Finetuning Foundational ECG Models to Detect Chagas Disease

Kelvin Nguyen[1], Andy Smithwick[1], Maxwell Loetscher[1], Shadi Manafi[1], Zaniar Adarlan[1], Saman Parvaneh[1]

[1]Edwards Lifesciences, Irvine, USA

Abstract

Chagas disease is a parasitic illness spread by triatomine bugs that can cause serious heart problems. Mass serological testing is costly, so automated early detection using ECGs is desirable as it is scalable and cost-effective. As part of the George B. Moody PhysioNet Challenge 2025, our team (Chagas_detector) utilized an ECG foundational model (ECG-FM) pretrained on 1.5 million ECGs, and finetuned only an added classification head, leveraging the model's existing feature extraction capabilities, to detect Chagas disease from 12-lead ECGs. The pipeline included signal preprocessing by re-sampling signals to 500 Hz, applying a 0.5 Hz high-pass butterworth filter, followed by powerline filtering, which was then passed into a frozen ECG-FM that encodes the signal into a 768-dimension vector. A classification MLP head, which combines this vector with patient age and sex, was then finetuned to output a binary probability for Chagas disease. We train on all of the PTB-XL, SaMi-Trop, and CODE-15% records. Our model received a Challenge score of .323 (ranked ?) on the hidden validation set.

1. Introduction

Our team joined the George B. Moody PhysioNet Challenge 2025 to develop open-source algorithms that automatically detect Chagas disease from ECG data [1, 2]. Although serological testing is the standard practice for confirming Chagas disease, ECG-based detection helps to classify patients to accommodate limited serological testing capacities. To this end, our team, Chagas_detector, developed a pipeline that takes a 12-lead ECG and combines traditional ECG pre-processing techniques with a finetuned 90.4 million parameter deep-learning ECG foundational model (ECG-FM) [3] to output a probability for Chagas disease. This pipeline is trained on open source ECG data provided by the challenge organizers [4–8].

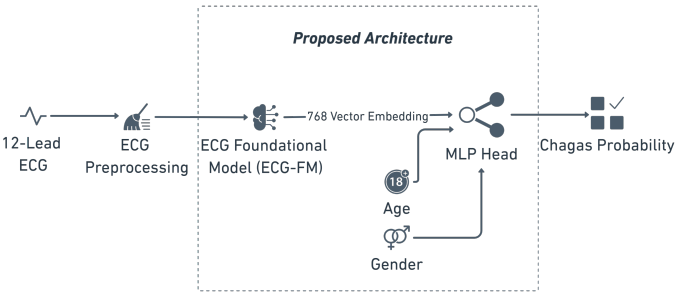

Figure 1. Proposed Architecture

2. Methods

Our proposed architecture is shown in figure 1. The pipeline consists of three parts: preprocessing the ECG signal, encoding the signal using the ECG-FM model [3], and passing in the resulting embedding along with demographic data into a multilayer perceptron (MLP) classification head to output a binary probability for Chagas. For training, we only update the MLP head, keeping the ECG-FM model frozen. See table 1 for a table of our hyperparameters.

2.1. Data

To train our pipeline, we used the PTB-XL (f_s=500Hz), SaMi-Trop (f_s=400Hz), and CODE-15% (f_s=400Hz), datasets where f_s is a sampling frequency [4–6] . To find the best architecture, we created a stratified internal training, internal validation, and internal test splits, with an 80%/10%/10% distribution, respectively; however the final submitted model had the internal validation set combined with the internal test set to form one larger internal validation set.

2.2. Preprocessing

Since ECG-FM is trained on ECGs with a sampling frequency of 500 Hz, we resampled our input signals to 500 Hz. We then pass the signal into the Python package Neu-

rokit2 [9] for ECG preprocessing which applies a 0.5 Hz high-pass butterworth filter and powerline filtering. We also obtain the age and sex of the patient represented by the ECG from the associated header file, and we normalize the age using the statistics obtained by the training data. We obtain a 10 second window of the recording, where we truncate the end of the recording if it is too long, or pad the end if it is too short. This window as well as the demographic features are then passed on to the next step of the pipeline.

2.3. Architecture and Finetuning

Table 1. Hyperparameters for ECG-FM Finetuning, BCE means Binary Cross Entropy.

Hyperparameter	Value
Data & Preprocessing	
Unified Frequency	500 Hz
Window Time	10 seconds
Sequence Length	5000
Windowing Method	Pad/Truncate End
Batch Size	16
Model & Architecture	
Freeze Encoder	'True'
Demographic Features	Age, Sex
Training & Optimization	
Epochs	20
Loss Function	BCE
Initial Learning Rate	1×10^{-4}
Linear Learning Rate Start Factor	1×10^{-6}
Linear Learning Rate End Factor	1
Warmup Steps	700
Final Learning Rate	1×10^{-7}
Dropout Rate	0.1
Checkpoint Monitor	Val Challenge Score

In this paper, we utilize the ECG-FM model [3], a foundational ECG model pretrained on 1.5 million ECG records. We specifically use a model version already finetuned on the MIMIC-IV-ECG dataset [10]. Our proposed architecture includes ECG-FM followed by a MLP classification head with ECG-FM's 768-dimension embedding and patient age and sex as input. More specifically, this classification head first applies dropout to the resultant 768-dim vector returned by ECG-FM with rate 0.1. Then, we project the combination of this vector plus the demographic features to a dimension of 256, apply ReLU, apply another dropout with the same rate of 0.1, and then reduce the dimensions to 1, our final probability for Chagas.

