Detecting Chagas Disease Using a Vision Transformer–based ECG Foundation Model

Lore Van Santvliet[1], Phu Xuan Nguyen[1], Bert Vandenberk[2,3], Maarten De Vos[1,4]

[1]STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium
[2]Department of Cardiology, University Hospitals Leuven, Leuven, Belgium [3]Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium [4]Department of Development & Regeneration, KU Leuven, Leuven, Belgium

Abstract

Foundation models (FMs) are reshaping machine learning and, by extension, computational cardiology. By exploiting large, heterogeneous, and possibly unlabeled datasets through self-supervised learning, these models scale to large model parameter counts and provide highly expressive feature extraction capabilities. Such pretrained feature extractors are particularly promising for downstream applications in low-data settings, including rare disease detection.

In this work, we demonstrate the use of a vision transformer–based FM for clinical 12-lead electrocardiograms (ECGs) as a prescreening tool for Chagas disease, in the context of the "Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025" (team Biomed-Cardio). Using cross-validation on the public training data, our method achieves 0.490±0.008 for the challenge metric. On the hidden validation set, it reaches 0.445, placing [...] out of [...] teams.

These results highlight the value of extensive pretraining for learning robust ECG representations, and their effectiveness in downstream Chagas detection. At the same time, this work underscores the challenges of applying FMs to unseen datasets, where distribution shifts and other pitfalls must be carefully addressed.

1. Introduction

Foundation models (FMs) are increasingly applied in computational cardiology to address complex downstream tasks such as electrocardiogram (ECG) classification [1]. These models are typically pretrained using self-supervised learning, exploiting large, heterogeneous datasets that may be unlabeled or labeled for unrelated tasks. During pretraining, the goal is to learn robust, general-purpose feature representations. Only in the subsequent fine-tuning stage are task-specific labels required. This partial decoupling of task-agnostic feature learning and task-specific learning enables the training of larger and more expressive models than would be feasible with supervised learning alone, where limited labeled data constrain model complexity.

Chagas disease, a parasitic condition endemic to Latin America, is associated with electrophysiological abnormalities [2]. Whereas serological testing remains the gold standard for diagnosis, ECGs can be used for non-invasive population screening, in order to optimize the allocation of limited testing capacity. Motivated by the robust and transferable feature representations learned by ECG FMs, we investigate the use of a vision transformer (ViT)–based FM [3], integrating feature representations from all intermediate encoder layers, to detect Chagas disease from clinical 12-lead ECGs in the 2025 George B. Moody PhysioNet Challenge [4, 5].

2. Methods

2.1. FM terminology

The term "foundation model" has been used with varying definitions [1]. In this work, we adopt two criteria: (1) pretraining with self-supervised learning, and (2) extensive pretraining on a dataset larger than those used in downstream applications. Our FM meets both criteria: it was pretrained using generative self-supervised learning, and the pretraining dataset contained 400 365 samples, indeed exceeding the downstream training set of 369 267 samples in size, although only by a limited margin.

2.2. FM pretraining

We used a one-dimensional (ViT)–based FM composed of an encoder with 12 transformer blocks and a decoder

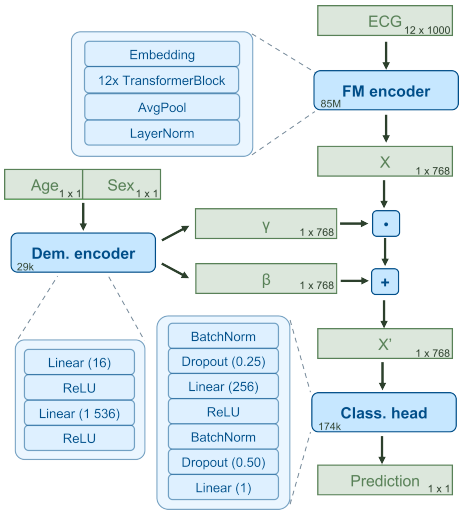

Figure 1. Architecture of the classification model. Its three main constituents, the foundation model (FM) encoder, demographic (dem.) encoder, and classification (class.) head, are depicted in blue, with the number of model parameters noted in the bottom left corner. Architectural details (output dimension for linear layers and dropout probability for dropout layers) are added between brackets. Inputs and outputs of each part are depicted in green, with their size noted in the bottom right corner. X and X' denote the original and modified feature vector, respectively.

