SwissBeatsNet: A Multilead Masked Autoencoder for Chagas Disease Detection

Lucas Erlacher[1*], Andrea Agostini[1*], Samuel Ruiperez-Campillo[1*],
Ece Ozkan[1,2†], Thomas M. Sutter[1†], Julia E. Vogt[1†]

[1]Department of Computer Science, ETH Zurich, Switzerland
[2]Department of Biomedical Engineering, University of Basel, Switzerland

[*] Shared first authorship; [†] Shared last authorship

Abstract

Chronic Chagas Disease is a parasitic cardiomyopathy often causing arrhythmias, conduction defects, and heart failure, yet early ECG diagnosis remains difficult, especially in low-resource. We propose SwissBeatsNet, a multilead masked-autoencoder (MLMAE) framework that treats ECGs as synchronized channels to learn intra-lead temporal dynamics and inter-lead spatial dependencies. Self-supervised pretraining using CODE-15%, SaMi-Trop, and PTB-XL datasets reconstructs randomly masked windows while incorporating an alignment objective to enhance representation learning. We then freeze the encoder and train an ensemble of linear heads to predict a Chagas disease score. On the hidden Physionet 2025 test set, the selected SwissBeatsNet model achieves a score of 0.425. We ranked the 4th position as a team on the leaderboard.

1. Introduction

Chagas disease, caused by the protozoan *Trypanosoma cruzi*, remains a major public health challenge in Latin America [1], and is increasingly detected in non-endemic regions such as North America and Europe [2]. A common chronic infection manifestation is chronic Chagas cardiomyopathy (CCC), a parasitic cardiomyopathy that often leads to conduction disease, arrhythmias, and heart failure [3]. Early detection of cardiac involvement is essential yet difficult, specially in primary care and resource-limited settings, because routine electrocardiogram (ECG) interpretation lacks sensitivity to detect subtle early abnormalities [4].

Recent advances in self-supervised learning (SSL) enable the extraction of informative representations from large unlabeled corpora, reducing dependence on expert annotations and potentially improving transfer to downstream tasks. Within SSL, contrastive learning (CL) aligns semantically related segments while separating unrelated ones to structure latent spaces [5, 6]. Time-series masked autoencoders (MAEs) [7] could complement this by reconstructing segments to capture local and global context. Applied to multichannel ECGs [8], MAE-style objectives can leverage cross-lead synchrony; however, many existing approaches still treat leads independently, ignoring meaningful joint temporal–spatial dependencies.

In Chagas disease, automated ECG analysis has evolved from handcrafted features and classical classifiers [9] to deep learning using CNN/LSTM models trained on raw signals [10]. Despite promising, these methods often require substantial labeled data and may generalize poorly across populations and acquisition protocols.

In this work, we present *SwissBeatsNet*, a multilead MAE (MLMAE) framework for Chagas disease detection from multi-lead ECGs. Treating each lead as a synchronized channel in a multivariate time series, our model jointly learns of intra-lead temporal dynamics and inter-lead spatial dependencies. During self-supervised pretraining on a combination of the CODE-15%[11], SaMi-Trop[12], and PTB-XL [13] datasets, SwissBeatsNet reconstructs randomly masked temporal windows (inspired by MAEs [14]) and employs an auxiliary alignment objective to sharpen representations and mitigate shortcut learning.

Our contributions are threefold: (i) joint spatio-temporal modeling across ECG leads within a unified MLMAE backbone; (ii) integration of MAE reconstruction with alignment regularization to couple context modeling and discriminative structure; and (iii) multi-dataset pretraining to enhance robustness and cross-cohort generalization.

This study is part of the George B. Moody PhysioNet Challenge 2025 [15], hosted on PhysioNet [16].

