Ranking Aware Loss for CNN–Based Chagas Disease Detection from ECGs

Florian Herzler*[1], Nabil Jabareen*[2], Sören Lukassen[2]

[1] Free University Berlin,
[2] Berlin Institute of Health at Charité Berlin
These authors contributed equally.

Abstract

Our approach to the PhysioNet Challenge 2025 centers on a custom Ranking Aware Tversky (RAT) loss, explicitly designed to align optimization with the competition metric: true positive rate among the top 5% of ranked predictions (TPR@5%). While standard objectives such as Binary Cross-Entropy and Focal Loss optimize average accuracy, they fail to prioritize the high-confidence predictions critical for this task. RAT introduces a differentiable soft top-k weighting that emphasizes the most confident predictions, penalizes overconfident false positives through entropy regularization, and stabilizes training with a BCE anchor. To improve representation learning under severe class imbalance, we incorporate lightly weighted supervised contrastive learning, which further enhances intra-class cohesion and inter-class separation. Combined with a ResNet-18 backbone augmented by Group Normalization and Squeeze-and-Excitation modules, RAT consistently outperformed baseline losses in local validation, achieving the highest TPR@5%. These results highlight the importance of explicitly rank-aware loss design for ranking-based evaluation metrics in imbalanced clinical datasets. Our team CHA-CruzControl received an official challenge score of 0.05; however, this reflects a submission bug that caused all-negative predictions during inference rather than the true capability of our model.

1. Introduction

In clinical risk prediction tasks, accurate identification of a small set of high-confidence positives can be more valuable than optimizing overall accuracy. The PhysioNet [1] Challenge 2025 [2,3] embodies this setting by evaluating submissions with the true positive rate among the top 5% of predictions (TPR@5%). Recent efforts to address class imbalance in medical data have proposed asymmetric objectives, such as the Tversky [4] and Focal losses [5], yet these approaches still treat all predictions uniformly and fail to align with ranking-based metrics. Moreover, severe label imbalance exacerbates the risk of overconfident false positives, undermining performance in practice.

2. Material & Methods

The challenge training data combined multiple heterogeneous ECG datasets with substantial differences in population composition, recording protocols, and label quality. These variations introduce distributional shifts between subsets, while the overall dataset is highly imbalanced, with Chagas-positive cases representing only a small minority. Together, this heterogeneity and skewed class distribution necessitated careful preprocessing and training strategies robust to both imbalance and domain shift.

Demographic variables such as age and sex were also unevenly distributed across datasets and conditions. We did not include them as model inputs in order to avoid learning spurious demographic shortcuts, focusing instead on ECG-derived features.

2.1. Preprocessing

Raw ECG recordings from CODE-15, SaMi-Trop, and PTB-XL [6–8] varied in duration and exhibited common artifacts (baseline wander, high-frequency noise). To reduce inter-dataset variability and focus the model on disease-related morphology, we applied a uniform pipeline to all samples.

Iteration 1 (v1). We computed per-lead global mean and standard deviation across the entire dataset and applied per-lead z-normalization to every recording. As a qualitative sanity check of cross-dataset harmonization (not used for model selection), we embedded channel-level summary features with UMAP (10 statistics × 12 leads: mean, SD, min, max, peak-to-peak, interquartile range, skewness, kurtosis, RMS, zero crossings).

Iteration 2 (v2). To further improve inter-dataset harmonization we applied a zero-phase band-pass filtering (0.5–45 Hz) and a per-lead soft clipping to the 1st–99th empirical percentiles of each dataset. For all further experiments, v2 was used. We did not conduct formal AUROC or TPR@5% comparisons between v1 and v2; the transi-

ISSN: 2325-887X DOI: 10.22489/CinC.2025.243

tion was motivated by the need to standardize input characteristics across datasets based on qualitative assessments. A systematic quantitative comparison will be included in future work.

2.2. Model Architecture

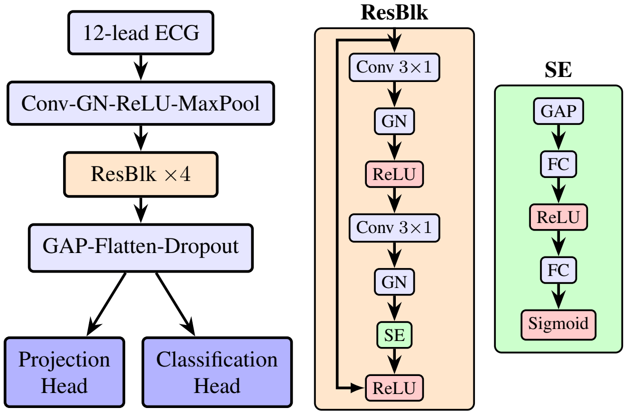

Figure 1. Network architecture supervised and contrastive learning for ECG classification. Left: model overview with classification head and projection head for contrastive loss. Middle: residual block with skip connection and Squeeze–Excitation (SE) block. Right: SE block specifications.

