Foundation Model–Driven High-Confidence Electrocardiogram-Based Chagas Disease Detection

Pindong Chen[1], Bo Yu[1], Wenlong Wu[1]

[1]Xiaomi Health Lab, Beijing, China

Abstract

As part of the George B. Moody PhysioNet Challenge 2025, team MIWEAR developed an approach for detecting Chagas disease from 12-lead electrocardiograms (ECGs). Using ResNet-18 with five-fold cross-validation, we estimated sample confidence and curated a subset by retaining only high-confidence positives and negatives. A large-scale ECG foundation model pretrained on over ten million recordings was then fine-tuned, alongside EfficientNet-B0 and ResNet-18 trained on the curated data. Model predictions were fused by averaging. Cross-validation confirmed that confidence-based sampling improved the performance. The standalone ECG foundation model achieved a Challenge score of 0.379 on the hidden validation set, ranking 7th, underscoring strong transferability under distribution shifts. A fusion model guided by the foundation model reached the highest score of 0.400 on the training set, demonstrating the value of integrating complementary architectures to boost accuracy and reduce variance. These findings show that foundation models provide a reliable backbone, while fusion enhances stability, offering a competitive strategy for ECG-based Chagas disease detection.

1. Introduction

We participated in the 2025 George B. Moody PhysioNet Challenge, which invited teams to develop automated, open-source algorithms for identifying Chagas disease from electrocardiograms (ECGs) [1, 2]. While serological testing is the gold standard for diagnosis, ECG-based interpretation provides a scalable and cost-effective screening alternative, particularly in resource-constrained settings.

The availability of large-scale public ECG databases, including CODE-15, SaMi-Trop, PTB-XL, REDS-II, and ELSA-Brasil [3–7], has enabled the development of data-driven approaches for this task. However, these datasets differ substantially in labeling quality, class balance, and demographic coverage. In particular, weakly labeled samples from large cohorts pose challenges for effective model training, as naive use of these data may amplify label noise.

Our team, MIWEAR, designed an approach that combines confidence-guided sample selection with deep neural networks to address these issues. Instead of discarding weakly labeled data entirely, we sought to extract reliable subsets by leveraging prediction confidence. This strategy allowed us to mitigate noise while still benefiting from the scale of large databases. We then integrated pre-trained ECG foundation models with conventional deep architectures, and employed ensembling to enhance predictive robustness.

In this paper, we describe our methodology in detail, present results from cross-validation and hidden validation evaluation, and discuss the advantages and limitations of our approach in the context of Chagas disease detection.

2. Methods

2.1. Data Preprocessing

All recordings from the training databases were first parsed into a unified metadata table containing the recording length, source, age, sex, and diagnostic label. To ensure data quality, we excluded samples shorter than 2900 samples (corresponding to approximately 7.25 s at 400 Hz).

Because the CODE-15% subset was both substantially larger than the other datasets and contained weaker labels, we applied random undersampling to reduce its prevalence in the training pool. Specifically, a fixed fraction of CODE-15% samples was retained, while all samples from the other databases were preserved. This step aimed to alleviate dataset imbalance and reduce the influence of noisy or uncertain labels.

For demographic attributes, we mapped sex into binary form (male = 1, female = 0) and retained patient age as a continuous feature. When demographic information was missing or ambiguous, we set its value to a missing indicator rather than discarding the record, in order to maximize data usage.

Each electrocardiogram (ECG) signal was then standardized into a 12-lead format (I, II, III, aVR, aVL, aVF, V1–V6). We reordered channels accordingly and discarded non-standard leads. To mitigate baseline wander and high-frequency artifacts, we applied median filtering to each lead. Signals were subsequently resampled to 400 Hz to unify sampling frequency across databases.

After resampling, amplitudes were normalized on a per-lead basis using min–max scaling,

$$x' = \frac{x - \min(x)}{\max(x) - \min(x) + \epsilon}, \tag{1}$$

where $\epsilon = 10^{-5}$ prevents division by zero. This approach scales each lead into $[0, 1]$ while preserving inter-lead dynamics. To account for noisy outliers, we additionally replaced undefined values with zeros.

Finally, each signal was truncated or zero-padded to a fixed length of 4096 samples (≈10.2 s), ensuring consistent input dimensions for model training. This representation provides sufficient temporal context while controlling memory footprint. The resulting dataset consisted of a tensor with shape $(N, 12, 4096)$, accompanied by diagnostic labels and demographic covariates. To address class imbalance, we computed positive class weights as the ratio of negative to positive samples and applied them during loss calculation.

2.2. Confidence-Based Sample Selection

Label noise and heterogeneity across datasets can severely affect supervised training. To mitigate this, we used a ResNet-18 trained with five-fold cross-validation to generate probability estimates for all samples in the first stage. Let p_i denote the probability of sample i being positive. The selection criteria were:

$$t_+ = Q_5(p), \quad t_- = Q_{95}(p) \tag{2}$$

$$P = \{i \mid y_i = 1,\ p_i \in [t_+, 1]\} \tag{3}$$

$$N = \{j \mid y_j = 0,\ p_j \in [0, t_-]\} \tag{4}$$

$$N_{final} = 95 \times P \tag{5}$$

where Q_{95} and Q_5 are the 95th and 5th percentiles of the distribution of predicted probabilities. P and N are the selected positive and negative samples. To balance the dataset, the number of final negatives $|\mathcal{N}|$ was chosen as $95 \times |\mathcal{P}|$. This procedure ensured that the curated subset contained only highly reliable labels.

