Transformer–xLSTM Ensembles for ECG-Based Chagas Disease Detection

Angus Nicolson[1], Riccardo Lunelli[1], Samuel Martin Pröll[1], Nadja Gruber[1], Axel Bauer[2], Clemens Dlaska[1]

[1] Digital Cardiology Lab, University Clinic of Internal Medicine III. Medical University of Innsbruck, Austria
[2] University Clinic of Internal Medicine III, Medical University of Innsbruck, Austria

Abstract

Automated ECG-based detection of Chagas disease, the focus of the 2025 PhysioNet Challenge, presents two major challenges: noisy supervision due to the self-reported weak labels in the CODE-15% dataset and severe class imbalance (2% prevalence). We address both issues through large-scale pretraining and dataset-asymmetric finetuning.

We combine the complementary strengths of attention-based models and recurrent architectures by pretraining multiple foundation models – masked autoencoding transformers and xLSTMs trained with simDINOv2. This enables the learning of low-level ECG representations without relying on label quality. During finetuning, we utilise the known disparity in label noise between datasets by applying smooth labelling to the CODE-15% dataset, where the labels are self-reported, but not to the PTB-XL or Sami-Trop datasets where the labels are more reliable.

Our team (DlaskaLabMUI) ranked 2nd out of 40 *teams, with a score of* 0.283 *on the hidden test set.*

1. Introduction

In the 2025 PhysioNet Challenge [1–3], the focus was electrocardiogram (ECG)-based detection of Chagas disease. ECGs are desirable for Chagas classification as they are a low-cost, non-invasive tool that could inform the use of limited serological testing capacities.

The challenge presents two substantial hurdles. First, CODE-15% contains labels derived from self-reported diagnoses, which introduces substantial label uncertainty. Second, the prevalence of Chagas disease within the training set is approximately 2%, creating a highly imbalanced classification task. These conditions require approaches that are robust to weak supervision and rare positives.

In this work, we combine large-scale self-supervised (SSL) pretraining of conceptually different foundation models with dataset-specific fine-tuning that accounts explicitly for label quality and class imbalance.

2. Methods

Data For pretraining we used the CODE dataset of 8M ECGs [4]. CODE is a large dataset collected by the Telehealth Network of Minas Gerais (TNMG), Brazil [5]. The xLSTM model was additionally pretrained on INCART [3], Chapman [6] and Ningbo [7]. For finetuning we used the datasets provided in the 2025 PhysioNet Challenge [1–3]: CODE-15% [8], PTB-XL [9], and Sami-Trop [10].

We split patients into train/val/test (85.5%/4.5%/10%) and used results on the validation set for model selection. In order to avoid confusion with the hidden validation and hidden test sets these internal sets will be named the 'internal validation' and 'internal test' sets.

Model Architectures and Pretraining Previous work uses convolutional neural networks (CNNs) for Chagas classification [11]. We instead use an ensemble of vision transformers [12] and xLSTMs [13], leveraging their complementary strengths in large-scale pretraining and temporal modelling.

Transformers. We adopt masked autoencoding [14, 15], where an encoder-decoder architecture reconstructs an input signal from a masked set of patches – see Figure 1 for an overview. First, the input signal is flattened and split into non-overlapping patches. These patches are then linearly encoded into tokens and a positional embedding is added. Each patch contains information from just one lead and hence self-attention is applied across both time and leads. During pretraining, 80% of tokens are randomly masked and the encoder processes only the visible subset. The decoder reconstructs the full signal from visible embeddings and shared mask tokens, with MSE loss against the original ECG. Positional embeddings are reapplied at the decoder to localise mask tokens. Compared to ST-MEM [15], which uses lead-wise decoding, our design uses a simpler positional embedding and relies on high mask ratios to enforce non-trivial learning. During finetuning, the classifier (CLS) token is passed to a linear head to predict Chagas likelihood.