We adapted a ResNet-18 backbone [9] to one-dimensional convolutions for 12-lead ECG classification. ResNet-18 was chosen for its simplicity and computational efficiency, and because it has previously shown strong performance in ECG-based tasks [10]; deeper variants were not explored in this work. Batch normalization occasionally inhibited learning in our imbalanced setting, likely due to uneven statistics across batches. Therefore, all batch normalization layers were replaced with Group Normalization (GN) [11] (group size $=$ 8). Squeeze-and-Excitation (SE) modules [12] were inserted in each residual block, following the original design with a reduction ratio of 16. Finally, in addition to the standard classification head, we appended a two-layer MLP projection head for supervised contrastive training. Supervised contrastive training was motivated by the heavily imbalanced class distribution, aiming to learn generalizable ECG features. The final architecture is illustrated in Figure 1.

2.3. Training and Hyperparameters

We trained for 25 epochs with a learning rate of 1×10^{-4} using AdamW [13] and a One-Cycle scheduler (PyTorch Lightning [14]); longer runs did not yield additional gains, so we favored more repeats over more epochs. Data were split into 5 folds stratified by age, sex, and Chagas label.

To address imbalance, we enforced 10% positives per batch via a `WeightedRandomSampler`—with batch size 256 this yields 26 positives per batch, providing a stable minority gradient and enough positive pairs for SupCon while staying closer to the real (2%) prior than 50/50 oversampling—and, within each class, we oversampled strong labels (SaMi-Trop positives, PTB-XL negatives) to 25% to prioritize high-confidence supervision without letting a single source dominate.

Augmentation was intentionally minimal to align with our denoising preprocessing: segment masking ($\leq$100 samples), lead dropout (p=0.1), Gaussian noise (σ=0.01), and time warping ($\leq$15%).

2.4. Baselines

The evaluation metric for this challenge was the true positive rate (TPR) among the top 5% of ranked predictions (TPR@5%). This metric prioritizes highly confident correct predictions while strongly penalizing overconfident false positives, making it poorly aligned with standard loss functions such as Binary Cross-Entropy (BCE) or Focal Loss, which optimize overall accuracy rather than top-ranked precision. Additionally, the dataset exhibits severe class imbalance, further complicating optimization.

We first evaluated two baseline losses: **Binary Cross-Entropy (BCE)**, the conventional choice for binary classification and **Focal Loss** [5], designed to emphasize hard examples by down-weighting easy negatives ($\alpha = 0.9$, $\gamma = 1.5$). While these loss functions are common choices for binary classifications, they do not sufficiently align with TPR@5%, as all predictions contribute equally regardless of their ranking.

2.4.1. Supervised Contrastive Loss

To improve discriminative representation learning under severe class imbalance, we incorporate contrastive loss strategies [15], applied on the projection head. These losses encourage intra-class cohesion and inter-class separation, complementing the main classification objective. The contrastive losses were added to all tested loss functions as a weighted sum:

$$\mathcal{L} = \mathcal{L}_{\text{Supervised}} + \lambda_3 \mathcal{L}_{\text{Contrastive}} \quad (1)$$

As contrastive loss functions we used the Prototype loss (PT) and the supervised contrastive loss (SC) as defined in [15]. Prototypes were initialized randomly, and the projection head was used only during training.

We investigated different weighting strategies when combining them the contrastive loss with the supervised loss. We compared two approaches:

- Magnitude matching: down-weighting the contrastive loss such that its scale matches that of the supervised loss.

- Full contribution: assigning equal weight ($\lambda_3 = 1$) to both losses, even when their absolute scales differed by an order of magnitude.

This analysis allowed us to assess whether the contrastive losses improved performance only when carefully balanced against the classification objective, or whether stronger unweighted contributions were beneficial.

3. Ranking-Aware Tversky Loss

To address the TPR@5% metric, we developed a loss guided by three principles: prioritizing the most confident predictions through a soft top-k mechanism, mitigating overconfidence in negative predictions via entropy regularization, and promoting robust feature learning with auxiliary representation-based objectives, while maintaining training stability with a BCE anchor.

