Reliability-Aware Hierarchical Learning for Chagas Detection from Electrocardiogram under Expert Label Scarcity

Hao Wen[1], Jingsu Kang[2]

[1]College of Science, China Agricultural University, Beijing, China
[2]Tianjin Medical University, Tianjin, China

Abstract

Aim: We present a reliability-aware hierarchical learning framework for ECG-based Chagas cardiomyopathy screening in the George B. Moody PhysioNet Challenge 2025 by Team Revenger, aiming to maximize positive case retrieval under prevalence constraints.

Methods: The 12-lead ECGs were resampled to 400 Hz, bandpass filtered (0.5–45 Hz), and z-score normalized. We used a ResNet model integrated with squeeze-and-excitation (SE) modules for binary classification. To address severe class imbalance and the scarcity of expert-confirmed labels, we applied stratified upsampling and reliability-weighted label smoothing to prioritize expert-confirmed positives over self-reported ones. Model training used an asymmetric loss to further penalize false negatives and was optimized with AdamW and a OneCycle learning rate scheduler. Model selection was based on the Challenge score from an internal hold-out subset.

Results: On the hidden validation set, our method received a Challenge score of 0.245 (rank 187 / 373). In cross-validation on the public training data, our approach achieved a Challenge score of 0.451.

Conclusion: The proposed method shows effective performance for ECG-based Chagas screening, and highlights potential for improving detection accuracy and reliability in resource-limited scenarios.

1. Introduction

Addressing underdiagnosis of Chagas disease through scalable ECG-based screening is the focus of the 2025 George B. Moody PhysioNet Challenge [1, 2]. Enabled by aggregated multi-cohort ECG datasets [3–7], the Challenge frames a multi-source learning setting with heterogeneous label reliability and severe class imbalance.

In this work, we propose a reliability-aware hierarchical framework that prioritizes expert-confirmed labels and mitigates severe class imbalance within a deep ECG model, with optimization aligned to prevalence-constrained sensitivity objectives.

2. Methods

2.1. Datasets and Preprocessing

We used three ECG datasets for model training, with substantial differences in sample size, Chagas prevalence, and label provenance as summarized in Table 1.

Dataset	Size	Chagas rate	Label provenance
SaMi-Trop	1 631	100 %	expert-confirmed
CODE-15%	345 779	1.795 %	self-reported
PTB-XL	21 799	0 %	N/A

Table 1: Dataset statistics and label provenance. Chagas rate is the proportion of recordings labeled positive in each dataset. N/A indicates that confirmed Chagas cases are not expected (non-endemic population).

All ECGs were uniformly resampled to 400 Hz, bandpass filtered (0.5–45 Hz), and z-score normalized to zero mean and unit variance computed as in Eq. 1:

$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}}{\boldsymbol{\sigma}_{\mathbf{x}}} \tag{1}$$

where $\mathbf{x}$ is the original ECG signal, $\boldsymbol{\mu}_{\mathbf{x}}$ and $\boldsymbol{\sigma}_{\mathbf{x}}$ are the mean and standard deviation of $\mathbf{x}$, respectively. We excluded ECGs shorter than 1200 samples to ensure inputs contain enough cardiac cycles for stable model analysis.

2.2. Reliability-Aware Hierarchical Supervision

We introduce a hierarchical supervision scheme that encodes source reliability through stratified label smoothing and adaptive upsampling. Three reliability levels are defined: (1) expert-confirmed positives (SaMi-Trop, maximal trust), (2) self-reported samples (CODE-15%, both