2. Methods

2.1. Data Sources

Let $\mathcal{X} = \{\mathbf{X}^{(i)}\}_{i=1}^{N_D}$ describe a dataset consisting of N_D ECG recordings where $\mathbf{X}^{(i)} \in \mathbb{R}^{|\mathbb{L}| \times N_L}$ is a multi-lead ECG signal of N_L samples with $\mathbb{L}$ being the set of available leads. Hence, we have $\mathbf{X}^{(i)} = \{\mathbf{x}_l^{(i)}\}_{l \in \mathbb{L}}$. In this work, we used three corpora: CODE-15%–a subset of 350,000 Brazil-

ian primary-care ECGs, 10s at 400 Hz, annotated with diverse cardiac diagnoses [11]; SaMi-Trop–a Chagas-specific cohort from Minas Gerais, with confirmed seropositive participants and clinical metadata, 10s at 300 Hz, acquired via the Brazilian Telehealth Network [12]; and PTB-XL–a public German dataset spanning a broad diagnostic spectrum, 10s at 100 or 500 Hz, annotated by cardiologists, used to increase morphological diversity during pretraining [13]. For both pretraining and finetuning, we pooled all three datasets and harmonized signals by resampling to 250 Hz and fixing the length to N_L =2,250 samples via trimming or symmetric zero-padding; each lead then underwent a 0.5 Hz high-pass Butterworth filter, 50 Hz powerline removal, and lead-wise z-score normalization. For the given datasets, $\mathbb{L} = \{l_I, l_{II}, l_{III}, l_{a_{VR}}, l_{a_{VL}}, l_{a_{VF}}, l_{V_1}, ..., l_{V_6}\}$.

2.2. SwissBeatsNet Architecture

We develop an SSL framework based on MAEs [14] tailored to multilead ECGs. SwissBeatsNet combines masked-signal reconstruction with an alignment objective [17] to capture intra-lead temporal dynamics and inter-lead spatial relationships. The pipeline has two stages: (1) self-supervised pretraining on unlabeled ECGs, and (2) supervised finetuning with a binary Chagas classifier–see overview in Figure 1.

2.2.1. Pretraining Stage

The proposed model follows the vision transformer architecture [18] and the pretraining introduced in [17], i.e., the reconstruction of missing input patches based on available input patches using an encoder E_ϕ and a decoder D_θ. We first split each lead signal into P non-overlapping patches $\mathbf{s}_{l_p}$ such that $\mathbf{x}_l = [\mathbf{s}_{l_1}, \ldots, \mathbf{s}_{l_P}]$ and subsequently transform the patches to d_t-dimensional tokens using a learnable linear projection. For this, we randomly mask $m = 90\%$ of input tokens and only input the remaining tokens to the encoder E_ϕ. The decoder D_θ predicts the sample values of the masked tokens based on the encoded input tokens. Both the encoder E_ϕ and the decoder D_θ are shared across the individual leads, i.e., there is only one encoder and one decoder, and the leads are reconstructed independently.

The pretraining loss $\mathcal{L}$ becomes: $\mathcal{L} = \mathcal{L}_R + \beta\mathcal{L}_A$, where a scaling parameter β controls the contribution of the contrastive term to the overall loss.

Reconstruction Objective. The pretraining objective minimizes the following reconstruction loss $\mathcal{L}_\mathrm{R}$:

$$\mathcal{L}_R = \frac{1}{|\mathbb{L}|}\frac{1}{N_L}\sum_{l\in\mathbb{L}}\sum_{p=1}^{P}||\mathbf{s}_{l_p} - \hat{\mathbf{s}}_{l_p}||_2^2, \tag{1}$$

where $\hat{\mathbf{s}}_{l_p}$ is the reconstruction of the respective input patch.

Alignment Loss. To encourage cross-lead consistency within the same record, we regularize with $\mathcal{L}_\mathrm{A}$ [17], instantiated either as MSE between per-lead embeddings or as a contrastive objective.

For the MSE loss, i.e., $\mathcal{L}_A = \mathcal{L}_{MSE}$, we calculate the average mean squared error across all pairs of per-lead embeddings, where each per-lead embedding $\mathbf{e_l}$ is obtained by averaging the token embeddings of the lead, i.e., $\mathbf{e}_l = \frac{1}{P}\sum_{p=1}^{P} E_\phi(\mathbf{x}_{l_p})$:

$$\mathcal{L}_{MSE} = \frac{1}{|\mathbb{L}|^2}\frac{1}{d_t}\sum_{l,k\in\mathbb{L}}||\mathbf{e}_l - \mathbf{e}_k||_2^2. \tag{2}$$

