Domain-Adversarial Pretrained Encoder for ECG-Based Chagas Disease Screening

Tianzheng Dong[1], Xinqi Bao[1], Jia Bi[2], Saikat Chatterjee[1]

[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]Rutherford Appleton Laboratory, United Kingdom

Abstract

Chagas disease (ChD) remains a major health burden in Latin America, and scalable screening tools are required to pre-select high-risk patients for confirmatory testing. Electrocardiograms (ECGs) are widely available, but machine learning models trained on heterogeneous datasets are vulnerable to domain bias. A domain-aware framework was proposed using a transformer encoder initialized with released pretrained weights, combined with a four-class reformulation of SaMi-Trop and PTB-XL cohorts and a domain-adversarial head to promote domain-invariant features. In local five-fold cross-validation, the model achieved an average challenge score (Recall@5%) of 0.824 (best 0.852, with AUROC/AUPRC of 0.883/0.852). Predicted probabilities remained conservative, avoiding extreme confidence while maintaining strong ranking. On the hidden external test set, the challenge score decreased to 0.270, indicating sensitivity to domain shift. These results demonstrate that the transformer encoder structure with multi-class adversarial training improve local robustness, but domain-aware validation and dataset diversification are necessary for generalizable ChD ECG screening.

1. Introduction

Chagas disease (ChD) is a neglected tropical disease caused by the parasite *Trypanosoma cruzi*, primarily endemic in Latin America but increasingly global due to human migration [1]. An estimated 6–8 million people are affected worldwide, and about 30% develop chronic cardiac complications contributing to morbidity and mortality [2, 3]. The diagnostic gold standard is serological testing, but its high cost, limited accessibility, and low throughput restrict large-scale deployment [4]. Electrocardiograms (ECGs) are routine, low-cost, and non-invasive, and ChD frequently manifests with conduction abnormalities such as right bundle branch block, prolonged QRS, and ST–T changes [5, 6]. These characteristics make ECG a promising modality for scalable screening, motivating machine learning approaches for automated detection.

Prior work on ECG-based ChD screening has followed two main directions: pipelines using handcrafted features with conventional classifiers [7–9], and end-to-end convolutional neural networks (CNNs) trained directly on raw signals [4, 10]. Reported local accuracies were often high at around 90%, with AUROC in the range of 0.80–0.97 [8–10]. However, performance typically attenuates under external validation due to dataset heterogeneity and inconsistent evaluation protocols. To address this limitation, the PhysioNet/Computing in Cardiology Challenge 2025 provided three open-source 12-lead ECG datasets, offering a common benchmark for fair and reproducible evaluation of Chagas screening algorithms. Nevertheless, strong-label cohorts contain only confirmed ChD cases, while healthy controls must be drawn from external databases. This coupling of disease status and dataset origin introduces domain bias that can confound model training, underscoring the need for domain-robust methods.

This study presents a domain-aware framework, based on a pretrained transformer encoder in [11]. A four-class reformulation and adversarial training are involved to mitigate shortcut learning from dataset artifacts. The approach highlights label–source coupling as a primary confounder and demonstrates improved robustness for ECG-based ChD screening across heterogeneous databases.

2. Methodology

2.1. Database and Pre-processing

The training data in this study come from the three databases specified by the Challenge: CODE-15% [12], SaMi-Trop [13], and PTB-XL [14]. These databases vary substantially in number of records, recording length, sampling frequency, and the reliability of diagnostic labels. Their key characteristics are summarized in Table 1.

In this study, the CODE-15% dataset was excluded because its Chagas disease labels are based on self-reports and therefore lack reliable verification. By contrast,

SaMi-Trop labels are confirmed by serological testing and were regarded as high-confidence annotations. For PTB-XL, negative cases were retained despite the absence of serological confirmation, given the negligible background prevalence of Chagas disease in Europe. Consequently, all SaMi-Trop and PTB-XL records were included in the current training.

Table 1. Summary of the Challenge training datasets.

Dataset	CODE-15%	SaMi-Trop	PTB-XL
Recordings	343,424	1,631	21,799
Duration (s)	7.3 - 10.2	7.3 - 10.2	10
Fs(Hz)	400	400	500
#Pos	6,561	1,631	0
#Neg	336,863	0	21,799
Label strength	weak	strong	strong

To harmonize the heterogeneous sources, a standardized pre-processing pipeline was applied. Eight standard leads (I, II, and V1–V6) were retained in fixed order, and a 7.0 s centered segment was extracted from each record. This selection reduces data redundancy and computational load while retaining all linearly independent information, since leads III, aVR, aVL, and aVF can be derived from I and II [15]. Signals were band-pass filtered between 0.5 and 45 Hz, resampled to 500 Hz, and standardized by z-score normalization. During training, random 5-s crops were used, whereas validation and inference relied on centered windows. Records with incomplete signals or that failed filtering were excluded.