For finetuning, we freeze the ECG-FM model, and only allow the MLP classification head to be updated. We freeze the ECG-FM model to prevent catastrophic forgetting, as the pretrained model has already been trained on an ECG dataset nearly 4 times larger than our ours.

We train the model for 20 epochs with binary cross-entropy loss (BCE), and use the true positive rate in the top 5% of predicted probabilities (the Challenge metric) as our validation metric. We save the checkpoint from the epoch that scores the highest on our metric, and chose 20 because although the validation challenge score plateaued after epoch 13, we wanted to be absolutely sure about model convergence. We use the AdamW optimizer [11], starting at the conservative 1e-4. To reduce overfitting, we first apply a learning rate warmup of 700 steps, with a starting factor of 1e-6 (meaning our actual starting learning rate is 1e-10) and after the warmup period, we use a cosine annealing learning rate scheduler [12], with an ending learning rate of 1e-7. We trained with batch size 16.

2.4. Comparison with Other Challenger Methods

Using a pretrained encoder backbone is a technique used by other teams as well; however, they pretrain only on the given training data itself. Since the challenge dataset only provides approximately 350,000 records, we reasoned that using a pretrained model with exposure to a dataset over 4 times larger would contribute to better performance, cut training time required to pretrain our model, and would save precious training data for finetuning.

3. Results

The challenge score in the internal test set and challenge validation set were 0.425 and 0.323, respectively (table 2).

Table 2. Challenge scores for our selected entry (team Chagas_detector), including the ranking of our team on the hidden test set. We used stratified 80/10/10 train/val/test split on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

Training	Validation	Test	Ranking
0.425	0.323	??	??

4. Discussion and Conclusions

One of the challenge's largest issues was label reliability; the CODE-15% labels are considered weak as they are not serologically validated, while also making up 93.60% of the training set; on the other hand, SaMi-Trop has serologically validated labels, and PTB-XL has patients from geographical regions with negligble rates of Chagas. As such, we experimented with training on just the PTB-XL

and SaMi-Trop datasets, with 5-fold training. This massively underperformed, as we received an official validation score of 0.081, compared to the 0.323 when adding CODE-15% and training without k-fold.

Another attempt to overcome the label reliability issue of CODE-15% involved the fact that the dataset has cardiologist-validated labels for right bundle branch block (RBBB) and first-degree atrioventricular block (1dAvb), which are very strong indicators of Chagas [13]. As such, we implemented multi-task learning, where for CODE-15% records, the model has to predict Chagas as well as RBBB and 1dAvb to ensure generalizability, and the weighted loss across all three tasks are used as the final batch loss. However, the model underperformed on the internal test set, getting a top score of 0.349 when using all 3 datasets, BCE loss, and 0.6, 0.2, and 0.2 task loss weighting on the Chagas, RBBB, and 1dAvb loss, respectively, compared to 0.426 from the main architecture.

Our internal test scores were also consistently higher than the official validation scores (see table 3). Notably, models trained without the weakly-labeled CODE-15% data performed best internally but underperformed on the official validation set, suggesting a domain shift between our internal splits and the hidden challenge data. We also observed that unfreezing ECG-FM led to lower results on our internal sets, which is probably because the useful pretrained knowledge was being overwritten.

Our model could also benefit from a more hypertuned sequence duration, as well as random padding or random windowing. We also believe that finding strong features besides demographics could improve the model; we did also try selecting our own wide features, ranked for importance using a random forest classifier; however, feature extraction proved to take too long (over 72 hours for the entire dataset), so it was not feasible under the challenge rules. Since successful models predicting other conditions have benefited greatly from strong feature selection, such as the PRNA model from the 2020 PhysioNet Challenge [14], which won first place in predicting 27 heart conditions, many of which were indicative of Chagas (such as RBBB and 1dAvb), we believe more work in extracting features will drastically improve our architecture.

We also explored other architectures. We tried ensembling a custom squeeze and excitation model (SE) with our main proposed finetuned ECG-FM model. The SE model by itself with multi-task learning had an internal test score of 0.540 and an official validation score of 0.079, but when ensembled with our finetuned ECG-FM model, the official validation score was 0.250, our third best performing model. This improvement over just the SE model highlights the importance of the rich pretraining knowledge of the ECG-FM model in reducing the effects of overfitting.

We also tried a loss function optimizing for the challenge's true positive rate metric by using a custom ranking loss which we call "Percentile Ranking Loss." This involves maintaining a running score threshold t which estimates the top 5% (the Challenge metric cutoff) of all prediction scores seen through training across all batches and epochs. We use an exponential moving average to combine every batch's distribution of scores with other batches, to smooth out outlier batches. Then, if positive samples are below t plus some hyperparameter margin m, they are penalized, and if negative samples are above $t - m$, they are penalized. Since t is inaccurate early in training since has not seen many batches, during the warmup period, batches calculate and use their own local version of the moving average in place of t, while t continues to be updated. Unfortunately, this underperformed internally relative to traditional binary cross entropy, as you can see in the "Percentile" loss category in table 3.

