Auxiliary Pretraining and Fine-Tuning Across Heterogeneous Datasets for ECG-Based Chagas Disease Detection

Bjørn-Jostein Singstad[1,2,3], Nikolai Olsen Eidheim[4], Amila Ruwan Guruge[5], Ola Marius Lysaker[5], Vimala Nunavath[4]

[1]Akershus University Hospital, Medical Technology and E-health, Lørenskog, Norway
[2]University of Oslo, Institute of Clinical Medicine, Oslo, Norway
[3]Vestfold Hospital, Department of Radiology, Tønsberg, Norway
[4]University of South-Eastern Norway, Department of Science and Industry Systems, Kongsberg, Norway
[5]University of South-Eastern Norway, Department of Process, Energy and Environmental Technology, Porsgrunn, Norway

Abstract

Chagas disease (American trypanosomiasis) is a neglected tropical disease caused by the parasite Trypanosoma cruzi. The disease can cause cardiac damage to humans known as chronic Chagas cardiomyopathy (CCC), manifesting as conduction blocks, arrhythmias, heart failure, and sudden death. The CODE-15% dataset contains more than 300.000 12-lead electrocardiogram (ECG) recordings, but the labeled data in this dataset are mostly weak, relying heavily on self-reported medical histories. We introduce auxiliary pretraining and subsequently perform fine-tuning on SaMi-Trop, which includes serologically verified Chagas patients, and PTB-XL, assumed to contain non-Chagas patients. The results show that the proposed model, when pretrained on the CODE-15% dataset and then fine-tuned with SaMi-Trop and PTB-XL, attained an AUROC of 0.69, an AUPRC of 0.22 on internal validation, and a challenge metric of 0.040 on hidden validation. Conversely, training only on CODE-15% and SaMi-Trop yielded an AUROC of 0.81, an AUPRC of 0.41 on internal validation, and a mean challenge score of 0.226 on the hidden test data, resulting in a 12th place in the challenge. These findings highlight a key limitation as the proposed pretraining strategy on auxiliary labels from CODE-15% and fine-tuning on PTB-XL and SaMi-trop offered no benefit and underperformed relative to conventional methods.

1. Introduction

Chagas disease (American trypanosomiasis) is a neglected tropical disease caused by the parasite Trypanosoma cruzi. The disease is widespread in Latin America and has spread globally via migration. Chronic Chagas disease can lead to cardiac damage known as chronic Chagas cardiomyopathy (CCC), manifesting as conduction blocks, arrhythmias, heart failure, and sudden death [1]. Unfortunately, most infected individuals remain undiagnosed and untreated. In many countries, less than 10% of Chagas cases are detected and even often as little as below 1% [2]. Early diagnosis is crucial, as antiparasitic treatment in the indeterminate phase can prevent progression to cardiomyopathy. However, screening for Chagas currently requires serological tests that are impractical for broad population screening due to cost and infrastructure needs. The 12-lead electrocardiogram is a commonly used and cost-effective diagnostic tool that can be beneficial in the screening process for Chagas disease. Characteristic ECG anomalies, such as a right bundle branch block (RBBB) combined with a left anterior fascicular block, atrioventricular conduction blocks (AVB), and various arrhythmias, including atrial fibrillation (AF), atrial flutter, and ventricular extrasystoles, are frequently observed in patients with CCC [3].

Prior work has shown that deep neural networks can detect hidden diseases or patient attributes from ECG signals, such as predicting a patient's age and sex [4] or silent conditions [5], with high accuracy. Recently, Ribeiro et al. developed a deep learning model to detect Chagas disease from ECG [6] and achieved an area under the ROC curve (AUC) of 0.80 on the internal validation set, demonstrating the promise of AI for ECG-based Chagas detection. However, performance dropped on external data (AUC 0.59–0.68) and the detection of Chagas in early, non-CCC cases, remains limited. The authors noted that improving data quality and incorporating additional patient information, like epi-

demiological risk factors, could enhance early detection. A major challenge is the limited availability of extensive, high-quality labeled data for Chagas disease. Although there are large ECG databases such as the CODE dataset, which contains millions of ECG records, the labels for Chagas disease in these datasets are mostly weak, relying heavily on self-reported medical histories. In contrast, smaller cohorts like SaMi-Trop provide strong labels confirmed by serology but cover only Chagas patients, lacking negative examples. In our participation as the Cha-Cha-Chagas team in the 2025 George Moody Challenge [7–9], we aim to tackle these data limitations. Our approach involves mitigating the limitations associated with weak labels by employing auxiliary pretraining that make use of more dependable labels.