2.2. Multi-class scheme and Domain Adversarial Neural Network (DANN)

Preliminary experiments showed that direct binary ChD classification led to rapid overfitting despite unified pre-processing, indicating hidden discrepancies between datasets, likely from acquisition hardware or electrode placement [16, 17]. To mitigate this, a Domain Adversarial Neural Network (DANN) [18] with a gradient reversal layer (GRL) was used to encourage domain-invariant features by jointly optimizing task and domain objectives.

Because ChD-positive and negative samples originate exclusively from SaMi-Trop and PTB-XL, direct adversarial training risks suppressing true disease features. To decouple this, a four-class scheme was adopted: both datasets were split into `normal_ecg` and `abnormal_ecg` based on expert annotations. This allows adversarial training to target dataset bias while forcing the encoder to capture clinically meaningful ECG morphology.

The shared encoder produces a window-level representation, with two linear heads on top: a task head projecting to the four-class label space, and a domain head predicting dataset source via the GRL. Table 2 summarizes the distribution of the training classes.

Table 2. Distribution of the four classes.

Class	Count	Proportion
`CH_NORM`	286	1.22%
`CH_ABN`	1345	5.74%
`NC_NORM`	9514	40.61%
`NC_ABN`	12285	52.43%

2.3. Encoder and Domain-Adversarial Heads

The proposed model comprises a shared transformer encoder and two lightweight classification heads (Fig. 1). Each 8-lead ECG window was divided into non-overlapping 50-sample patches, producing 400 tokens in total. A linear projection mapped each patch to a 768-dimensional embedding, and a two-dimensional sinusoidal positional encoding over the (lead $\times$ time) grid was added to preserve temporal and spatial order.

The encoder consists of 12 transformer blocks with 16-head self-attention and feed-forward networks (expansion ratio 4), using pre-norm LayerNorm. A structured attention mask was applied, permitting unrestricted temporal interactions within each lead and synchronized cross-lead interactions at the same time index. After the final block, token-wise mean pooling produced a fixed-length 768-dimensional representation $\mathbf{z}$ for the ECG window.

On top of the encoder, two linear heads operate in parallel. The task head maps $\mathbf{z}$ to logits for four ECG classes, while the domain head receives a copy of $\mathbf{z}$ through a GRL and predicts dataset source. The GRL behaves as identity in the forward pass but scales encoder-side gradients by $-\lambda$ during backpropagation, with λ increasing along a logistic ramp up to 0.5. At evaluation, $\lambda = 0$, disabling the adversarial pathway. Both heads are single linear layers to concentrate representational capacity in the encoder.

2.4. Training and Evaluation

The transformer encoder was initialized with weights pretrained in a JEPA framework [11], which used self-supervised learning on more than 170k quality-controlled ECGs from multiple public datasets. This initialization provides an inductive bias toward ECG morphology and supports transferable representation learning.

To address class imbalance, weighted sampling and class-specific loss weights were applied, inversely proportional to empirical class frequencies and normalized to unit mean across classes. Fine-tuning was performed with all parameters trainable for up to 10 epochs using the AdamW optimizer. The learning rate was 2.5×10^{-5} for the task

