Fine-Tuning a Pretrained ECG Foundation Model for Chagas Disease Detection

Yongchao Long[1,2,*], Jinshuai Gu[1,2,*], Mingke Yan[2,*], Rafael da Costa Silva[7], Deyun Zhang[3], Shijia Geng[3], Jun Li[2], Qinghao Zhao[4], Diego Furtado Silva[7,†] Yuxi Zhou[1,†], Shenda Hong[2,5,6]

[1]Department of Computer Science, Tianjin University of Technology, Tianjin, China
[2]National Institute of Health Data Science, Peking University, Beijing, China
[3]HeartVoice Medical Technology, Hefei, China
[4]Department of Cardiology, Peking University People's Hospital, Beijing, China
[5]Institute of Medical Technology, Health Science Center of Peking University, Beijing, China
[6]Institute for Artificial Intelligence, Peking University, Beijing, China
[7]Computer Science Department, University of São Paulo, São Carlos, Brazil

Abstract

While large-scale ECG collections are common for prevalent diseases, high-quality, strongly-labeled data for rare conditions such as Chagas disease remains extremely scarce, presenting significant challenges for artificial intelligence diagnostic systems. As part of the George B. Moody PhysioNet Challenge 2025, team ChagasExplorers proposes a transfer-learning methodology built on the pre-trained ECGFounder model. Deep features are extracted from a fully frozen foundation model, fused with demographic information, and fed to a lightweight multilayer perceptron for classification. Our method achieved an official score of 0.108 on the test set (rank 38/41). This paper details the methodology, summarizes performance across challenge stages, and provides visual insights into the model's decision behaviour.

1. Introduction

Chagas disease, also known as American trypanosomiasis, is a neglected tropical disease caused by the protozoan parasite *Trypanosoma cruzi* and has long been overlooked by the World Health Organization. The disease is primarily endemic to Latin America, where approximately six million people are estimated to be infected. However, with increasing global migration, its prevalence has gradually extended to non-endemic regions including North America and Europe. Chagas disease poses a serious threat to public health, with its primary cardiac complication, Chagas cardiomyopathy (CCM), representing the leading cause of death among infected individuals [1].

*Contributed equally: Mingke Yan, Jinshuai Gu and Yongchao Long.
†Corresponding authors: Yuxi Zhou, Diego Furtado Silva

In this context, the ECG has emerged as an ideal tool for large-scale screening of cardiac abnormalities and early identification of high-risk individuals due to its low cost, non-invasive nature, and ease of deployment. However, the development of AI models for this purpose faces significant data challenges, a situation reflected in the training data provided for The George B. Moody PhysioNet Challenge 2025 [2, 3].

To address these challenges, we propose an effective transfer learning strategy based on the pre-trained ECG foundation model, ECGFounder [4]. The main contribution of this paper is the proposal and validation of an efficient "feature extraction + lightweight classification" pipeline. This approach demonstrates the potential of leveraging large-scale foundation models for knowledge transfer in processing ECG data for rare diseases.

2. Methods

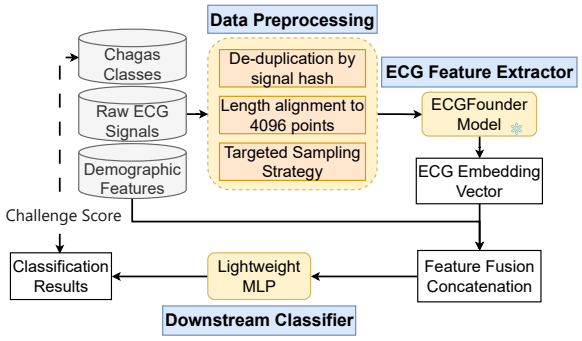

Figure 1. Schematic diagram of our proposed pipeline.