2. Methods

2.1. Data

The datasets used are composed of multi-lead electrocardiogram (ECG) recordings derived from three primary sources: CODE-15% [10], SaMi-Trop [11] and PTB-XL [12, 13]. Each ECG record included raw waveform data alongside structured metadata, which contained variables such as patient age, sex, data origin, and binary outcome labels indicating the presence or absence of Chagas disease. It is important to note that in the CODE-15% dataset, these Chagas disease labels were based on self-reports. Conversely, in the SaMi-Trop dataset, the Chagas labels were determined through serological testing, making them more reliable compared to the self-reported labels in CODE-15%. However, it should be considered that the CODE-15% dataset might include ECG recordings from individuals at an earlier stage of the disease. This can potentially offer an opportunity to identify Chagas disease at an earlier phase. The PTB-XL dataset, comprising 21,799 12-lead ECG recordings, was collected in Germany throughout the 1990s. Given the time period and geographical location of these recordings, it is plausible to assume the absence of Chagas disease patients within this dataset and the recordings can serve as control samples for the conducted study.

2.2. Data Preprocessing

The signals underwent a resampling process to achieve a frequency of 100 Hz and were uniformly adjusted to a duration of 7 seconds. This 7-second duration was chosen because it represents the shortest recording present in the dataset, specifically from the SaMi-Trop dataset. Extending these recordings via zero-padding to a length of 10 seconds was avoided because such an action could unintentionally result in data leakage. As SaMi-Trop exclusively consists of positive cases, and the model might unintentionally learn to classify all zero-padded ECGs as Chagas disease due to the presence of these confounding zero-padded signals.

The selection of ECG leads was limited to limb leads (I,II and III) and precordial leads (V1-V6), because the augmented limb leads are mathematically derived from the frontal leads, thus offering no additional information beyond what is already provided by the frontal leads themselves.

2.3. Model Architecture

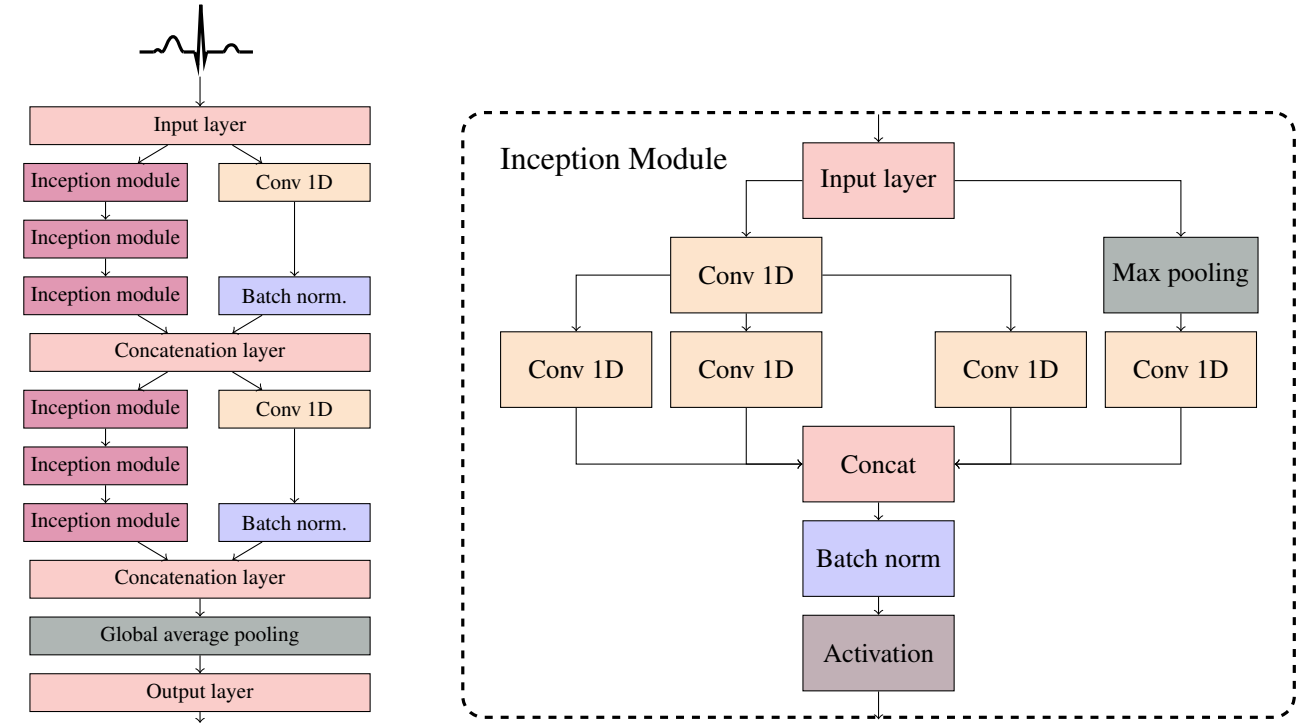
We implemented a 1D convolutional neural network with an Inception Time architecture, which combines residual connections, maintaining stable optimization in deeper networks, and parallel convolutional filters of varying receptive fields, to capture temporal features of various lengths [14]. Figure 1(a) illustrates the architectural design of the Inception Time network. Meanwhile, Figure 1(b) provides a detailed view of an individual inception module, thereby offering a closer examination of its components and structure.