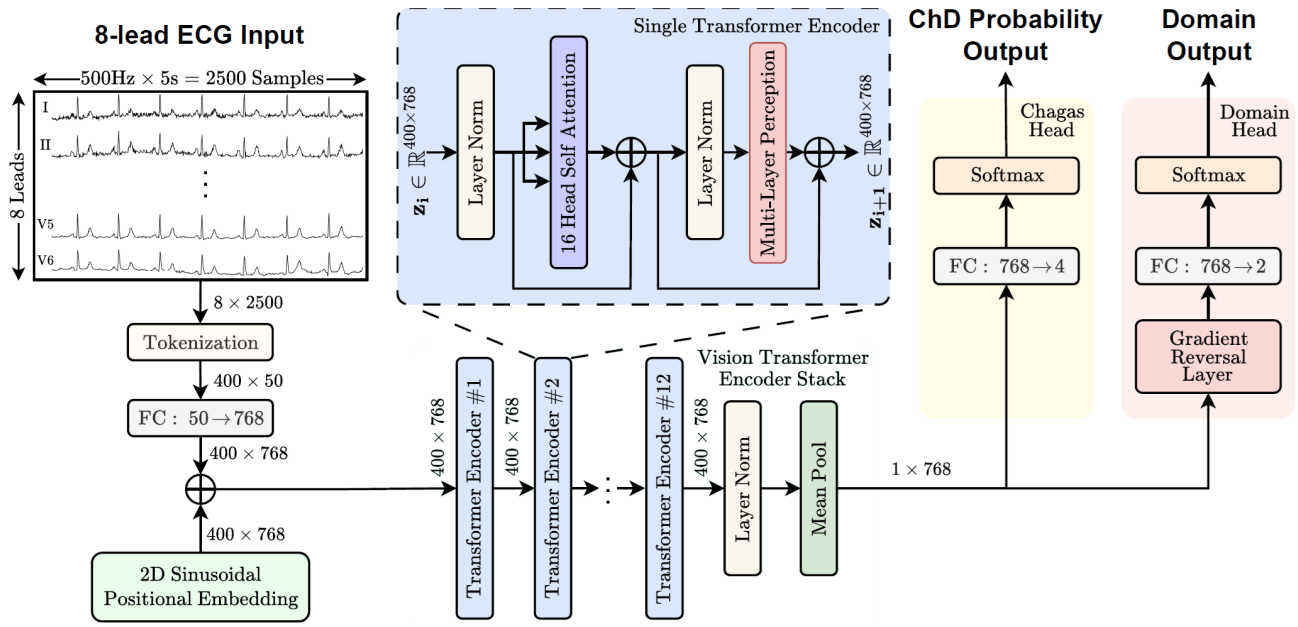

Figure 1. Overall architecture of the proposed model. '$\mathbf{FC} : \mathbf{x} \rightarrow \mathbf{y}$' denotes a fully connected layer mapping an $\mathbf{x}$-dimensional input to $\mathbf{y}$ output units with a bias term included.

and domain heads, while encoder parameters were scaled by 0.1. A scheduler with three epochs of linear warm-up followed by cosine decay was used.

The gradient-reversal coefficient λ was gradually increased during training according to a logistic ramp:

$$\lambda(p) = \frac{2}{1 + \exp(-10p)} - 1,$$

where p denotes normalized training progress. The maximum value of λ was 0.5, applied both in the GRL and as the domain loss weight.

During inference, each record was divided into fixed-length windows. Inference used top-k averaging across windows ($k = 25\%$). For the four-class model, ChD probability was computed as a weighted sum of subclass probabilities ($0.2 \times \texttt{CH\_NORM} + 0.8 \times \texttt{CH\_ABN}$).

Model performance was monitored at each epoch using AUROC, average precision, and accuracy, while model selection was based on recall at the top 5% (Recall@5%) ChD probabilities, consistent with the Challenge metric. Early stopping with five-epoch patience was applied, and the final model was chosen by maximizing Recall@5% with low domain accuracy.

3. Results

The framework was evaluated by five-fold cross validation consistent with the Challenge protocol. Since the objective is to rank subjects by predicted ChD probability, the primary endpoint was the Challenge score, defined as recall among the top 5% of ranked subjects. AUROC and AUPRC were reported as secondary reference metrics.

Across folds, the model achieved a mean Challenge score of 0.824, with the best fold reaching 0.852. AUROC and AUPRC averaged 0.883 and 0.852, respectively, indicating strong ranking performance and consistent discrimination across folds.

Evaluation on the hidden external test set yielded a Challenge score of 0.270, reflecting the challenge of generalization under unseen domains.

4. Discussion and Conclusion

This study presented a domain-aware framework for ChD screening from 12-lead ECGs, integrating a pretrained transformer encoder, a four-class reformulation, and adversarial training. In local cross-validation, the model achieved strong ranking performance with conservative probability calibration, suggesting that adversarial learning can suppress dataset-specific artifacts while retaining clinically meaningful ECG patterns. However, performance dropped substantially on the external hidden test set, revealing limited robustness under domain shift.

Two factors likely explain this decline. First, ChD labels are tightly coupled with dataset provenance, creating residual confounding despite adversarial training and reflecting differences in acquisition devices, preprocessing, and cohort composition. Second, the scarcity of serology-confirmed ChD cases with normal ECGs, combined with treating non-ChD abnormalities as negatives, restricts the

ability to disentangle ChD-specific signatures from generic abnormalities.

Future work should address these issues by expanding dataset diversity, introducing augmentations that approximate acquisition variability, and adopting stronger domain adaptation strategies such as conditional DANN or multi-task objectives with auxiliary ECG diagnoses.

In conclusion, while the proposed framework improves in-domain calibration and reduces reliance on dataset artifacts, its external degradation underscores the need for source-aware validation and broader data coverage. Addressing these limitations will be critical for developing reliable, scalable ECG-based screening of Chagas disease.

Acknowledgments

This work was supported by Digital Futures, Stockholm, Sweden. The first author acknowledges support from the Digital Futures Summer Internship program, and the corresponding author is a Digital Futures Research Fellow.

References

[1] Gascon J, Bern C, Pinazo MJ. Chagas disease in spain, the united states and other non-endemic countries. Acta tropica 2010;115(1-2):22–27.