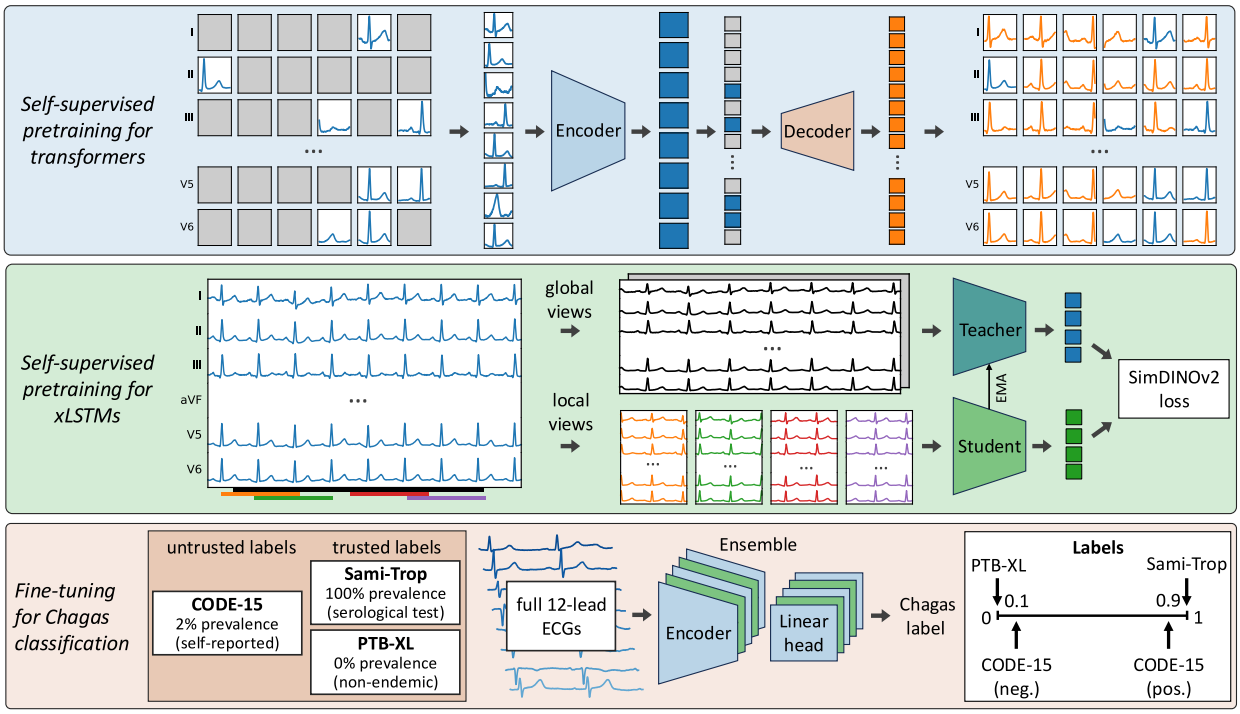


Figure 1. Transformer models are pretrained using masked autoencoding (top), where patches are randomly masked before being passed to the encoder and then reconstructed by the decoder. The xLSTM model is pretrained using SimDINOv2, a teacher-student SSL learning paradigm (middle), where the student predicts the features of the moving average teacher. During finetuning (bottom), there is an asymmetric treatment of labels, with smoothing applied to self-reported labels (CODE-15%) but not reliable labels (PTB-XL and Sami-Trop).

xLSTMs. xLSTMs extend classic LSTMs with sLSTM and mLSTM blocks that improve scalability and gradient flow on long sequences [13]. Vision-LSTM later introduced bidirectional variants as an alternative to vision transformers [16]. Here we use **xECG** [17], a recently proposed foundation model which combines these ideas by stacking alternating sLSTM and mLSTM layers in a bidirectional pattern (i.e., s, s, m, m, s, s, ...), where each layer processes the sequence in the opposite direction to the previous one. Unlike the transformers, which patchify each lead independently, xECG patches span *all 12 leads*. Pretraining follows SimDINOv2 [18], a stable non-contrastive teacher–student SSL framework with multi-view augmentations. For pretraining details see the xECG paper [17].

