimized the neural networks with SAM [5] with $\rho = 0.1$, as it has been shown to help with class imbalance and label noise [5, 14]. The base optimizer was AdamW with a learning rate of $5 \cdot 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay of 0.01. The loss function was an unweighted addition of three losses: A standard cross-entropy loss between the labeled samples and their Chagas predictions, the FixMatch unsupervised loss [7], and the domain adversarial loss [6]. We employed a weighted sampler for the labeled batches, which drew samples from SaMi-Trop, CODE-15%, and PTB-XL with equal probability. The models were trained for 20 epochs. Freezing the *ECGFounder* weights improved the optimization's efficiency, at the cost of limiting the adaptability to the training datasets. This design choice was the main difference between our approach and others in the challenge.

2.8. Test-Time Adaptation

We finally improved the performance of the Chagas classifier on distribution-shifted datasets with the entropy minimization method COME [8]. In particular, we minimized the entropy of the MIMO ensemble's mean prediction, effectively not only minimizing the entropy of each individual prediction, but also minimizing their disagreement. We performed one optimization step per sample with the AdamW optimizer with a learning rate of $3.125 \cdot 10^{-5}$ and did not filter samples based on entropy. If a sample consisted of multiple segments, the prediction with the highest Chagas probability was taken.

3. Results

We validated our approach with stratified five-fold cross validation using the labeled subset. The unlabeled CODE-15% subset was included in every training fold and used by the FixMatch algorithm. We report the mean challenge and F_1 score for our algorithm and some ablations in Table 1. We furthermore report the five-fold cross validation, hidden validation, and hidden test set challenge scores for our chosen entry in Table 2. The chosen entry of our team, *DEbuggers*, achieved a challenge score of 0.144 on the hidden test set, ranking $35/40$.

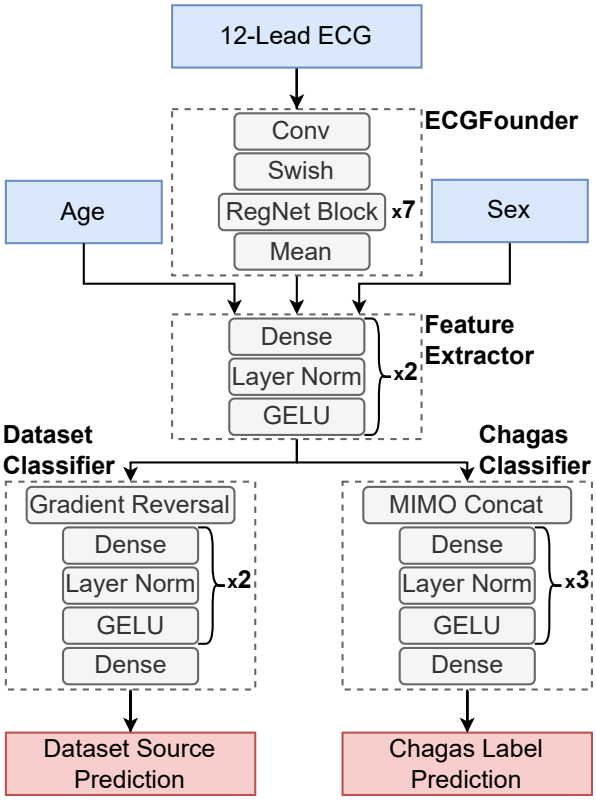

Figure 1. Complete model architecture. The color blue refers to inputs, red to outputs, and gray to model layers. We refer to Li et al. [4] for the RegNet block's design.

4. Discussion and Conclusions

The challenge and F_1 scores in Table 1 showed conflicting results. We observed that the local challenge score overestimated the performance of our approach, while the F_1 score was a better measure of our progress. We therefore based our analysis on the F_1 scores in Table 1. The ablations of our algorithm showed that SAM [5, 14] and MIMO [16] resulted in the largest improvements. Cross-validation under-estimated the influence of domain adversarial learning [6] and data augmentations, as they improved generalization and therefore prevented overfitting on the local dataset. Our approach performed reasonably well on the hidden validation set, being 0.108 behind the best entry, but heavily deteriorated on the hidden test set. We hypothesize that freezing the *ECGFounder* weights hindered the adaptation of our model to Brazilian ECG datasets. COME failed to mitigate this distribution shift, as our Hackathon entry with the same algorithm without COME performed slightly better with a score of 0.150 [8]. Further improvements may therefore be achieved by unfreezing the *ECGFounder* weights and training on more unlabeled data, e.g. by including the full CODE dataset.