[2] Saraiva RM, Mediano MFF, Mendes FS, Sperandio da Silva GM, Veloso HH, Sangenis LHC, da Silva PS, Mazzoli-Rocha F, Sousa AS, Holanda MT, Hasslocher-Moreno AM. Chagas heart disease: An overview of diagnosis, manifestations, treatment, and care. World Journal of Cardiology 2021;13(12):654–675.

[3] Chadalawada S, Rassi Jr. A, Samara O, Monzon A, Gudapati D, Vargas Barahona L, Hyson P, Sillau S, Mestroni L, Taylor M, da Consolação Vieira Moreira M, DeSanto K, Agudelo Higuita NI, Franco-Paredes C, Henao-Martínez AF. Mortality risk in chronic chagas cardiomyopathy: a systematic review and meta-analysis. ESC Heart Failure December 2021;8(6):5466–5481. Epub 2021-10-30.

[4] Jidling C, Gedon D, Schön TB, Oliveira CDL, Cardoso CS, Ferreira AM, Giatti L, Barreto SM, Sabino EC, Ribeiro AL, et al. Screening for chagas disease from the electrocardiogram using a deep neural network. PLoS Neglected Tropical Diseases 2023;17(7):e0011118.

[5] Brito BOdF, Ribeiro ALP. Electrocardiogram in chagas disease. Revista da Sociedade Brasileira de Medicina Tropical 2018;51(05):570–577.

[6] Rojas LZ, Glisic M, Pletsch-Borba L, Echeverría LE, Bramer WM, Bano A, Stringa N, Zaciragic A, Kraja B, Asllanaj E, 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.

[7] Bao X, Hu F, Xu Y, Trabelsi M, Kamavuako E. Paroxysmal atrial fibrillation detection by combined recurrent neural network and feature extraction on ecg signals. In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies. 2022; 85–90.

[8] Hevia-Montiel N, Perez-Gonzalez J, Neme A, Haro P. Machine learning-based feature selection and classification for the experimental diagnosis of trypanosoma cruzi. Electronics 2022;11(5):785.

[9] Cornejo DR, Ravelo-García A, Alvarez E, Rodríguez MF, Díaz LA, Cabrera-Caso V, Condori-Merma D, Cornejo MV. Deep learning and permutation entropy in the stratification of patients with chagas disease. In 2022 Computing in Cardiology (CinC), volume 498. IEEE, 2022; 1–4.

[10] Hussein AF, Al-Neami AQ, Habash QA. Efficient deep learning for chagas cardiomyopathydetection: Data-driven ecg feature reduction. Al Furat Journal of Innovations in Engineering Applications 2025;1(1):18–18.

[11] Weimann K, Conrad TO. Self-supervised pre-training with joint-embedding predictive architecture boosts ecg classification performance. Computers in Biology and Medicine 2025;196:110809.

[12] Ribeiro AH, Paixao GMM, Lima EM, Horta Ribeiro M, Pinto Filho MM, Gomes PR, Oliveira DM, Meira Jr W, Schon TB, Ribeiro ALP. Code-15%: a large scale annotated dataset of 12-lead ecgs, 2021. URL https://doi.org/10.5281/zenodo.4916206.

[13] Ribeiro ALP, Ribeiro AH, Paixao GMM, Lima EM, Horta Ribeiro M, Pinto Filho MM, Gomes PR, Oliveira DM, Meira Jr W, Schon TB, Sabino EC. Sami-trop: 12-lead ecg traces with age and mortality annotations, 2021. URL https://doi.org/10.5281/zenodo.4905618.

[14] Wagner P, Strodthoff N, Bousseljot R, Samek W, Schaeffter T. Ptb-xl, a large publicly available electrocardiography dataset, 2022. URL https://doi.org/10.13026/kfzx-aw45. RRID:SCR_007345.

[15] Jaros R, Martinek R, Danys L. Comparison of different electrocardiography with vectorcardiography transformations. Sensors Jul 2019;19(14):3072. URL https://doi.org/10.3390/s19143072.

[16] Ong Ly C, Unnikrishnan B, Tadic T, Patel T, Duhamel J, Kandel S, Moayedi Y, Brudno M, Hope A, Ross H, et al. Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. NPJ digital medicine 2024;7(1):124.

[17] Bao X, Abdala AK, Kamavuako EN. Estimation of the respiratory rate from localised ecg at different auscultation sites. Sensors 2020;21(1):78.

[18] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V. Domain-adversarial training of neural networks. Journal of machine learning research 2016;17(59):1–35.

Address for correspondence:

Xinqi Bao
Division of Information Science and Engineering,
School of Electrical Engineering & Computer Science,
KTH Royal Institute of Technology,
xba@kth.se