2.4. Experimental Setup and Training

In order to determine the efficacy of using auxiliary pretraining, an evaluation was conducted comparing two approaches. In the first approach, we used auxiliary pretraining followed by fine-tuning on binary Chagas disease labels, and in the second approach, we trained the model from scratch solely on binary Chagas disease labels.

2.4.1. Auxiliary Pretraining and Fine-Tuning

During the pretraining phase, the model was consistently trained on the CODE-15% dataset, where various attributes, namely age, gender, first-degree AVB, RBBB, left bundle branch block, sinus bradycardia, sinus tachycardia and atrial fibrillation were employed as target labels. The training extended over 15 epochs, and binary cross-entropy loss was utilized for all classes, except for age, for which the mean squared error was applied. After completing the auxiliary pretraining, we froze all layers in the network, removed the final layer, and replaced it with an unfrozen single neuron with sigmoid activation. We then fine-tuned the model on the SaMi-Trop and PTB-XL datasets. This newly integrated output layer underwent further training for an additional 30 epochs. To tackle the issue of class imbalance ($N_{\text{Chagas}} \ll N_{\text{non-Chagas}}$), each mini-batch was carefully stratified to ensure an equal representation of Chagas and non-Chagas cases. The number of iterations conducted per epoch was calculated as $round(\frac{N}{B})$, with N representing the total count of ECG recordings, and B being the designated size for each mini-batch. Because of this batch balancing method, certain Chagas cases were sampled multiple times within a single epoch.



(a) The high-level architecture of the implemented Inception Time model.

(b) The architecture incorporated within each inception module.

Figure 1: Overview of the Inception Time architecture and a detailed examination of the inception module model.

2.4.2. Conventional Training

In the conventional training methodology, the model was exclusively trained on the Chagas label by utilizing the three available datasets. Specifically, we conducted experiments where various dataset was held out at a time to evaluate its impact. The training extended over 15 epochs, employing binary cross-entropy as the loss function, along with a balanced batch strategy, similar to the technique employed in the first training approach.

All experiments, both in first and second apporach, were performed using AdamW optimizer [15] with $\eta = 0.001$ and a mini-batch size of 32. 15% of the dataset was always withheld as validation data and to ensure the best-performing model was retained we saved the set of weights that achieved the highest validation area under the precision-recall curve (AUPRC) during the 15 epochs[1].

3. Results and Discussions

To evaluate model performance, we employed four distinct dataset configurations on the external hidden validation set. Among these, two configurations underwent internal validation, utilizing AUROC and AUPRC. The model that achieved the best result on the external validation was applied to three external test sets. Details of both the internal and external validations, along with the associated tests, are documented in Table 1. Notably, our top-performing model achieved challenge scores of 0.296, 0.247, and 0.136 across the three distinct test sets, resulting in an average challenge score of 0.226. This performance placed our team in 12th place among a total of 40 qualified teams.