with 4 blocks. It processes 12-lead ECGs of length 1 000 samples and sampling frequency 100 Hz, using a patch size of 50 samples. The encoder outputs a 768-dimensional feature vector, which is subsequently fed into the decoder.

Pretraining followed the spatio-temporal masked ECG modeling (ST-MEM) framework [6], in which 75% of the input patches were randomly masked. The FM was trained to reconstruct the masked patches using a mean squared error loss. For further details on the FM architecture, and pretraining datasets and procedure, we refer to [3].

2.3. FM fine-tuning

The final classification model used the FM encoder as a feature extractor. Using the Aggregation-of-Layers (AoL) scheme [3], intermediate feature vectors were extracted from all 12 encoder layers and aggregated via average pooling to produce a 768-dimensional feature representation. This vector was then shifted and scaled by the output of a demographic encoder, a multi-layer perceptron (MLP) that receives sex (binary encoding) and age (in centuries) as input. An MLP with hidden dimension of 512 was used as a classification head. The architectures are illustrated in Figure 1.

Model fine-tuning for the Chagas detection task was performed using the public training datasets provided by the challenge: PTB-XL [7], CODE-15% [8], and SaMi-Trop [9]. ECG lengths vary between 5 and 11 seconds, and all ECGs, with original sampling frequencies of 400 or 500 Hz, were resampled to 100 Hz. Label certainty differs across datasets: negative PTB-XL labels and positive SaMi-Trop labels are considered highly reliable, reflecting the disease localization and serological testing, respectively, whereas mixed CODE-15% labels are self-reported and thus less reliable. To account for this uncertainty, we introduced soft labels for CODE-15%, setting positive labels to 0.8 and negative labels to 0.2, while retaining strong labels (0 and 1) for PTB-XL and SaMi-Trop.

We used a weighted binary cross-entropy loss function for fine-tuning, assigning a weight of 5 to positive cases to address class imbalance. Additionally, we implemented weighted sampling, oversampling positive Chagas samples by a factor of 5. The fine-tuning phase started with frozen-backbone training, in which only the classification head was trained for 2 000 iterations, using a learning rate of 2×10^{-4} and batch size 64. Next, all model weights were unfrozen, and the full model was fine-tuned for 12 000 iterations with a learning rate of 2×10^{-5} and the same batch size.

Augm.	Property	Sampling interval	Unit
Powerline	f	$[50 \pm 0.2] \cup [60 \pm 0.2]$	Hz
inter-	ϕ	$[0, 2\pi]$	rad
ference	SNR	$[15, 30]$	dB
Cropping	L_1	$[5.65, 10]$	s
Shifting	L_2	$[0 \pm \min(1, (10 - L_1))]$	s

Table 1. Implementation details of augmentation (augm.) techniques. f, ϕ and SNR refer to frequency, phase and signal-to-noise ratio of the powerline interference, respectively. L_1 depicts the remaining signal length after cropping, and $L2$ depicts the shifting length along the time axis. Variables are drawn from the indicated intervals using uniform sampling.

2.4. Data augmentation

A key difficulty in this challenge is the final out-of-distribution evaluation on hidden validation and test datasets with unknown origin. Subtle dataset-specific cues, particularly the frequency of powerline interference, might be beneficial for performance in the training dataset, but possibly lose all value for detection in new datasets.