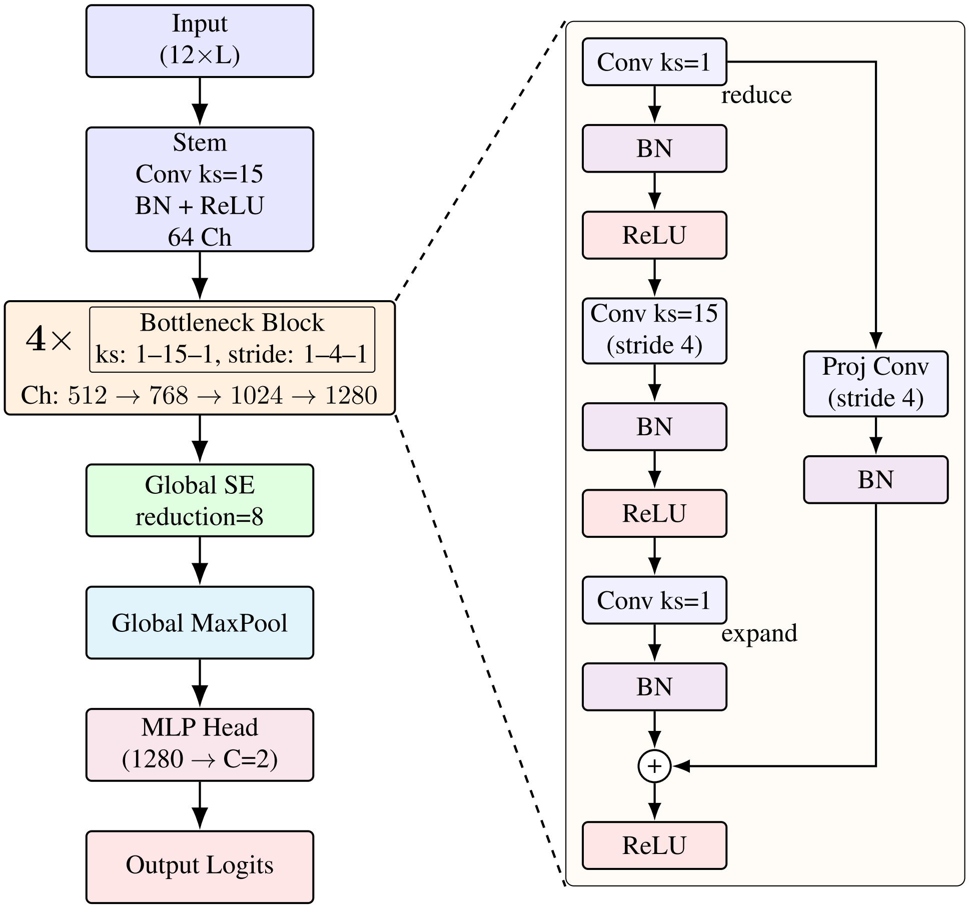

Figure 1: Model architecture. Left: overall network: a stem (Conv1d, kernel size (ks) 15, 64 channels (Ch), Batch Normalization (BN), ReLU) followed by four bottleneck residual blocks, a global squeeze-and-excitation (SE) module, global max pooling, and an MLP head producing $C = 2$ logits. Channel widths shown ($512 \to 768 \to 1024 \to 1280$) are the expanded channels. Right: internal bottleneck structure (1–15–1 pointwise–temporal–pointwise). The middle convolution of kernel size 15 uses a stride of 4 for temporal downsampling; kernel size 1 convolutions reduce and then expand channels, and a projection convolution (kernel size 1, stride 4) aligns resolution and width for the residual path. For clarity, dropout layers present in the implementation are omitted. Abbreviations: ks kernel size; Ch channels; BN Batch Normalization; SE squeeze-and-excitation; MLP multi-layer perceptron.

positives and negatives, higher uncertainty), and (3) non-endemic negatives (PTB-XL, very low true prevalence but still mildly regularized). Given a one-hot label $\mathbf{y}$ ($[0, 1]$ for positives, $[1, 0]$ for negatives) and number of classes $C = 2$, the smoothed target is computed as in Eq. 2:

$$\tilde{\mathbf{y}} = (1 - \varepsilon) \cdot \mathbf{y} + \frac{\varepsilon}{C} \cdot \mathbf{1}, \qquad (2)$$

where ε is the smoothing factor which depends on the reliability level: 0.0 (SaMi positives), 0.6 (CODE-15% positives & negatives), 0.2 (PTB-XL negatives). This attenuates overconfident gradients for noisier or potentially misreported labels while preserving sharp supervision on expert-confirmed cases.

Severe class imbalance was mitigated by upsampling positives during training: positive samples from CODE-15% were upsampled by a factor of 3, and those from SaMi-Trop by 12. No upsampling was applied to PTB-XL, which contains no positives. We chose these factors after reviewing hidden validation scores from multiple submissions, as shown in Table 2.

Smoothed labels and upsampling strategies for each dataset are summarized in Table 3.

2.3. Model Architecture

We build upon the 1D ResNet ECG classifier of Ribeiro et al. [3] and introduce three modifications.