The setup that relied solely on the use of the SaMi-Trop and CODE-15% datasets, with conventional training methods, demonstrated superior performance compared to the results obtained from our proposed approach, which incorporated auxiliary training techniques. Several factors could explain the underperformance by the proposed method. The first possible explanation is related to the pretraining with auxiliary labels may not have been relevant to the subsequent Chagas-label training, resulting in the frozen feature extractor failing to extract relevant features during finetuning. A second and more plausible explanation is related to the finetuning, where all negative labels are sourced from PTB-XL and all positives are from SaMi-Trop, two separate datasets collected at different times and locations (SaMi-Trop from Brazil during 2011-2012 and PTB-XL from Germany during 1989-1996). Differences in ECG

[1]The code referenced in this paper can be accessed here: `https://github.com/Bsingstad/GMC2025`

Data used for pretraining	Data used for Fine-tune or conventional training	AUROC on internal validation data	AUPRC on internal validation data	Challenge score on hidden validation data	Challenge score on hidden test set		
					REDS-II	SaMi-Trop	ELSA-Brasil
CODE-15%	SaMi-Trop and PTB-XL	0.69	0.22	0.040	-	-	-
-	SaMi-Trop and PTB-XL	-	-	0.066	-	-	-
-	CODE-15%	-	-	0.143	-	-	-
-	SaMi-Trop and CODE-15%	0.81	0.41	0.316	0.296	0.247	0.136

Table 1: Performance comparison of models trained with different pretraining and fine-tuning datasets. AUROC and AUPRC are reported on internal validation data, while the challenge metric reflects performance on hidden validation data.

acquisition and selection bias may introduce unintended artifacts, which could be learned from a model and make it act as a dataset classifier rather than a Chagas classifier.

4. Conclusions and Future Work

We investigated whether auxiliary pretraining on weakly labeled ECG data from CODE-15% could improve the detection of Chagas disease when combined with fine-tuning on serologically verified patients and presumed non-Chagas controls. Contrary to our expectations, this method underperformed relative to conventional training approaches, especially when training only using CODE-15% and SaMi-Trop. These results highlight the model's susceptibility to the dataset's composition, particularly when positive and negative labels originate from distinct populations, acquisition methods, or temporal contexts. Future research should consider these biases and develop strategies to mitigate the influence of acquisition artifacts when combining datasets from varied global regions and populations.

Acknowledgments

This work was supported by the Norwegian Health Association (#29038 - Artificial intelligence-enabled ECG).

References

[1] Roman-Campos D, Marin-Neto JA, Santos-Miranda A, Kong N, D'Avila A, Rassi A. Arrhythmogenic manifestations of chagas disease: Perspectives from the bench to bedside. Circulation Research May 2024;134(10):1379–1397.

[2] Editor. Putting chagas disease on the global health agenda. BMC Medicine May 2023;21. ISSN 1741-7015.

[3] 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 June 2018;12(6). ISSN 1935-2727.

[4] Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, et al. Age and sex estimation using artificial intelligence from standard 12-Lead ECGs. Circulation Arrhythmia and Electrophysiology September 2019;12(9). ISSN 1941-3084.

[5] Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet London England September 2019;394(10201):861–867. ISSN 1474-547X.

[6] Jidling C, Gedon D, Schön TB, Oliveira CDL, Cardoso CS, Ferreira AM, et al. Screening for chagas disease from the electrocardiogram using a deep neural network. PLOS Neglected Tropical Diseases July 2023;17(7).

[7] Reyna MA, Koscova Z, Pavlus J, Saghafi S, Weigle J, Elola A, et al. Detection of chagas disease from the ECG: the George B. Moody physionet challenge 2025, 2025. URL `https://arxiv.org/abs/2510.02202`. _eprint: 2510.02202.

[8] 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. In Computing in Cardiology 2025, volume 52. 2025; 1–4.

[9] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, physiotoolkit, and physionet. Circulation June 2000;101(23):e215–e220.

[10] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications April 2020;11(1). ISSN 2041-1723.

[11] Cardoso CS, Sabino EC, Oliveira CDL, Oliveira LCd, Ferreira AM, Cunha-Neto E, et al. Longitudinal study of patients with chronic chagas cardiomyopathy in Brazil (samitrop project): A cohort profile. BMJ Open May 2016;6(5). ISSN 2044-6055, 2044-6055.

[12] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data May 2020;7(1):154. ISSN 2052-4463.

[13] Wagner P, Strodthoff N, Bousseljot R, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3), 2022.

[14] Ismail Fawaz H, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, et al. InceptionTime: Finding AlexNet for time series classification. Data Mining and Knowledge Discovery November 2020;34(6):1936–1962. ISSN 1573-756X.

[15] Loshchilov I, Hutter F. Decoupled weight decay regularization, January 2019. ArXiv:1711.05101 [cs].

Address for correspondence:

Bjørn-Jostein Singstad
Halfdan Wilhelmsens alle 17, 3103 Tønsberg, Norway
bjosin@siv.no