Hyperparameters. For transformers, we pretrain a base (86M parameters) and large (177M) model with AdamW, where in each case, the decoder was an order of magnitude smaller than the encoder (see Table 1). This makes the MAE pretraining more efficient, as only the decoder requires the full sequence of tokens. We pretrain an xECG model (57M parameters), using the hyperparameters in Table 1. We preprocessed ECGs by resampling to a consistent frequency depending on the model. For xECG, we apply augmentations including drop-lead (0.2), jitter (0.1), amplitude scaling (0.1), and batch-wise baseline shuffle. The transformers use bandpass filtering (0.5-60 Hz), lead-wise Z-score normalisation, and for a consistent input length of 3072 samples in each lead (7.68s), we employ random padding and random cropping. Pretraining took 3 days (base transformer and xECG) to 13 days (large transformer) on an NVIDIA H100 GPU.

Finetuning During supervised training, we differentiated between datasets based on label noise:

- CODE-15%: Labels derived from self-reports. We applied label smoothing (smoothing factor $\alpha = 0.2$) to account for uncertainty.
- Sami-Trop: Serologically confirmed labels. No label smoothing was applied.
- PTB-XL: Assumed negative labels – sourced from a non-endemic country. No label smoothing applied.

To address class imbalance, we oversampled positive cases during minibatch construction to achieve an effective prevalence of 5%. We hypothesise this works well because it matches the challenge score metric's threshold. Both the label smoothing and oversampling were optimised based on internal test set challenge scores.

We found an ensemble of models performed optimally and used performance on the internal validation and internal test sets to select the optimal set of hyperparam-

Hyperparameter	Base Transformer	Large Transformer	xECG
frequency	400	150	100
patch size	256	128	25 (×12)
input signal length	7.32 s	7.32 s	10 s
total patches	144	84	40
mask ratio	0.80	0.80	0.30
number of heads	12	12	4
layers	12	16	9
dimensions	768	960	1024
projected-dim	3072	3840	2048
register tokens	4	4	0
parameters	86M	177M	57M
decoder layers	4	4	-
decoder dimensions	384	384	-
decoder projected-dim	1536	1536	-
decoder nhead	6	6	-
decoder parameters	7M	7M	-
epochs	120	200	50
warmup epochs	1	1	5
learning rate	0.0001	0.001	0.0001
weight decay	0.05	0.05	0.04
final wd	0.05	0.05	0.4
gradient clipping	0.5	0.5	3.0
batch size	2048	1024	512
pretraining time (H100)	3 days	13 days	3 days
initial EMA	-	-	0.99
final EMA	-	-	1.0
global crops	-	-	2
global crop size	-	-	0.8
local crops	-	-	4
local crop size	-	-	0.4

Table 1. SSL hyperparameters. projected-dim refers to the number of dimensions in the MLP/mLSTM layers.

Hyperparameter	Model 1	Model 2	Model 3	Model 4
pretrained model	B-Transformer	L-Transformer	xECG	xECG
max epochs	12	40	12	12
early stopping patience	5	8	5	5
monitor metric	AUPRC	Challenge	AUPRC	AUPRC
warmup epochs	1	1	1	1
learning rate (lr)	0.00005	0.00005	0.00001	0.00001
linear lr	0.00005	0.00005	0.0001	0.0001
hard restarts	0	4	0	0
hard restart decay	-	0.5	-	-
weight decay	0.05	0.05	0.05	0.05
drop path	0.2	0.5	0.5	0.5
layer-wise lr decay	0.75	0.75	0.75	0.75
gradient clipping	0.5	0.5	0.5	0.5
batch size	112	112	128	128
final layernorm	-	-	True	False
batchnorm output	-	-	False	True

Table 2. Finetuning hyperparameters. Model 2 is trained twice with different random seeds for the train/val split.

eters (Table 2). An ensemble of 5 models, where the logit outputs are averaged, gave a good trade-off between training/inference time and performance. All models were trained with a cosine annealing learning rate, with Model 2 using hard restarts once the learning rate had reduced to 0. In each finetuning run, the final model is the model which achieved the highest validation performance by either area under the precision recall curve (AUPRC) or the challenge metric (Table 1).