The contrastive loss, i.e., $\mathcal{L}_A = \mathcal{L}_C$, is computed as in [17]:

$$\mathcal{L}_C(\mathcal{X}) = -\frac{1}{2N_D}\sum_{l\in\mathbb{L}}\sum_{i=1}^{N_D}\Gamma(\mathcal{X}, l, i), \tag{3}$$

where

$$\Gamma(\mathcal{X}, l, i) = \log\left(\frac{\Lambda(\mathbf{e}_l^{(i)}, \mathbf{e}_{\bar{l}}^{(i)})}{\sum_{\substack{k=1\\k\neq i}}^{N_D}\Lambda(\mathbf{e}_l^{(i)}, \mathbf{e}_l^{(k)}) + \Lambda(\mathbf{e}_l^{(i)}, \mathbf{e}_{\bar{l}}^{(k)})}\right),$$

where $\Lambda(\mathbf{e}_l^{(i)}, \mathbf{e}_{\bar{l}}^{(i)}) = \exp\left(\mathrm{sim}(\mathbf{e}_{\mathrm{l}}^{(\mathrm{i})}), \mathbf{e}_{\bar{\mathrm{l}}}^{(\mathrm{i})})/\tau\right)$, $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity, τ a temperature parameter, and $\bar{l}$ defines a lead not equal to l. All pairs of leads from the same ECG recording i are treated as a positive pair. Embeddings from different studies $j \neq i$, regardless of the lead, serve as negative samples.

2.2.2. Finetuning Stage

We adapt the pretrained, label-agnostic encoder E_ϕ to Chagas prediction via either of the following two classification networks. *(i) Per-lead ensemble:* per-lead heads H_{ϕ_l} consume the embedding of the [CLS] token and apply a linear projection (halving dimensionality), ReLU, dropout, and a final linear layer to produce lead-wise disease scores - the record-level score is then given by the mean over all leads-wise scores. *(ii) Fused MLP:* concatenated per-lead [CLS] embeddings are transformed into a record-level disease score by a 4-layer MLP (where each layer halves dimensionality and is composed of linear projection, ReLUm and dropout).

2.3. Training Details

All runs (pretraining and finetuning) used batch size 128, AdamW, and a multi-GPU Distributed Data Parallel strategy. In the pretraining part, we trained ViT-Tiny and ViT-Base [18]. ViT-Tiny ran 120 epochs with linear warm-up of the learning rate (LR) followed by cosine decay; we use a

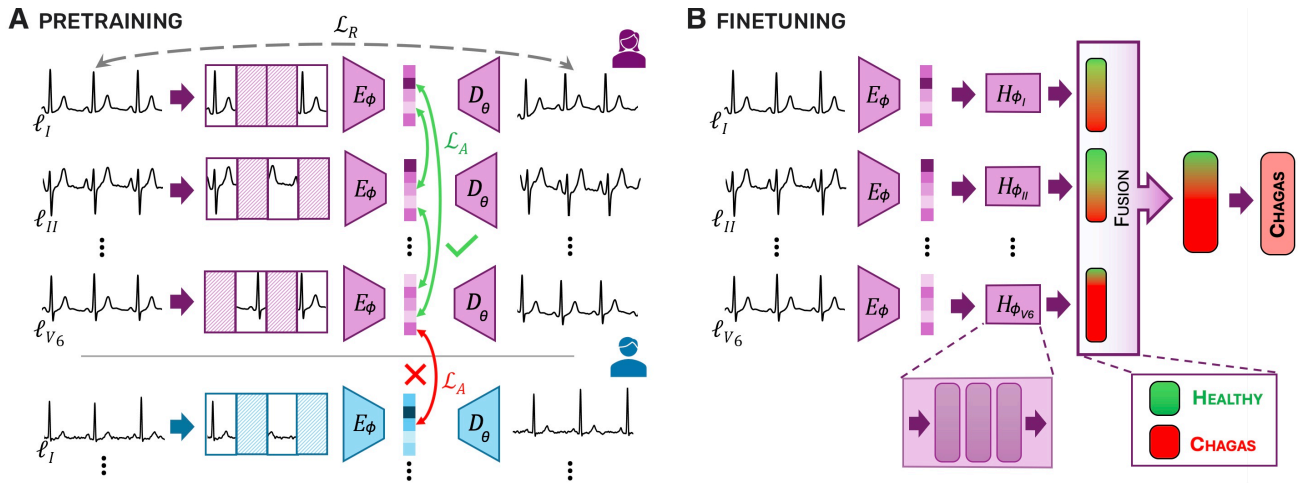

Figure 1. Overview of SwissBeatsNet architecture. Pretraining is shown on the left and downstream finetuning on the right.