Algorithm	Score	F_1
ECGFounder + MLP Classifier	0.712	0.511
+ Augmentations	0.731	0.471
+ Domain adversarial learning	0.765	0.285
+ Sharpness aware minimization	0.762	0.505
+ MIMO classifier	0.730	0.574
+ FixMatch	0.744	0.567
+ COME = Final algorithm	0.585	0.440

Table 1. Stratified five-fold cross validation results for our algorithm and its ablations. The mean Challenge and F_1 scores are reported.

Training	Validation	Test	Ranking
0.585 ± 0.077	0.350	0.144	35/40

Table 2. Challenge scores for our selected entry (team DEbuggers), including the ranking of our team on the hidden test set. We report five-fold cross validation results with one standard deviation on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

Acknowledgments

This work was funded by the European Union's Horizon Europe programme under the Grant Agreement no. 101137278.

References

[1] 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.

[2] 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. DOI 10.48550/arXiv.2510.02202.

[3] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[4] Li J, Aguirre A, Moura J, Liu C, Zhong L, Sun C, Clifford G, Westover B, Hong S. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. arXiv preprint arXiv241004133v4 2024.

[5] Foret P, Kleiner A, Mobahi H, Neyshabur B. Sharpness-aware minimization for efficiently improving generalization. In International Conference on Learning Representations. 2021.

[6] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V. Domain-adversarial training of neural networks. Journal of Machine Learning Research 2016;17(59):1–35.

[7] Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, Cubuk ED, Kurakin A, Li CL. FixMatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems 2020; 33:596–608.

[8] Zhang Q, Bian Y, Kong X, Zhao P, Zhang C. COME: Test-time adaption by conservatively minimizing entropy. In International Conference on Learning Representations. 2025.

[9] Cardoso C, Sabino E, Oliveira C, de Oliveira L, Ferreira A, Cunha-Neto E, Bierrenbach A, Ferreira J, Haikal D, Reingold A, Ribeiro A. Longitudinal study of patients with chronic chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. BMJ Open 2016;6(5):e0011181.

[10] Ribeiro A, Ribeiro M, Paixão G, Oliveira D, Gomes P, Canazart J, Ferreira M, Andersson C, Macfarlane P, Meira WJ, Schön T, Ribeiro A. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications 2020;11(1):1760.

[11] 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; 7:154.

[12] Nunes M, Buss L, Silva J, Martins L, Oliveira C, Cardoso CS BB, Ferreira A, Oliveira L, Bierrenbach A, Fernandes F, Busch M, Hotta V, Martinelli L, Soeiro M, Brentegani A, Salemi V, Menezes M, Ribeiro A, Sabino E. Incidence and predictors of progression to chagas cardiomyopathy: Long-term follow-up of trypanosoma cruzi-seropositive individuals. Circulation 2021;144(19):1553–1566.

[13] Pinto-Filho M, Brant L, Dos Reis R, Giatti L, Duncan B, Lotufo P, da Fonseca M, Mill J, de Almeida M, MacFarlane P, Barreto S, Ribeiro A. Prognostic value of electrocardiographic abnormalities in adults from the Brazilian longitudinal study of adults' health. Heart 2021;107(19):1560–1566.

[14] Shwartz-Ziv R, Goldblum M, Li Y, Bruss CB, Wilson AG. Simplifying neural network training under class imbalance. Advances in Neural Information Processing Systems 2023; 36:35218–35245.

[15] Moody G, Muldrow W, Mark R. The MIT-BIH noise stress test database. Computers in Cardiology 1984;11:381–384.

[16] Havasi M, Jenatton R, Fort S, Liu JZ, Snoek J, Lakshminarayanan B, Dai AM, Tran D. Training independent subnetworks for robust prediction. In International Conference on Learning Representations. 2021.

Address for correspondence:

Jad Haidamous
Merckstrasse 25, 64283 Darmstadt
haidamous@kismed.tu-darmstadt.de