To reduce overfitting, regularisation was applied. We used a linearly decaying drop path [19], where transformer/xLSTM blocks are randomly skipped during training. Similarly, a layer-wise learning rate decay was applied so that larger updates were made to later layers.

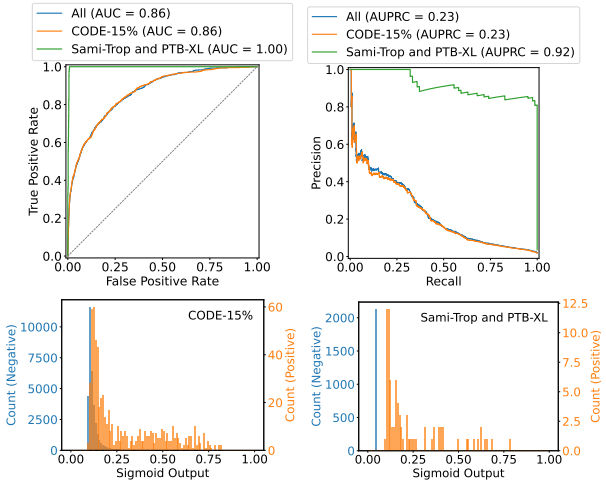


Figure 2. AUC (top left), AUPRC (top right) and predictions (bottom) for Model 1 on the internal test set.

Training	Validation	Test	Ranking
-	0.440	0.283	2/40

Table 3. Challenge scores for our selected entry, including the ranking of our team (DlaskaLabMUI) on the hidden validation set. We used repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

3. Results

Our method ranked 2nd on the challenge leaderboard with a score of 0.283 on the hidden test set.

We found performance improvements on the internal test set, mainly based on CODE-15%, did not translate well to the hidden validation set. Table 4 contains the results for several individual models and ensembles on the internal test set and hidden validation set. The selected ensemble had a challenge score 0.023 higher than an ensemble of five Model 1s, but only an increase of 0.004 on the hidden validation set.

To evaluate the importance of ensembling, we submitted an ensemble of five Model 1s and saw an increase of 0.026 on the hidden validation set. However, the small increase in internal test set challenge score when ensembling Model 2 suggests this increase will depend on the underlying model. By choosing multiple different hyperparameters and architectures we aim to increase the diversity within the ensemble and improve performance.

Figure 2 details the results for Model 1 on the internal test set which achieved an area under the receiver operator characteristic curve (AUC) of 0.86. Notably, the model was near perfect for the trusted labels in Sami-Trop and PTB-XL with an AUC of 1.00 and AUPRC of 0.92. The effect of the soft labels for CODE-15% can be seen in the distributions of sigmoid outputs in Figure 2.

Model Name	Training	Internal Test	Validation
Model 1	0.403 ± 0.022	0.475	0.410
Model 1 Ensemble	-	0.489	0.436
Model 2	0.458 ± 0.017	0.484	-
Model 2 Ensemble	-	0.490	-
Selected Ensemble	-	0.511	0.440

Table 4. Challenge scores on the internal test, hidden validation, and 5-fold cross validated training sets.

4. Discussion and Conclusions

The results suggest that large-scale SSL pretraining of both transformers and xLSTMs can effectively capture ECG structure and generalise across datasets with noisy labels. However, SSL tends to scale with dataset size and diversity [20] so future work should explore expanding the pretraining across more datasets. Fine-tuning with dataset-specific soft labels improved performance, particularly on the datasets for which we had reliable labels. Oversampling positives helped mitigate extreme class imbalance, although often led to overfitting in development which required use of early stopping criteria and regularisation.