(1) Bottleneck residual blocks. We replace basic ResNet blocks with bottleneck blocks of kernel sizes

Upsample factor		Challenge score
CODE-15%	SaMi-Trop	
-	-	0.239*
3	7	0.212
3	12	**0.245**
10	120	0.210
6	36	0.221

Table 2: Representative upsampling schemes and corresponding Challenge scores on the hidden validation set. The model and training strategies used were the same. "-" indicates no upsampling.
* obtained during the unofficial phase.

Dataset	Upsampling	Smoothed labels	
	factor	negative	positive
SaMi-Trop	12	N/A	[0, 1]
CODE-15%	3	[0.7, 0.3]	[0.3, 0.7]
PTB-XL	1	[0.9, 0.1]	N/A

Table 3: Smoothed labels (computed from Eq. 2) and upsampling strategies for each dataset.

1–15–1 (pointwise–temporal–pointwise). The middle temporal convolution applies a stride of 4 for downsampling; the two convolutions (kernel size 1) first reduce the number of channels and then expand them with an expansion factor of 4. A projection convolution (kernel size 1, stride 4) is used in the residual branch whenever temporal resolution or channel width changes. Across the four blocks, the reduced (bottleneck) channel widths are $128 \to 192 \to 256 \to 320$, yielding expanded output widths $512 \to 768 \to 1024 \to 1280$.

(2) Global squeeze-and-excitation (SE). After the final bottleneck block, a single global SE module (reduction ratio 8) [8] performs temporal average pooling to a channel descriptor, applies a two-layer bottleneck multi-layer perceptron (MLP) $1280 \to 160 \to 1280$ with ReLU and sigmoid gating, and rescales the feature map channel-wise.

(3) Global pooling head for variable input length. Instead of flattening a fixed-length feature map as in the original baseline, we apply global max pooling over the remaining temporal dimension, yielding a 1280-dimensional vector irrespective of input length L. This vector is fed to a lightweight two-layer classification MLP: a hidden fully connected layer ($1280 \to 1024$) with non-linear activation and dropout (rate 0.2), followed by a final linear layer ($1024 \to 2$) producing class logits.

Stem and regularization. A stem Conv1d (kernel size 15, stride 1, 64 channels) with BatchNorm and ReLU precedes the bottleneck stack. Within each bottleneck block, we apply BatchNorm+ReLU after the first two convolutions and dropout (rate 0.2) after each of those activations. All convolutions use "same" padding to preserve temporal length before downsampling operations.

The overall model architecture is illustrated in Fig. 1.

2.4. Training and Implementation Setups

We employed an asymmetric loss (ASL) [9] to complement the reliability-aware label smoothing strategy, jointly addressing the challenges of severe class imbalance. Let $\mathbf{z} = (z_0, z_1)$ denote the logits and $p = \operatorname{softmax}(\mathbf{z})_1$ the predicted probability of the Chagas-positive class. The ASL is defined in Eq. 3 with separate focusing parameters for positives and negatives and a clipped negative probability term:

$$\begin{aligned} L = {} & -y \cdot (1-p)^{\gamma_+} \log(p) \\ & - (1-y) \cdot (p_m)^{\gamma_-} \log(1-p_m), \end{aligned} \tag{3}$$

where y is the (smoothed) positive-class target probability, $p_m = \max(p - m, 0)$, $(\gamma_+, \gamma_-) = (1, 4)$ and margin $m = 0.05$. We train for 30 epochs with batch size 128 using the AdamW optimizer (initial learning rate 1×10^{-4}, peak 6×10^{-4} under a OneCycle scheduler, weight decay 1×10^{-2}). Early stopping (patience 10 epochs, monitored on a fixed 20% internal hold-out subset) selects the final model via the Challenge metric. Each training segment is a uniform random crop (or center trim if shorter) of length 4096 samples. The full implementation, including model construction, data pipeline, and optimization utilities, is based on the `torch-ECG` framework [10].

3. Results

The Challenge score of our team "Revenger" on the hidden validation set was 0.245, ranking 187th among 373 submissions. The score on the internal hold-out of the public training data, the hidden validation score, and the validation ranking are summarized in Table 4.

