Biomarker-Based Pretraining for Chagas Disease Screening in Electrocardiograms

Elias Stenhede[1,2], Arian Ranjbar[1,2]

[1] Medical Technology & E-health, Akershus University Hospital, Lørenskog, Norway
[2] Faculty of Medicine, University of Oslo, Oslo, Norway

Abstract

Data-driven Chagas disease screening via electrocardiograms (ECGs) is limited by scarce and noisy labels in existing datasets. We propose a biomarker-based pretraining approach, where an ECG feature extractor is first trained to predict percentile-binned blood biomarkers from the MIMIC-IV-ECG dataset, using a bin-smoothing regularization to handle the sparsity introduced by binning. The pretrained model is then fine-tuned on Brazilian datasets (CODE15% and SaMi-Trop) for Chagas detection. Our 5-model ensemble, developed by the Ahus AIM team, achieved a challenge score of 0.412 on the hidden validation set, ranking 5/66 in Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025. Source code and model weights are shared on GitHub: https://github.com/Ahus-AIM/physionet-challenge-2025.

1. Introduction

This work was developed as part of the George B. Moody PhysioNet Challenge 2025, which aims to advance automated ECG-based screening methods for Chagas disease.

Chagas disease, caused by the parasite Trypanosoma cruzi, remains a significant health burden in Central and South America [1]. If left untreated, the infection can result in potentially life-threatening cardiac, digestive, and neurological complications. While treatments exist to slow cardiovascular damage, serological testing to detect the parasite is inaccessible for much of the at-risk population. In this context, the use of AI-interpreted electrocardiograms (ECGs) is a promising, resource-efficient option for large-scale screening.

Openly accessible ECG databases have been recorded in regions affected by Chagas disease [2, 3], but annotation validity is limited, reducing the effectiveness of traditional supervised deep learning approaches.

2. Methods

The following sections describe data sources, preprocessing steps, biomarker-based pretraining strategy, subsequent fine-tuning, and the model architecture used for the challenge.

2.1. Data sources

The MIMIC-IV-ECG [4] is collected across Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA, spanning the period 2008 to 2019. The ECGs in this dataset can be connected to biomarkers by connecting them with the MIMIC-IV dataset [5]. Biomarker usage in clinical practice is highly skewed, with most tests used infrequently and a few used routinely. To limit the number of tests, selection was based on both prevalence and clinical relevance. The tests used in pretraining are presented in Table 1. The datasets used for fine-tuning are both collected in Brazil, where Chagas disease is endemic. The CODE15% dataset is collected by Telehealth Network of Minas Gerais in the period 2010 to 2016, and is paired with self-reported Chagas labels. In contrast, the SaMi-Trop dataset only contains patients with chronic Chagas cardiomyopathy. All utilized datasets are presented in Table 2.

2.2. Preprocessing steps

Preprocessing involves standardizing ECG signals, resolving label inconsistencies, and normalizing biomarker values for pretraining. Details for each component are provided in the subsections below.

2.2.1. Electrocardiograms

Across all datasets, ECGs are resampled to 400 Hz, followed by dropping leads I, II, III, and aVR, as they are linear combinations of aVL and aVF defined by Einthoven's Law. During training, a single two-second snippet is randomly extracted from each ECG. All snippets are normal-

Table 1. Biomarkers included for model pretraining (alphabetical).

Biomarker	Clinical domain
Albumin	Liver function
Calcium, Total	Electrolytes
Creatinine	Renal
Hematocrit	Hematology
Hemoglobin	Hematology
INR(PT)	Coagulation
NTproBNP	Cardiac
Potassium	Electrolytes
Red Blood Cells	Hematology
Troponin T	Cardiac
Urea Nitrogen	Renal

Table 2. Datasets used for model development. For the pretraining dataset, not all ECGs were included, as each ECG had to occur within 24 hours of a blood test.

Dataset	# ECGs	# Patients	Step
MIMIC-IV-ECG	523,275	102,511	Pretraining
CODE15%	345,779	233,770	Fine-tuning
SaMi-Trop	1,959	1,959	Fine-tuning

ized to have zero mean and unit variance before being fed into the model.

2.2.2. Chagas labels

In the CODE15% dataset, a total of 1,825 patients have ECGs labelled both as Chagas-positive and negative. In an attempt to reduce label noise, all ECGs for these patients are labelled using the average positive proportion of ECGs. The aim is to reduce noise, without discarding these patients, to preserve sample size while minimizing label uncertainty.

2.2.3. Biomarkers

Normalizing biomarker values is challenging due to differing distributions and the need to preserve clinically relevant cut-offs. We address this by replacing each value with its percentile rank (binned into 100 intervals). The label sparsity introduced by increasing the number of bins is further discussed in Section 2.3.1. Biomarker results are matched to ECGs by timestamp; ECGs without a paired result within 24 hours are excluded from pretraining.