To align optimization with TPR@5%, we introduce a soft weighting scheme that emphasizes top-ranked predictions. Let $p_i = \sigma(x_i)$ denote the predicted probability for sample i, and $y_i \in \{0, 1\}$ its label. Define $\hat{p_i}$ as the top-k predictions, where k corresponds to 5% of the batch size. We compute a soft mask:

$$m_i = \sigma\left(\frac{p_i - \min(\hat{p_i})}{0.5}\right), \tag{2}$$

which assigns higher weights to samples near the top-k threshold while maintaining differentiability. The weighted counts of true positives (TP), false positives (FP), and false negatives (FN) are:

$$\mathrm{TP} = \sum_i m_i p_i y_i,$$
$$\mathrm{FP} = \sum_i m_i p_i (1 - y_i),$$
$$\mathrm{FN} = \sum_i m_i (1 - p_i) y_i.$$

These terms are integrated into the Tversky index loss [4] making it a rank aware:

$$\mathcal{L}_{\mathrm{RankTversky}} = 1 - \frac{\mathrm{TP} + \varepsilon}{\mathrm{TP} + \alpha\,\mathrm{FP} + \beta\,\mathrm{FN} + \varepsilon}, \tag{3}$$

We set $\alpha = 0.6$ and $\beta = 0.4$ to penalize false positives more strongly in line with the class imbalance. To encourage diverse predictions, we introduce an entropy term for a batch of size N, that punishes predicted probabilities close to 0.5:

$$\mathcal{L}_{\mathrm{Entropy}} = \frac{\sum_i \mathrm{p_i} \log \mathrm{p_i} + (1 - \mathrm{p_i}) \log(1 - \mathrm{p_i})}{N} \tag{4}$$

Finally, we add a BCE term to maintain gradient stability and avoid degenerate minima during early training.

The complete objective is a Ranking-Aware Tversky Loss (RAT):

$$\mathrm{RAT} = \mathcal{L}_{\mathrm{RankTversky}} + \lambda_1 \mathcal{L}_{\mathrm{Entropy}} + \lambda_2 \mathcal{L}_{\mathrm{BCE}}, \tag{5}$$

where $\lambda_1 = 0.005$ and $\lambda_2 = 0.1$ controls auxiliary loss weighting.

This formulation directly targets TPR@5% by concentrating on the top-ranked predictions, reducing overconfidence in negatives. Compared to conventional Tversky Loss, our approach selectively emphasizes high-confidence true positives while incorporating regularization and stability mechanisms.

4. Results

The results of all experiments and contrastive loss ablations are summarized in Table 1. All metrics reported are threshold independent and are computed on a local validation fold using bootstrapping.

Table 1. Classification performance across loss functions with supervised contrastive (SC) and prototype (PT) strategies under different weightings. Metrics: challenge score (TPR@5%), AUROC, and AP, reported as percentages. Shown is the mean ± SD over 10^3 bootstrap runs on the validation part of the training set. For each loss, the best configuration is underlined; the overall best is in **bold**.

Loss	Contrastive	TPR@5%	AUROC	AP
BCE		40.54±1.21	$\underline{81.38\pm0.59}$	$\underline{18.48\pm0.99}$
+	0.05×SC	40.39±1.20	80.74±0.59	17.91±0.94
+	1.00×SC	39.54±1.16	81.21±0.58	16.42±0.84
+	0.05×PT	$\underline{41.06\pm1.17}$	81.15±0.58	18.04±0.92
+	1.00×PT	38.50±1.15	80.77±0.56	16.33±0.89
Focal		40.21±1.20	80.90±0.59	17.78±0.98
+	0.01×SC	40.04±1.11	81.33±0.55	16.27±0.86
+	1.00×SC	38.39±1.10	80.92±0.56	15.59±0.81
+	0.01×PT	$\underline{41.19\pm1.22}$	$\underline{\mathbf{81.67}\pm0.59}$	$\underline{\mathbf{18.77}\pm1.01}$
+	1.00×PT	36.70±1.14	79.90±0.60	15.01±0.79
RAT (ours)		41.77±1.18	81.14±0.58	$\underline{17.64\pm0.93}$
+	0.05×SC	$\underline{\mathbf{42.03}\pm1.23}$	81.42±0.58	17.38±0.92
+	1.00×SC	41.09±1.19	81.29±0.59	16.76±0.84
+	0.05×PT	41.81±1.17	$\underline{81.58\pm0.58}$	17.17±0.91
+	1.00×PT	40.57±1.16	80.61±0.58	16.69±0.88

4.1. Contrastive Loss Consistently Improves Ranking Performance

Across all evaluated objectives, adding a weighted contrastive term ($\lambda_3 < 1$, see Section 2.4.1) improved performance on the challenge metric TPR@5%. In combination with BCE loss, contrastive learning did not yield measurable gains in AUROC or AP, suggesting its benefit is primarily in ranking the most confident positives rather than improving overall discrimination. Similarly, when combined with our proposed RAT loss, contrastive learning enhanced TPR@5% and AUROC, but had little effect on AP.