The overall framework of our proposed method is illustrated in Figure 1. Given the limited number of strongly-

labeled samples and the significant noise within the training data, we designed a two-stage modular pipeline to mitigate the risk of overfitting that could arise from end-to-end training of a large model. This pipeline first utilizes a large-scale pre-trained foundation model to convert raw ECG waveform signals into information-rich feature vectors. Subsequently, these features are fused with patient demographic information. Finally, a lightweight machine learning model performs the definitive diagnostic classification.

2.1. Datasets and Preprocessing

Our research data was sourced from the official challenge dataset, which includes records from multiple sources such as CODE-15% [5], SaMi-Trop [6], and PTB-XL [7]. During data loading, we first performed **de-duplication** on all records based on a hash of the signal content to eliminate redundant data. For constructing the training and validation sets, we employed a specific **data sampling recipe**, precisely extracting a specified number of positive and negative samples from each data source to control the composition and distribution of the training data.

To enable the model to process these heterogeneous ECG signals, we designed a standardized preprocessing workflow. To ensure all signals input to the model had uniform dimensions, we implemented a length alignment step, standardizing the length of each 12-lead signal to 4096 sampling points. Shorter signals were zero-padded at the end, while longer signals were centrally cropped.

2.2. ECG Feature Extractor (ECGFounder)

We selected ECGFounder as the core ECG feature extraction module [4]. ECGFounder is a foundation model pre-trained on over ten million ECG recordings, with a core architecture based on the Net1D network and a RegNet design. Our key strategy in this task was to keep its parameters **fully frozen**, preventing it from participating in gradient updates during training. It was used solely as a powerful feature extraction engine, converting the preprocessed ECG signals into high-dimensional embedding vectors.

2.3. Multimodal Feature Fusion

To integrate information from different modalities, we directly **concatenated** the ECG embedding vectors extracted from ECGFounder with the demographic features (age and sex) obtained from patient metadata at the feature level. This created a comprehensive multimodal feature vector containing both physiological information from the ECG and background information about the patient.

2.4. Downstream Classifier (MLP)

We chose a lightweight **Multi-Layer Perceptron (MLP)** as the final classifier. The input dimension of the MLP matches that of the fused multimodal feature vector. It contains two hidden layers and uses Dropout as a regularization technique to prevent overfitting. The MLP was selected for its simple architecture, fast training speed, and suitability for processing information-dense feature vectors that have already been extracted by a large model.

2.5. Training Details and Evaluation

We trained only the downstream MLP classifier. To address the class imbalance, where positive samples are far fewer than negative ones, we used a **Binary Cross-Entropy with Logits Loss** function with weights for the positive class. For the optimizer, we selected AdamW and employed a learning rate scheduler to dynamically adjust the learning rate.

To further enhance model stability and prevent overfitting, we implemented two regularization strategies in the training loop: **Gradient Clipping**, which limits the norm of the gradients to a maximum value of 1.0, and an **Early Stopping** strategy based on the validation set loss. Training would terminate prematurely if the validation loss did not improve within a pre-set number of patience epochs, and the best-performing model state was saved.

The model's performance was evaluated using the official scoring metric of the competition: the **Challenge Score**, defined as the proportion of true positive cases (TPR) found within the top 5% of patients with the highest predicted probabilities. Additionally, we calculated traditional metrics such as AUROC and AUPRC for supplementary assessment.

3. Results

3.1. Model Performance Evaluation

We first conducted a series of ablation studies using cross-validation on the training set to determine the optimal model strategy. As shown in Table 1, our final adopted solution, the "ECGFounder Features + MLP" strategy, achieved the best cross-validation performance, reaching a Challenge Score of 0.353 and an AUROC of 0.812. Note that these cross-validation results are not comparable to official challenge scores.

The model showed strong performance during development but declined substantially on the test set, achieving an official score of 0.108 (rank 38/41).

Table 1. Ablation using cross-validation on training set: *Z*=frozen params, *F*=fine-tuned; Recipe=C (Code-15%), S (Sami-Trop), P (PTB-XL).