2.3. Training strategy

All model training is performed on the setup detailed in Table 3. Across all model training, the Muon optimizer is used for parameters with dimension $\geq$2, as it has proved to converge faster and yield better results [6]. For the remaining parameters, i.e., biases and parameters in normalization layers, Adam is used [7]. The same learning rate is used for both optimizers. For clarity, the pretraining and fine-tuning training recipes are described one at a time.

Table 3. Development environments and hardware.

Component	Specification
System	Debian 12
CPU	Intel Core i9-14900KF
RAM	2×48 GB; 4800 MT/s
GPU	NVIDIA GeForce RTX 5090
CUDA version	12.9
Programming language	Python 3.12
Deep learning framework	PyTorch 2.7.1+cu128

2.3.1. Biomarker based pretraining

The pretraining task is formulated as a classification problem for each biomarker, where the target is the percentile rank of the test result. Most ECGs are not paired with all selected biomarkers; in these cases, the loss is not computed for those missing biomarker-ECG pairs. Using classification instead of regression enables the model to predict a probability distribution over possible blood test values, which provides richer supervision and reduces bias toward the mean value of each biomarker. Specifically, the model outputs logits with shape $\mathbb{R}^{\text{batch\_size}\times 100\times \mathrm{T}}$ with T denoting the number of tests. Cross-entropy loss is computed over the 100 percentile bins for each available biomarker. The parameters used in pretraining are listed in Table 4.

Table 4. Parameters used during biomarker-based pretraining.

Parameter	Value
Batch size	64
Optimizer	Muon and Adam
Muon momentum	0.95
Learning rate	0.0037
Loss function	Cross-entropy

To mitigate the label sparsity introduced by percentile binning, we use a decoupled regularization step applied af-

ter each optimizer step. In essence, we enforce the inductive bias of neighbouring bins sharing directions in weight space. We treat the final fully-connected layer's weight as

$$W \in \mathbb{R}^{(100\times \mathrm{T})\times \mathrm{F}}$$

where F is the feature dimension of the activations just before the last layer and T the number of tests. After the usual gradient-based update, we apply an in-place smoothing update with $\alpha = \eta\,\beta$, where η is the learning rate and β a fixed bin-smoothing factor. For each weight row W_i, we then perform

$$W_i \leftarrow \begin{cases} (1-\frac{\alpha}{2})\,W_1 + \frac{\alpha}{2}\,W_2, & i = 1, \\ (1-\alpha)\,W_i + \frac{\alpha}{2}\,(W_{i-1} + W_{i+1}), & 2 \le i \le 99, \\ (1-\frac{\alpha}{2})\,W_{100} + \frac{\alpha}{2}\,W_{99}, & i = 100. \end{cases}$$

Because this smoothing runs after back-propagation and does not contribute to any gradients, it adds negligible overhead. The bin-smoothing factor β is set to 1 in this work, and it can be optimized as a hyperparameter.

2.3.2. Fine-tuning for Chagas screening

The fine-tuning step starts by initializing a feature extractor with the weights obtained in the pretraining step, and dropping the last linear layer mapping to biomarkers, instead replacing it with a new randomly initialized layer that will be trained to output logits corresponding to the probability of Chagas disease. Although the CODE15% and SaMi-Trop datasets contain both self-reported and strong labels, they are pooled. To ensure good generalization, 5-fold cross-validation is used to train 5 models, with each model being selected at the epoch where the validation loss is lowest. Parameters used in this step are detailed in Table 5.

Table 5. Parameters used during fine-tuning.

Parameter	Value
Batch size	128
Optimizer	Muon and Adam
Muon momentum	0.95
Learning rate	0.001
Loss function	Binary cross-entropy

2.4. Model architecture

The model is based on the InceptionTime architecture [8], with minor modifications. Before the inception blocks, we include two convolutional layers followed by batch normalization and GELU activation. Each convolution uses a kernel size of 5 and a stride of 2, which ensures that the receptive field within the network is sufficiently large. The parameters for the InceptionTime network are summarized in Table 6.

Table 6. Parameters for the InceptionTime network.

Parameter	Value
Number of blocks	6
Kernel sizes	9, 19, 39
Number of filters	32
Bottleneck channels	32

2.5. Final ensemble

During inference, each ECG is divided into ten overlapping two-second segments, which are individually standardized and processed by the ensemble models. The logits produced by each model are transformed with a sigmoid function, and the resulting probabilities are averaged across all segments and ensemble members.

3. Results

The validation loss for biomarker-based pretraining reached its minimum after three epochs. When ranking biomarkers by validation-set perplexity, the model predicted NT-proBNP most accurately, followed by albumin, hemoglobin, and troponin. Predicted probability distributions for selected biomarkers are illustrated in Figure 1.