The 1st place team (Biomed-Cardio) used a similar transformer-based architecture and masked autoencoding pretraining to our transformer models [21], but with two key additions: (1) multi-layer feature extraction and (2) inclusion of patient information, i.e., age and sex. Future work should explore the benefits of using patient metadata and whether it is more or less susceptible to data drift.

We present an ensemble taking advantage of the different representations of attention-based transformers and recurrent xLSTMs. Through self-supervised learning and label-noise-aware fine-tuning we address noisy supervision and class imbalance in Chagas disease classification.

Acknowledgments

We thank the authors of CODE [4] for making their dataset available.

References

[1] Reyna MA, Koscova Z, Pavlus J, Weigle J, et al. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025. In: CinC 2025. vol. 52; 2025. p. 1-4.

[2] Reyna MA, Koscova Z, Pavlus J, Saghafi S, et al. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025. arXiv preprint. 2025. arXiv:2510.02202.

[3] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):e215-20.

[4] Ribeiro ALP, Paixão GMM, Gomes PR, Ribeiro MH, et al. Tele-electrocardiography and bigdata: The CODE (Clinical Outcomes in Digital Electrocardiography) study. J Electrocardiol. 2019;57S:S75-8.

[5] Alkmim MB, Figueira RM, Marcolino MS, Cardoso CS, et al. Improving patient access to specialized health care: The Telehealth Network of Minas Gerais, Brazil. Bulletin of the WHO. 2012;90(5):373-8.

[6] Zheng J, Zhang J, Danioko S, Yao H, et al. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Sci Data. 2020;7(1):48.

[7] Zheng J, Chu H, Struppa D, Zhang J, et al. Optimal multi-stage arrhythmia classification approach. Sci Rep. 2020;101:1-17.

[8] Ribeiro A, Ribeiro M, Paixão G, Oliveira D, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Comm. 2020;11(1):1760.

[9] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data. 2020;7:154.

[10] Cardoso C, Sabino E, Oliveira C, de Oliveira L, et al. Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. BMJ Open. 2016;6(5):e0011181.

[11] Jidling C, Gedon D, Schön TB, Oliveira CDL, et al. Screening for Chagas Disease from the electrocardiogram using a deep neural network. PLoS Negl Trop Dis. 2023;17(7):e0011118.

[12] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, et al. An image is worth 16x16 words: transformers for image recognition at scale. ICLR. 2021.

[13] Beck M, Pöppel K, Spanring M, Auer A, et al. xlstm: Extended long short-term memory. NeurIPS. 2024.

[14] He K, Chen X, Xie S, Li Y, et al. Masked autoencoders are scalable vision learners. CVPR. 2021.

[15] Na Y, Park M, Tae Y, Joo S. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. arXiv preprint. 2024. arXiv:2402.09450.

[16] Alkin B, Beck M, Pöppel K, Hochreiter S, et al. Vision-LSTM: xLSTM as generic vision backbone. ICLR. 2025.

[17] Lunelli R, Nicolson A, Pröll SM, Reinstadler SJ, et al. BenchECG and xECG: a benchmark and baseline for ECG foundation models. arXiv preprint. 2025. arXiv:2509.10151.

[18] Wu Z, Zhang J, Pai D, Wang X, et al. Simplifying DINO via coding rate regularization. arXiv preprint. 2025. arXiv:2502.10385.

[19] Huang G, Sun Y, Liu Z, Sedra D, et al. Deep networks with stochastic depth. ECCV. 2016:646-61.

[20] Siméoni O, Vo HV, Seitzer M, Baldassarre F, et al. DINOv3. arXiv preprint. 2025. arXiv:2508.10104.

[21] Santvliet LV, Nguyen PX, Vandenberk B, Vos MD. Detecting Chagas Disease using a vision transformer–based ECG foundation model. In: CinC 2025. vol. 52; 2025. p. 1-4.

Address for correspondence:

Angus Nicolson
angus.nicolson@i-med.ac.at

Clemens Dlaska
clemens.dlaska@i-med.ac.at