4.2. RAT Outperforms Standard Losses

Both BCE and Focal loss achieved comparable TPR@5% values in the range of 37–41%. In contrast, RAT consistently reached higher TPR@5% (41–42%) and showed the greatest robustness across different contrastive loss weightings. While RAT was the strongest performer on the challenge metric, it is noteworthy that Focal loss combined with prototype-based contrastive learning achieved the highest AUROC and AP overall. This highlights that optimizing for TPR@5% and optimizing for global discrimination metrics are not fully aligned, and that RAT is particularly effective for ranking-based evaluation.

5. Discussion

Prior Chagas–ECG work has largely trained CNNs with cross-entropy and reported global metrics (AUROC/AUPRC) rather than top-k recall; for example, Jidling et al. [10] trained a 12-lead CNN on CODE and SaMi-Trop and evaluated AUROC/AP, without rank-aware optimization. In contrast, our Ranking-Aware Tversky (RAT) loss directly targets the Challenge metric (TPR@5%) by prioritizing the highest-confidence slice of predictions under the 98:2 imbalance, yielding consistent gains over canonical objectives. This metric-aligned design is technically relevant—optimizing the part of the ranking that matters—and clinically practical for screening, where limited confirmatory testing capacity favors surfacing the most likely positives. Future work will validate on the hidden test set and systematically study weights for auxiliary losses.

References

[1] Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov P, Mark R, Mietus J, Moody G, Peng C, Stanley H. 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, Elola A, Hassannia MS, Campbell K, Bahrami Rad A, Ribeiro AH, Ribeiro AL, Sameni R, Clifford GD. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025. In Computing in Cardiology 2025, volume 52. 2025; 1–4.

[3] Reyna MA, Koscova Z, Pavlus J, Saghafi S, Weigle J, Elola A, Seyedi S, Campbell K, Li Q, Bahrami Rad A, Ribeiro A, Ribeiro ALP, Sameni R, Clifford GD. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025, 2025. URL https://arxiv.org/abs/2510.02202.

[4] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks, June 2017. URL http://arxiv.org/abs/1706.05721. ArXiv:1706.05721 [cs].

[5] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection, February 2018. URL http://arxiv.org/abs/1708.02002. ArXiv:1708.02002 [cs] version: 2.

[6] Ribeiro AH, Paixao GM, 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, June 2021. URL https://zenodo.org/records/4916206.

[7] Ribeiro ALP, Ribeiro AH, Paixao GM, 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, June 2021. URL https://zenodo.org/records/4905618.

[8] Wagner P, Strodthoff N, Bousseljot RD, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset.

[9] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. ArXiv:1512.03385 [cs].

[10] Jidling C, Gedon D, Schön TB, Oliveira CDL, Cardoso CS, Ferreira AM, Giatti L, Barreto SM, Sabino EC, Ribeiro ALP, Ribeiro AH. Screening for Chagas disease from the electrocardiogram using a deep neural network. PLOS Neglected Tropical Diseases July 2023;17(7):e0011118. ISSN 1935-2735. Publisher: Public Library of Science.

[11] Wu Y, He K. Group Normalization, June 2018. URL http://arxiv.org/abs/1803.08494. ArXiv:1803.08494 [cs].

[12] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks, May 2019. URL http://arxiv.org/abs/1709.01507. ArXiv:1709.01507 [cs].

[13] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization, January 2019. URL http://arxiv.org/abs/1711.05101. ArXiv:1711.05101 [cs].

[14] Falcon W, team TPL. PyTorch Lightning, April 2025. URL https://zenodo.org/records/15284694.

[15] Mildenberger D, Hager P, Rueckert D, Menten MJ. A Tale of Two Classes: Adapting Supervised Contrastive Learning to Binary Imbalanced Datasets, March 2025. URL http://arxiv.org/abs/2503.17024. ArXiv:2503.17024 [cs] version: 1.

Address for Correspondence:
Florian Herzler – florian.herzler@gmail.com
Berlin Institute of Health at Charité
Luisenstr. 65, 10117 Berlin