The fine-tuned model was evaluated using the official challenge metric, defined as the number of Chagas-positive cases in the top 5% of ECGs sorted by model-predicted risk, divided by the total number of cases. When evaluating the model using 5-fold cross-validation on the development set, a mean challenge score of 0.439 (SD = 0.010) and an AUC-ROC of 0.840 (SD = 0.008) were achieved. When deploying the ensemble on the hidden validation set, the ensemble achieved a challenge score of 0.412, resulting in a ranking of red 5/66, also displayed in Table 7. Notably, larger models resulted in better cross-validated scores on the development set, but not on the hidden validation set.

Table 7. Challenge score on the hidden validation set.

Team	Challenge score	Rank
Ahus AIM	0.412	5/66

4. Discussion

While traditional approaches to label noise include partial self-supervision, label correction techniques, early

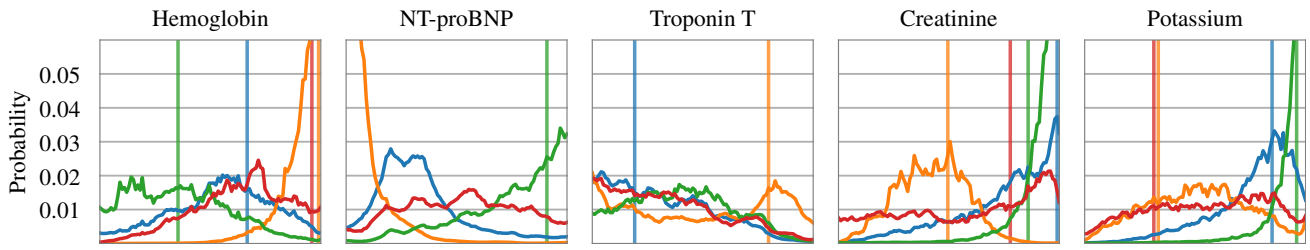

Figure 1. Shows the predicted probability distributions over blood test percentiles for four ECGs taken from four different patients (one colour per patient). The vertical lines represent the actual values for the same patients, measured within 24 hours of ECG recording. The selected biomarkers represent a subset of the biomarkers used in pretraining. As the probability distributions are defined over percentile bins, no units are displayed on the x-axis.

stopping, or robust loss functions, our method leverages clinical biomarkers to guide the model in learning physiologically relevant ECG features before fine-tuning for Chagas disease detection. One notable aspect of our study is the geographical and clinical diversity of the datasets. Pretraining was conducted on datasets collected in the USA, where Chagas disease is rare, whereas fine-tuning was performed on Brazilian datasets where the disease is endemic. This split underscores the importance of evaluating domain shift and raises questions regarding the generalizability of features learned from one population to another. The degree to which biomarker-driven pretraining can yield transferable and robust ECG representations across diverse settings remains an area for further study.

Future work should assess the transferability of this approach to other diseases and evaluate the impact of varying the set of biomarkers, bin sizes, and pretraining datasets. Additionally, systematic comparisons with other label noise mitigation strategies would help clarify the relative strengths and weaknesses of biomarker-based pretraining. This has not yet been done and is thus a limitation of the current study.

5. Conclusion

We present a biomarker-based pretraining approach for ECG-based Chagas disease screening, motivated by the need for improved feature extraction in settings with limited or noisy disease labels. We believe the same framework can be applied to other diseases, and it is especially beneficial in cases where labels are few or of limited validity.

6. Acknowledgements

We thank Akershus University Hospital for the funding that made this work possible.

References

[1] Cucunubá ZM, Gutiérrez-Romero SA, Ramírez JD, Velásquez-Ortiz N, Ceccarelli S, Parra-Henao G, Henao-Martínez AF, Rabinovich J, Basáñez MG, Nouvellet P, Abad-Franch F. The epidemiology of Chagas disease in the Americas. The Lancet Regional Health Americas September 2024;37:100881. ISSN 2667193X.

[2] Cardoso CS, Sabino EC, Oliveira CDL, De Oliveira LC, Ferreira AM, Cunha-Neto E, Bierrenbach AL, Ferreira JE, Haikal DS, Reingold AL, Ribeiro ALP. Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. BMJ Open May 2016; 6(5):e011181. ISSN 2044-6055, 2044-6055.

[3] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MPS, Andersson CR, Macfarlane PW, Meira Jr. W, Schön TB, Ribeiro ALP. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications April 2020;11(1):1760. ISSN 2041-1723.

[4] Gow B, Pollard T, Nathanson LA, Johnson A, Moody B, Fernandes C, Greenbaum N, Waks JW, Eslami P, Carbonati T, Chaudhari A, Herbst E, Moukheiber D, Berkowitz S, Mark R, Horng S. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset.

[5] Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV.

[6] Muon: An optimizer for hidden layers in neural networks | Keller Jordan blog.

[7] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization, January 2017. ArXiv:1412.6980.

[8] Fawaz HI, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, Webb GI, Idoumghar L, Muller PA, Petitjean F. InceptionTime: Finding AlexNet for Time Series Classification, December 2020. ArXiv:1909.04939.

Address for correspondence:

Elias Stenhede
Sykehusveien 25, 1478 Nordbyhagen, Norway
elias.stenhede@ahus.no