sigmoid schedule to anneal β from 0 to 1. We trained ViT-Base in total for 720 epochs. For the finetuning part, the classifier used a binary cross-entropy loss and a minority oversampling strategy to handle the class imbalance present in the datasets. We evaluated both a frozen-encoder setup (classifier training only) and finetuning of all parameters.

3. Results

Evaluation Protocol. We split the combined dataset into non-overlapping subsets at the record level, with 80% used for the training of the model, 10% for tracking overfitting during training (S_1), and 10% for assessing the generalization error of the trained model (S_2). During pretraining, we monitored reconstruction loss on both the training set and S_1 to assess potential overfitting. To evaluate the learned representations, we conducted linear probing on the encoder using six auxiliary (non-Chagas) diagnostic labels available in the CODE-15% dataset, where the mean AUROC across these labels served as a proxy for the quality of the representations. In the finetuning phase, we tracked binary cross-entropy loss and per-class (Chagas vs. non-Chagas) accuracy on the training set and S_1. Final predictive performance on the Chagas classification task was evaluated using the challenge score on S_1 and S_2 to assess generalization.

Implementation Details. We conducted two self-supervised pretraining runs: ViT-Tiny (patch size 90) using the alignment loss from Eq.3, and ViT-Base (patch size 50) using the loss from Eq.2. Key configuration values for the finetuning runs are summarized in Table 1. Note that in submission 2815, the ensemble's classification heads were restricted to a single linear layer. Table 1 furthermore reports model performance on the Chagas task across S_1 and S_2, with the final column displaying the challenge score on the official leaderboard set.

Results Analysis. Among our experiments, end-to-end finetuning achieved high specificity, but was outperformed overall by approaches that froze the encoder and trained only lightweight classification heads. Notably, submissions 2813 and 2816 (referring to official Challenge submission IDs) yielded the strongest results across multiple metrics. Submission 2813 excelled in sensitivity and demonstrated strong generalization from S_1 to S_2, while 2816 achieved strong specificity and slightly higher Challenge Scores on both S_1 and the leaderboard test set, despite using the smaller ViT Tiny backbone. However, given it's more stable generalization between S_1, S_2, and leaderboard test set, and its high sensitivity - which closely aligns with the Challenge Score objective - we ultimately decided to select submission 2813 for evaluation on the final test set. The relevant scores of this model are summarized in Table 2.

4. Discussion and Conclusion

Our findings highlight the promise of *SwissBeatsNet* for detecting Chagas disease from 12-lead ECGs, particularly in low-resource settings where early diagnosis remains a challenge. Although the achieved challenge score (0.425) reflects task difficulty, the model's ability to learn general cardiac patterns through self-supervised pretraining enhances its sensitivity to subtle pathological changes indicative of early-stage CCC. The combination of masked autoencoder pretraining with an alignment objective enables the model to capture both intra-lead dynamics and inter-lead spatial dependencies. This multilead design, trained across heterogeneous datasets, supports better generalization across populations and acquisition conditions, an essential feature for real-world deployment.