Table 3. Challenge scores for different architectures on internal test set, where Scenario 1 and 2 are training with all data and training data excluding CODE-15%, respectively (* indicates official validation score)

Encoder State/Loss	**Scenario**	**Internal Test Score**
Unfrozen/Percentile	1	0.335
Unfrozen/BCE	1	0.342
Unfrozen/BCE	2	0.515
Frozen/BCE	1	0.426 (*0.323)
Frozen/Percentile	2	0.485
Frozen/BCE	2	0.424
Frozen/BCE	K Fold with 2	0.534 (*0.081)

We experimented with the use of a wide and deep transformer. This model consisted of a traditional transformer model enhanced with RoPe attention [15]. When training with just the PTB-XL and SaMi-Trop datasets, we obtained an official validation score of 0.284 (our second highest score), with an internal validation score of 0.349, which underperforms our finetuned ECG-FM model.

We also explored using other encoder models. We tried using the PRNA model by passing our ECGs into the PRNA model and then feeding the output into XGBoost along with the 20 wide features that the model produces; however, this received an internal challenge score of 0.22. Some chronic cases of Chagas can present with normal ECG findings [16], making us believe that extracted features such as the outputs of the PRNA model might not be representative of all cases in available datasets. Therefore, a deeper and more complex feature extractor model similar to ECG-FM may have more power to detect patterns that are not captured by traditional ECG findings and features.

In an attempt to address this issue, we also experimented with a transferring the knowledge of the PRNA model to our more complex finetuned ECG-FM archi-

tecture. Specifically, during training, we added an MLP adapter on top of the PRNA model to project the 27 final logits down to a single value (the "soft target"). Then, we forward pass through our ECG-FM finetuning architecture as usual, and produce a Chagas probability, which is then compared to the soft target to calculate the distillation loss. The combination of the distillation loss plus the normal BCE loss between the Chagas probability and actual label forms the total loss. Finally, we only update the MLP head on both the ECG-FM model and the PRNA models. The top internal test score when we ran this knowledge distillation pipeline combined with the multi-task learning objective described above was 0.331. The higher score compared to the XGBoost model once again suggests that deeper, more complex networks are required for Chagas; however, we believe that the addition of the knowledge distillation performed worse than having just multi-task learning is either due to the limitations of the adapter, or perhaps we need to find another layer of the PRNA model to generate our soft target. In conclusion, we demonstrate the potential of adding classification heads on frozen foundational ECG models in detecting Chagas. We show across all of our tested methods that the massive amount of training data ECG-FM is exposed to allows it generalize to Chagas detection effectively while also mitigating the effects of overfitting that limits our other attempted methods. We provide different potential architectures, including the wide and deep transformer and the custom SE model. We finally discuss potential regularization techniques for training, including multi-task learning, ensembling, and knowledge distillation with PRNA.

Acknowledgments

We have no conflicts of interest to declare.

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] McKeen K, Masood S, Toma A, Rubin B, Wang B. Ecg-fm: An open electrocardiogram foundation model, 2025. URL https://arxiv.org/abs/2408.05178.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, et al. NeuroKit2: A python toolbox for neurophysiological signal processing. Behavior Research Methods feb 2021;53(4):1689–1696. URL https://doi.org/10.3758%2Fs13428-020-01516-y.

[10] Gow B, Pollard T, Nathanson LA, Johnson A, Moody B, Fernandes C, et al. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. PhysioNet 2023;URL https://doi.org/10.13026/4nqg-sb35.

[11] Loshchilov I, Hutter F. Decoupled weight decay regularization. In International Conference on Learning Representations. 2019; URL https://openreview.net/forum?id=Bkg6RiCqY7.

[12] Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. In International Conference on Learning Representations. 2017; URL https://openreview.net/forum?id=Skq89Scxx.

[13] Rojas LZ, Glisic M, Pletsch-Borba L, Echeverría LE, Bramer WM, Bano A, et al. Electrocardiographic abnormalities in chagas disease in the general population: A systematic review and meta-analysis. PLoS Neglected Tropical Diseases 2018;12(6):e0006567.

[14] Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S, et al. A wide and deep transformer neural network for 12-lead ecg classification. In 2020 Computing in Cardiology. 2020; 1–4.

[15] Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing 2024;568:127063. ISSN 0925-2312. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

[16] Almeida-Filho OC, Maciel BC, Schmidt A, Pazin-Filho A, Marin-Neto JA. Minor segmental dyssynergy reflects extensive myocardial damage and global left ventricle dysfunction in chronic chagas disease. Journal of the American Society of Echocardiography Jun 2002;15(6):610–616.

Address for correspondence:

Saman Parvaneh
1 Edwards Way, Irvine, CA, USA, 92614
parvaneh@ieee.org