To build in robustness against this potentially unreliable confounder, we implemented a powerline interference augmentation strategy during fine-tuning. More specifically, we randomly added synthetic powerline noise around either 50 or 60 Hz to training samples. The augmentation

Training	Validation	Test	Ranking
0.490 ± 0.008	0.445	[...]	[...]/[...]

Table 2. Challenge scores for our selected entry (team Biomed-Cardio), including the ranking of our team on the hidden test set. We used 5-fold cross validation on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

included variable frequency, phase and signal-to-noise ratios (SNRs) to mimic realistic powerline interference of moderate amplitude (see Table 1 for implementation details). The second and third harmonics of this powerline noise were also added, with a random phase and a SNR that is equal to half and one third of the main SNR, respectively. The augmentation was applied with a probability of 0.5, and identical noise was added across all 12 ECG leads.

Random cropping and temporal shifting of the ECG were implemented as additional augmentation strategies. Positive and negative temporal shifting are equivalent to adding zero padding prior to or after the signal, respectively. Implementation details are provided in Table 1. This augmentation was applied to all training samples during fine-tuning, identically across all 12 ECG leads to preserve physiological consistency. After cropping and shifting, all signals were zero-padded to reach a length of exactly 10 seconds. We also experimented with cutmix augmentation and manifold mixup, but these strategies did not yield noticeable improvements for this task and were therefore not included in the final implementation.

2.5. Model selection and evaluation

Our algorithm was evaluated through five-fold cross-validation on the challenge training set, with mean and standard deviation of the challenge metric, top5%-true positive rate (top5%-TPR), the area under the precision-recall curve (AUPRC), the area under the receiver operating characteristic curve (AUROC) and the F1 score reported across folds. For out-of-distribution evaluation on the hidden evaluation and test sets, we randomly divided the public training set into an internal training (80%) and holdout (20%) subset. The internal holdout subset was used for model selection, choosing the model with the highest top5%-TPR. The prevalence of Chagas in the top5% ranked by our model is calculated as top5%-TPR $\times$ $p/0.05$, where p is the endemic Chagas prevalence (assumed 0.02).

3. Results

Our ECG FM achieves an out-of-distribution validation set top5%-TPR of 0.445. The prevalence of Chagas in the top5% of this hidden validation set, provided by our algorithm, is 0.178. The training set top5%-TPR obtained via cross validation equals 0.490 ± 0.008 (0.381 ± 0.003 for frozen-backbone training only), AUPRC 0.252 ± 0.008, AUROC 0.867 ± 0.001, and F1 score 0.116 ± 0.002. The most important results are summarized in Table 2.

Between an initial submission, which already included cropping and shifting augmentations but none of the other described tweaks, and our final model, we increased our challenge score on the hidden validation set by 5% (from 0.395 to 0.445). In terms of performance within the public training set, the difference between this initial and final submission was notably less pronounced (only 2.7%).

4. Discussion and Conclusions

4.1. Utility of the ECG FM

Applying our Chagas detection model for prescreening, and selecting the identified top5% for testing, the proportion of infected individuals receiving serological tests would amount to 44.5%, representing a nearly nine-fold improvement over the random baseline level of 5%. The obtained Chagas prevalence of 0.178 in the top5% means that the expected number of individuals that need to be tested serologically, in order to detect one Chagas-infected individual, is decreased from 50 (with baseline prevalence 0.02) to 6, when using our prescreening tool. Put differently, the fraction of serological tests spent on uninfected individuals, rather than serving the identification of infected individuals, would decrease from 98% under random selection (with endemic prevalence 0.02) to 82.2% when guided by our algorithm, markedly improving the efficiency of limited testing resource usage.

The performance that was obtained when training solely the classification head, and keeping the FM backbone weights frozen, illustrate the strong feature extraction capabilities of the FM. Even though the masked modeling pretext task is unrelated to the Chagas detection task, the FM's feature representations after pre-training already proved particularly useful as-is for this downstream task. Nevertheless, end-to-end fine-tuning of the model still increased the top5%-TPR by approximately 11%.