We evaluated multiple sampling ratios (2%, 10%, 50%, 66%, and 100%) to understand the trade-off between sample reliability and diversity. Empirically, 50–100% sampling offered the best balance, while extremely low ratios reduced coverage.

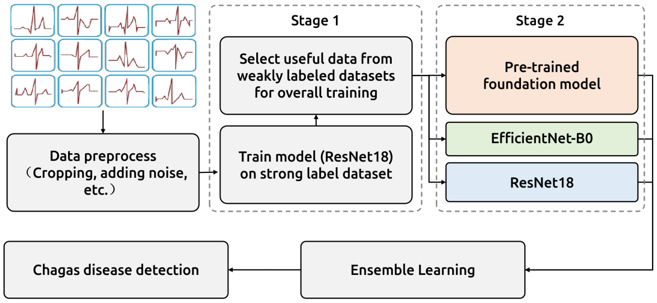
Stage 1
Select useful data from weakly labeled datasets for overall training
Train model (ResNet18) on strong label dataset
Data preprocess （Cropping, adding noise, etc.）
Stage 2
Pre-trained foundation model
EfficientNet-B0
ResNet18
Ensemble Learning
Chagas disease detection

Figure 1. Overview of the Model Framework.

2.3. Model Architectures

Three models were trained independently to capture complementary representations of the electrocardiogram (ECG) signals, the model structure is shown in Figure1:

- **Foundation Model**: The ECG foundation model was constructed using over ten million 12-lead electrocardiogram recordings from more than one million patients, annotated with 150 diagnostic labels [8]. A RegNet-based architecture was adopted to capture both temporal dynamics and spatial correlations across leads. The model was trained with a multilabel classification objective and included strategies to handle incomplete annotations and improve robustness. In addition, single-lead augmentation was incorporated to enhance adaptability for wearable and mobile applications, yielding expert-level performance across diverse diagnostic tasks and providing a versatile backbone for downstream use. Building on this foundation, we performed full fine-tuning of the network using our curated high-confidence dataset. Both the backbone parameters and the classification layer were updated during training, enabling the model to better adapt to the specific task of Chagas disease detection.
- **EfficientNet-B0**: A compact convolutional neural network (CNN) optimized for parameter efficiency [9]. We adapted this architecture for one-dimensional physiological signals by replacing image-based convolutions with temporal convolutions, allowing the model to extract multi-scale temporal features with minimal computational cost.
- **ResNet-18**: A residual CNN architecture [10] that facilitates gradient propagation through shortcut connections. We employed this network both as a baseline for model comparison and as a robust classifier for sample selection and downstream prediction.

2.4. Training Strategy

All the high-confidence samples were used to train the three models in the second stage. All models were trained with binary cross-entropy loss. The Adam optimizer was used with learning rate 10^{-3}, and a ReduceLROnPlateau scheduler adjusted the rate dynamically. Mini-batch size was 64–512 depending on GPU memory. To improve generalization, we applied the following augmentations:

- Random cropping within the 10-s window.
- Lead masking, where 1–2 channels were randomly dropped.
- Amplitude scaling, multiplying signals by factors between 0.9 and 1.1.

For downstream adaptation, we fine-tuned the ECG foundation model on three distinct categories of curated high-confidence datasets, allowing the network to adjust its representations to the specific distributions of Chagas-related signals.

2.5. Fusion Strategy

While each model demonstrates strong performance individually, their error patterns and feature representations differ, suggesting potential gains from combining their outputs. Inspired by recent ensemble learning studies [11], we designed a fusion strategy for the ECG foundation model, EfficientNet-B0 and ResNet-18. Specifically, we assigned the same weights to the three models in the aggregation process and used the auxiliary models to refine decision boundaries and reduce model-specific variance. This foundation model–guided ensemble leverages the strong generalization ability of the pretrained backbone while incorporating the diversity of lightweight CNNs, resulting in improved robustness and stability across folds. For inference, we applied the fusion to obtain the weighted probabilities.

3. Results

Table 1 summarizes our results. The baseline ResNet-18 trained on 10% CODE-15% samples without demographic information achieved a cross-validation (CV) score of 0.308 and 0.310 on the hidden validation set, demonstrating limited predictive capacity when trained with restricted data. Incorporating demographic features (e.g., age and sex) and increasing the proportion of training data provided modest improvements, with scores rising to 0.355 under a 50% sampling regime. These results suggest that demographic priors contain complementary information, but their contribution alone is not sufficient to close the performance gap.