Several limitations remain. Dataset biases, such as demographic and equipment variability, may influence model

ID	ViT	Classifier	Finetuning	Base LR	Epochs	AUROC		Sens.		Spec.		Score		
						S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2	Validation
2812	Base	Fused	Heads	1e-3	3	*0.87*	**0.86**	0.63	*0.65*	0.91	0.91	*0.49*	**0.53**	0.314
2813	Base	Ensemble	Heads	1e-3	1	*0.87*	**0.86**	**0.69**	**0.70**	0.86	0.86	0.45	*0.50*	*0.425*
2815	Base	Ensemble	All	1e-5	5	0.85	*0.84*	0.52	0.57	*0.92*	*0.92*	0.42	0.46	0.352
2816	Tiny	Ensemble	Heads	3e-3	15	**0.92**	0.83	*0.68*	0.47	**0.94**	**0.93**	**0.56**	0.40	**0.430**

Table 1. Key parameters and downstream performance of different submissions. "ID" refers to the leaderboard ID, "Sens." to Sensitivity, "Spec." to Specificity, "Score" to Challenge Score and "Validation" to the leaderboard score. The highest score of each column is marked in bold and the second highest in italics.

Training	Val.	Test	Val. Ranking	Test Ranking
0.500	0.425	-	4^{th}	-

Table 2. Challenge Scores of the selected entry. "Training" refers to the score on S_2, "Val." to the leaderboard score, and "Test" to the score of the final evaluation, to be filled after results are announced.

behavior. Additionally, our retrospective evaluation calls for prospective clinical validation to confirm its effectiveness in routine care. Future directions include integration with point-of-care ECG devices to support early screening in endemic regions and extending the framework to multi-disease detection for broader clinical impact.

In summary, this work introduces the first MMAE-based multilead ECG framework for Chagas detection, leveraging large-scale self-supervised pretraining and targeted finetuning. SwissBeatsNet demonstrates the feasibility of scalable, generalizable ECG-based diagnostics with meaningful public health implications.

Acknowledgements

The authors acknowledge funding from ETH Zurich, University of Basel, Innosuise, the Swiss AI Initiative, and the Swiss National Supercomputer Center in Switzerland.

References

[1] WHO. Chagas disease (also known as american trypanosomiasis), April 2025.

[2] Albajar-Viñas P, Jannin J. The hidden chagas disease burden in Europe. Eurosurveillance September 2011;16(38).

[3] Nunes MCP, Beaton A, Acquatella H, et al. Chagas cardiomyopathy: An update of current clinical knowledge and management: A scientific statement from the American Heart Association. Circulation September 2018;138(12).

[4] Rojas LZ, Glisic M, Pletsch-Borba L, et al. Electrocardiographic abnormalities in chagas disease in the general population: A systematic review and meta-analysis. PLOS Neglected Tropical Diseases June 2018;12(6):e0006567.

[5] Kiyasseh D, Zhu T, Clifton DA. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In International Conference on Machine Learning. PMLR, 2021; 5606–5615.

[6] Soltanieh S, Etemad A, Hashemi J. Analysis of augmentations for contrastive ECG representation learning. In 2022 Int Joint Conf on Neural Networks. 2022; 1–10.

[7] Chen J, Wu W, Liu T, et al. Multi-channel masked autoencoder and comprehensive evaluations for reconstructing 12-lead ECG from arbitrary single-lead ECG. npj Cardiovascular Health December 2024;1(1).

[8] Na Y, Park M, Tae Y, et al. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram, 2024.

[9] Sady CC, Ribeiro ALP. Symbolic features and classification via support vector machine for predicting death in patients with chagas disease. Computers in Biology and Medicine March 2016;70:220–227.

[10] Jidling C, Gedon D, Schön TB, et al. Screening for chagas disease from the electrocardiogram using a deep neural network. PLOS Neglected Tropical Diseases July 2023; 17(7):e0011118.

[11] Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications April 2020;11(1).

[12] Cardoso CS, Sabino EC, Oliveira CDL, et al. Longitudinal study of patients with chronic chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. BMJ Open May 2016;6(5):e011181.

[13] Wagner P, Strodthoff N, Bousseljot RD, et al. PTB-XL, a large publicly available electrocardiography dataset, 2022.

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

[15] Reyna MA, Koscova Z, Pavlus J, et al. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025. Computing in Cardiology 2025;52:1–4.

[16] Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[17] Agostini A, Laguna S, Ryser A, et al. Leveraging the structure of medical data for improved representation learning, 2025.

[18] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. CoRR 2020;abs/2010.11929.

Address for correspondence:

Samuel Ruiperez-Campillo (sruiperez@inf.ethz.ch)
Universitätstrasse 6, 8092 Zurich