Model	Recipe	Chal. Score	AUROC	AUPRC
Official Baseline	C+P+S	0.075	0.733	0.043
ECGFounder(Z)	P+S	0.119	0.792	0.138
ECGFounder(F)	C+S	0.241	0.523	**0.180**
ECGFounder(Z)	C+S	**0.353**	**0.812**	0.142

3.2. Visualization of Model Decision Behavior

To objectively present the model's decision-making patterns on the test set, we performed t-SNE dimensionality reduction on its output ECG features for visualization. Figure 2 displays the distribution of the test set samples in the feature space. In the plot, the color represents the predicted probability assigned by the model (yellower indicates a higher probability), while the shape denotes the ground truth label (circles for negative, crosses for positive). From the figure, it can be observed that the model tends to assign higher positive probabilities to samples located in a specific region of the feature space (the lower-left area).

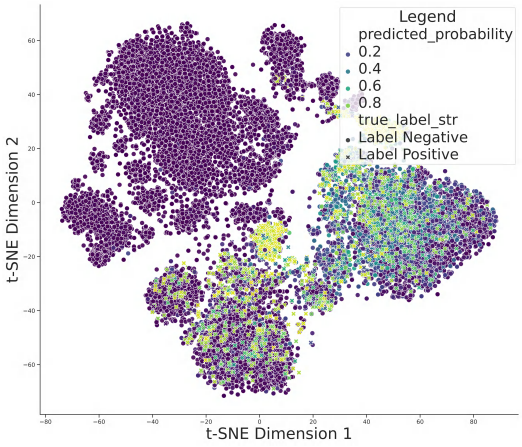

Figure 2. t-SNE feature visualization of the test set. The colors in the figure represent the model prediction probabilities, and the shapes represent the actual labels.

4. Discussions

The core finding of this study is the explanation for the drastic performance difference between development and test stages. This discrepancy does not stem from a flaw in the model itself but rather from a "data shortcut" problem caused by systematic bias within the training data.

First, we performed t-SNE dimensionality reduction on a mixture of training and validation samples and colored them according to their ground truth labels (Figure 3). As the figure clearly shows, the positive samples (red) and negative samples (blue) form two relatively independent and separable clusters in the feature space. On the surface, this appears to be an ideal classification scenario.

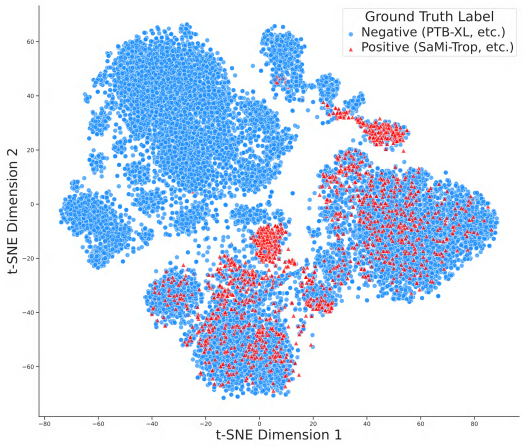

Figure 3. Distribution of true labels of training and validation set samples in the t-SNE feature space. Positive (red) and negative (blue) samples form clearly separable clusters.

However, the true nature of the problem is revealed when we color these same samples by their data source (Figure 4). The figure reveals that the large blue cluster of negative samples seen in Figure 2 is composed almost entirely of the PTB-XL dataset from Europe (green squares). Meanwhile, the red cluster of positive samples from Figure 2 spatially overlaps with all other datasets from Brazil (such as the blue dots from Code15 and the red triangles from SaMi-Trop).

This phenomenon provides a definitive explanation for the model's performance at different stages. The model did not learn the intrinsic biomarkers of Chagas disease; instead, it discovered a much simpler classification rule: to distinguish between ECGs from Europe and ECGs from Brazil. During development, this shortcut based on data origin was effective, as negative samples were predominantly from PTB-XL and all positive samples were from Brazil, which was reflected in the strong cross-validation results on the training set.