In contrast to shallow baselines, strategies that leveraged high-confidence sampling and pretrained representation models consistently demonstrated superior performance. In particular, our standalone ECG foundation model achieved a Challenge score of 0.379 on the hidden validation set, securing 7th place on the leaderboard. This result highlights the strong transferability of foundation-model–based representations for clinical ECG signals, even when trained under conditions of noisy labels and dataset heterogeneity. The foundation model's robustness suggests that large-scale pretraining captures fundamental electrophysiological patterns that can be effectively adapted to downstream diagnostic tasks.

Model	Training	Validation	Test
ResNet-s10	0.308	0.310	–
ResNet-s50-d	0.355	0.333	–
ResNet-2stage	0.371	–	–
EfficientNet-2stage	0.393	0.326	–
ECGFounder-2stage	0.391	**0.379**	–
Fusion	**0.400**	0.368	–

Table 1. Challenge scores for team MIWEAR across different model configurations. Training scores are obtained by 5-fold cross-validation (CV) on the public training data. Validation scores correspond to the official hidden validation set. Test scores and final rankings will be updated after the conference. ResNet-s10 indicates ResNet-18 trained with 10% randomly sampled CODE-15% data. ResNet-s50-d includes 50% sampled CODE-15% data with demographics. "2stage" refers to high-confidence sampling with a two-stage training scheme. ECGFounder-2stage denotes models initialized from ECG pretraining. Fusion (Confusion-2stage) indicates a weighted ensemble of multiple two-stage models.

Beyond single models, we investigated ensemble learning as a means to further improve generalization. A fusion model guided by the ECG foundation model achieved the highest observed Challenge score of 0.400 on the training set, surpassing all individual models in terms of peak accuracy. The fusion design integrated predictions from diverse architectures while assigning dominant weight to the ECG foundation model, thereby preserving its discriminative capacity while exploiting complementary inductive biases from other networks. Although the ensemble did not outperform the standalone foundation model on the hidden validation set, it exhibited improved robustness across cross-validation folds and reduced susceptibility to model-specific overfitting. We hypothesize that this trade-off—sacrificing a small amount of peak hidden-validation accuracy in exchange for greater fold-to-fold stability—may prove advantageous when models are deployed on unseen clinical cohorts, where distribution shifts and rare pathological patterns are common.

4. Discussion and Conclusions

Our results show that confidence-based sample selection is a practical and effective strategy for mitigating label noise in large-scale ECG datasets. By prioritizing high-confidence examples, the models were less affected by mislabeled or ambiguous data, leading to more stable training and consistent improvements across folds.

The ECG foundation model proved to be the most competitive single-model solution, achieving a Challenge score of 0.379 on the hidden validation set and ranking 7th overall. This emphasizes the strong transferability of pretrained ECG representations, which capture generalizable features even in the presence of distributional shifts.

Beyond individual models, a fusion model guided by the ECG foundation model reached the highest score of 0.400 on the training set. While its hidden validation performance did not surpass the standalone foundation model, the ensemble improved robustness across folds and reduced model-specific variance. This suggests that ensembles can complement foundation models, though more sophisticated fusion strategies (e.g., adaptive weighting or uncertainty-aware methods) may be required to fully exploit model diversity.

Incorporating demographic information such as age and sex yielded only limited benefits, likely due to incomplete or inconsistent metadata. Future work should investigate tighter multimodal integration, dynamic confidence thresholds for sample selection, and adaptive ensemble weighting to further enhance generalization. Despite these challenges, our framework achieved competitive performance, highlighting the promise of foundation-model–based approaches as a backbone for robust and generalizable ECG classification.

Taken together, these findings point to three overarching conclusions. First, foundation models provide a substantial performance advantage over conventional CNN baselines, underscoring their value in ECG representation learning. Second, data curation through high-confidence sampling is indispensable for mitigating label noise and ensuring reliable supervision in real-world clinical datasets. Third, fusion strategies enhance robustness and reduce variance, though their design requires careful consideration: an ensemble dominated by the foundation model improves stability but may dilute peak accuracy, while lighter fusions can preserve higher scores but remain less stable across folds. Future work should explore adaptive weighting schemes and hybrid architectures that reconcile these trade-offs, aiming to deployment reliability in clinical practice.

Acknowledgments

The authors declare that they have no conflict of interest.

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. Computing in Cardiology 2025;52:1–4.

[3] 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.

[4] 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.

[5] 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.

[6] Nunes M, Buss L, Silva J, Martins L, Oliveira C, Cardoso CS BB, 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.

[7] 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.

[8] Li J, Aguirre A, Moura J, Liu C, Zhong L, Sun C, et al. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. arXiv preprint arXiv241004133 2024;.

[9] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning. PMLR, 2019; 6105–6114.

[10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; 770–778.

[11] Wu W, Tan Y. Melicientnet: Harnessing mel-spectrograms and efficientnet architectures for predicting neurological recovery post-cardiac arrest. In 2023 Computing in Cardiology (CinC), volume 50. IEEE, 2023; 1–4.

Address for correspondence:

Wenlong Wu
33 Xi'erqi Middle Road, Haidian District, Beijing, China
wuwenlong3@xiaomi.com