Training	Validation	Test	Ranking
0.451 ± 0.005	0.245	TBA	187 / 373

Table 4: Challenge scores for our submitted entries (team "Revenger"). Training: internal hold-out mean ± std over repeated runs. Validation: best among 10 validation submissions. Test: to be announced. Ranking: position on the hidden validation leaderboard.

4. Discussion and Conclusions

The hidden validation Challenge score presented in Table 4 indicates that our proposed method is effective for

Chagas screening from ECGs, albeit with substantial room for improvement. The result demonstrates our model's ability to learn diagnostically relevant features from ECGs for this task under scarce and noisy supervision. This is achieved through reliability-aware label smoothing, which incorporates both label provenance and reliability instead of treating all positive labels uniformly. Together with the asymmetric loss and strategic upsampling, these results indicate that explicitly modeling label reliability helps stabilize the learning process more effectively than introducing additional architectural complexity. Overall, our approach offers a scalable and resource-efficient solution and aligns well with the Challenge's objective of identifying high-risk individuals under limited serological testing capacity.

However, the performance gap between our internal hold-out subset and the hidden validation set suggests two primary limitations. First, the reliability weights (smoothing factors) were pre-defined based on label provenance rather than learned from data. This static assignment cannot capture the inherent heterogeneity of label quality within each source. Second, positive upsampling was applied uniformly at the dataset level using fixed factors. This strategy overlooks variations in individual sample difficulty and does not adapt to the model's evolving confidence during training. Furthermore, we did not make use of the additional labels available in some datasets, such as arrhythmia labels, to design and implement auxiliary learning tasks that could have enhanced the performance of the main task, Chagas screening.

Future research directions will primarily focus on the development of a self-adaptive supervision framework. This includes dynamic weighting schemes for learning instance-specific reliability scores, moving beyond static smoothing factors; and adaptive sampling strategies that respond to the model's evolving confidence during training, offering a promising alternative to fixed upsampling factors. Data augmentation techniques, such as CutMix [11] and SMOTE [12], could be applied to further expand and diversify the positive samples, thereby enhancing the model's robustness against overfitting and improving generalization to underrepresented patterns. Additionally, a multi-task learning framework leveraging auxiliary arrhythmia labels could enhance feature representation and improve generalization for the primary Chagas screening task. Exploring self-supervised pre-training on large-scale unlabeled ECG data also represents a promising direction to learn more transferable representations before fine-tuning on the target task.

References

[1] 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.

[2] 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.

[3] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic Diagnosis of the 12-lead ECG Using a Deep Neural Network. Nature Communications 4 2020;11(1):1–9.

[4] Cardoso CS, Sabino EC, Oliveira CDL, de Oliveira LC, Ferreira AM, Cunha-Neto E, et al. Longitudinal Study of Patients with Chronic Chagas Cardiomyopathy in Brazil (SaMi-Trop Project): A Cohort Profile. BMJ Open 5 2016; 6(5):e011181. ISSN 2044-6055.

[5] 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(1):1–15.

[6] Nunes MCP, Buss LF, Silva JLP, Martins LNA, Oliveira CDL, Cardoso CS, et al. Incidence and Predictors of Progression to Chagas Cardiomyopathy: Long-Term Follow-Up of Trypanosoma Cruzi –Seropositive Individuals. Circulation 11 2021;144(19):1553–1566. ISSN 1524-4539.

[7] Pinto-Filho MM, Brant LC, dos Reis RP, Giatti L, Duncan BB, Lotufo PA, et al. Prognostic Value of Electrocardiographic Abnormalities in Adults from the Brazilian Longitudinal Study of Adults' Health. Heart 12 2020; 107(19):1560–1566. ISSN 1468-201X.

[8] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 6 2018; 7132–7141.

[9] Ridnik T, Ben-Baruch E, Zamir N, Noy A, Friedman I, Protter M, et al. Asymmetric Loss for Multi-Label Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 10 2021; 82–91.

[10] Wen H, Kang J. A Novel Deep Learning Package for Electrocardiography Research. Physiological Measurement 11 2022;43(11):115006. ISSN 1361-6579.

[11] Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision. Institute of Electrical and Electronics Engineers (IEEE), 10 2019; 6022–6031.

[12] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 6 2002; 16(1):321–357. ISSN 1076-9757.

Address for correspondence:

Hao Wen
No. 17, Qinghua East Road, Haidian District, Beijing, China
wenh06@cau.edu.cn,wenh06@gmail.com