Figure 2 further confirms this behavior. The regions where the model assigned high prediction probabilities (yellow and green dots) precisely cover the feature space occupied by the Brazilian datasets in Figure 4. In essence, the model was performing a "geographical classification" task rather than a "disease diagnosis" task. When the model encountered the test set, which likely did not contain this pronounced source-based discrepancy, the previously learned "shortcut" failed completely, causing a sharp drop

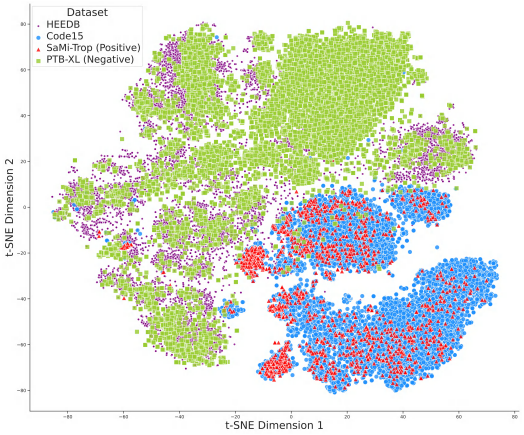

Figure 4. Distribution of different data sources in the t-SNE feature space. PTB-XL negative samples from Europe (green) form a completely separate cluster from all samples from Brazil (positive and negative).

in performance.

The case presented in this study clearly demonstrates that when dealing with heterogeneous data from multiple centers in the medical AI field, the issue of latent "data shortcuts" poses a fatal threat to a model's generalization capabilities. This serves as a warning that relying solely on a single, similarly distributed hidden validation set is far from sufficient in model development and evaluation. It is imperative to enhance the transparency of model decision-making and conduct robust testing for resilience against data distribution shifts.

5. Conclusion

This study proposed and validated a transfer learning method based on the pre-trained ECG foundation model, ECGFounder, for Chagas disease ECG screening. The method achieved an official score of 0.108 on the test set (rank 38/41). Through an in-depth visual analysis of the model's decision behavior and the dataset's distribution, we revealed that a significant discrepancy in data distribution was the key reason for the performance difference between development and test stages. The model exploited a "shortcut" based on data origin rather than learning the disease itself, resulting in severely insufficient generalization capabilities. This finding holds significant cautionary implications for AI modeling with multi-center, heterogeneous medical data. Since time was limited, our exploration of applying the ECG foundation model to Chagas disease is far from exhaustive. Future work could not only continue this exploration but also incorporate more advanced domain adaptation techniques and weak-label learning strategies to further enhance the model's generalization and robustness on complex real-world data.

References

[1] Sabino EC, Carmo M, Blum J, Molina I, Luiz A. Cardiac involvement in chagas disease and african trypanosomiasis. Nature Reviews Cardiology jul 2024;.

[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] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, physiotoolkit, and physionet. Circulation 2000;101(23):e215–e220. URL https://www.ahajournals.org/doi/abs/10.1161/01.CIR.101.23.e215.

[4] Li J, et al. An electrocardiogram foundation model built on over 10 million recordings. NEJM AI jun 2025;2(7).

[5] Ribeiro AH, Paixao GMM, Lima EM, Horta Ribeiro M, Pinto Filho MM, Gomes PR, et al. CODE-15%: a large scale annotated dataset of 12-lead ECGs (1.0.0), 2021. [Data set].

[6] Ribeiro ALP, Ribeiro AH, Paixao GMM, Lima EM, Horta Ribeiro M, Pinto Filho MM, et al. Sami-Trop: 12-lead ECG traces with age and mortality annotations (1.0.0), 2021. [Data set].

[7] Wagner P, Strodthoff N, Bousseljot R, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3), 2022. RRID:SCR_007345.

Address for correspondence:

Yuxi Zhou
Department of Computer Science, Tianjin University of Technology, Tianjin, China
joy_yuxi@pku.edu.cn