4.2. Limitations of the ECG FM

The overlap between pretraining and fine-tuning data is substantial, as both CODE-15% and PTB-XL appear in both stages. Further expansion and diversification of the pretraining dataset would likely strengthen its foundation character in two ways. First, a larger dataset would enable scaling up the encoder, facilitating the discovery of more subtle and complex ECG representations. Second, more diverse input data would potentially improve generalizability, making the model applicable across a wider range of

downstream tasks, demographic groups, and diseases.

More advanced aggregation methods than average pooling could be used for feature aggregation from all 12 FM encoder layers. The gating-based Mixture-of-Layers scheme, proposed in [3], is a promising alternative to the AoL scheme used in this work, potentially allowing even more optimized feature representations.

4.3. Mindful FM application

There is a notable difference in the effect of certain tweaks within the training set, compared to the hidden validation set. One possible explanation for this could be found in hidden confounders, and in particular a difference in the extent to which they are present in the challenge's training and validation set.

In this context, this challenge illustrates one of the key pitfalls when applying expressive deep learning models, caused by their black-box nature. Their high flexibility not only enables strong performance but also increases the risk of potentially unknowingly relying on such hidden confounders, which becomes particularly problematic when evaluating in out-of-distribution settings.

In the challenge, for instance, spurious cues such as the original sampling frequency or the presence and type of powerline interference could artificially boost in-distribution performance but fail under external evaluation. Addressing such vulnerabilities requires careful consideration of such confounders and implementation of mitigating techniques, which can be as simple as resampling and powerline noise augmentations. Explainability techniques could potentially reveal additional hidden confounders that may underlie the persistent gap between in-distribution and out-of-distribution performance, and could therefore be a valuable future extension of the FM.

4.4. Conclusions

We demonstrate the use of a ViT–based ECG FM, pretrained with an ST-MEM objective, for Chagas detection. Our approach integrates a demographic encoder to adapt FM features with age and sex information. To address class imbalance and label uncertainty, we combined weighted oversampling, a weighted binary cross-entropy loss, and soft labeling. Robustness was further enhanced through data augmentation strategies, including powerline interference, cropping and shifting.

Our algorithm achieved a 0.445 top5%-TPR in the George B. Moody Challenge of 2025, demonstrating the potential of a ViT–based ECG FM as non-invasive and scalable Chagas prescreening tool in resource-limited endemic regions. Beyond this specific task, our findings highlight the broader promise of FMs for computational cardiology, while also underscoring the need for careful fine-tuning and confounder control in real-world applications.

Acknowledgments

This research is funded by a PhD fellowship fundamental research from the Research Foundation - Flanders, Belgium (FWO) for L. Van Santvliet (FWO project number 1107725N); the Flanders AI Research Program, Belgium; a research grant for the project "Artificial Intelligence (AI) for data-driven personalised medicine" from FWO (FWO project number G0C9623N); the 'Bijzonder Onderzoeksfonds KU Leuven (BOF)' (BF-PhD: "VHeart FM: a foundation model for ECG analysis in Vietnam").

References

[1] Clifford G. Past, Present and Future Challenges in Sharing Science: From PhysioNet to Foundation Models. In Computing in Cardiology, volume 51. December 2024; 1–4.

[2] Marcolino MS, Palhares DM, Ferreira LR, Ribeiro AL. Electrocardiogram and Chagas Disease: A Large Population Database of Primary Care Patients. Global Heart September 2015;10(3):167.

[3] Nguyen PX, Phan H, Pham H, Chatzichristos C, Vandenberk B, De Vos M. Exploiting a Mixture-of-Layers in an Electrocardiography Foundation Model. In preparation.

[4] 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.

[5] 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.

[6] Na Y, Park M, Tae Y, Joo S. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. International Conference on Learning Representations January 2024;12.

[7] 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.

[8] 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.

[9] 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.

Address for correspondence:

Lore Van Santvliet
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
lore.vansantvliet@kuleuven.be