Sequential Deep Learning for Chagas Disease Screening: A CNN-BiLSTM Approach with an Attention Mechanism

Saber Jelodari Mamaghani, Adam Bokun, Heike Leutheuser

Ambient Assisted Living & Medical Assistance Systems, Department of Computer Science, University of Bayreuth, Bayreuth, DE

Abstract

Chagas disease detection from electrocardiograms (ECGs) is challenging due to subtle waveform changes hidden within complex signal patterns, variability in ECG morphology, and inherent noise, which demand robust automated diagnostic methods. For the 2025 George B. Moody PhysioNet Challenge, our team Chagas_UBT built an open-source ECG-based system to screen for Chagas disease from 12-lead ECGs. Our approach combines a convolutional neural network (CNN), bidirectional long short-term memory (BiLSTM) layers, and an attention layer. To address class imbalance, we use focal loss, ECG-specific data augmentation, mixup, and weighted sampling.

On the official hidden test set, our model achieved a Challenge score of 0.238, resulting in a rank of 10. This work contributes toward the earlier detection of Chagas disease, and our open-source implementation's low computational cost makes it practical for screening in resource-limited settings. The source code is available in our GitHub repository.

1. Introduction

Chagas disease is a chronic parasitic infection caused by *Trypanosoma cruzi* that is frequently underdiagnosed. Many infected individuals are asymptomatic for years before developing Chagas cardiomyopathy. Confirmatory serological tests capabilities are limited and can be overlooked when symptoms are absent or nonspecific. The George B. Moody PhysioNet Challenge 2025 focuses on detecting potential cases of Chagas disease from standard 12-lead electrocardiograms (ECGs) and helping prioritize patients for confirmatory serological testing and being treated in the early stages with a specific therapy [1, 2].

Our goal was to develop a computationally efficient and deployable algorithm for use with standard 10-second, 12-lead ECGs. We adopt a compact architecture that uses a convolutional neural network (CNN) for local morphology extraction, a bidirectional long short-term memory (BiLSTM) network for long-range temporal context, and an additive attention layer for time-wise aggregation. To handle strong class imbalance, we combine focal loss [3], ECG-specific data augmentation [4], mixup regularization [5], and weighted sampling, and we output calibrated risk probabilities to prioritize confirmatory testing.

2. Methods

2.1. Data Processing and Partitioning

In this research we used all available Challenge training datasets: CODE-15%, SaMi-Trop, and PTB-XL [6–8]. All sources provide standard 12-lead ECGs with labels as released by the respective datasets. The largest of these, CODE-15%, contains unvalidated labels derived from self-reported records. SaMi-Trop includes clinician-validated Chagas positives, and PTB-XL provides validated negatives. To approximate the hidden-test workflow, we reserved a small internal holdout by randomly sampling without replacement 100 Chagas positives records from SaMi-Trop and 100 negatives from PTB-XL. These 200 records were strictly excluded from training and model selection and were used only for internal testing.

To ensure consistency across all datasets before training the model, we standardized the recordings using Sami-Trop [7] as the reference, as it was the only source containing validated positive cases and had a sampling rate of 400 Hz. All other recordings, regardless of their original sampling rates (e.g., PTB-XL at 500 Hz [8]), were resampled to 400 Hz. Any signals with rates above or below this target were upsampled or downsampled accordingly. Then each recording was formatted as a $12 \times T$ matrix, where $T = 4000$ samples, corresponding to a 10-second window. Recordings shorter than 10 seconds were zero-padded, while those longer than 10 seconds were center-cropped. To ensure uniformity in channel structure, all signals were adjusted to have exactly 12 leads: recordings with fewer leads were zero-padded in the missing channels, and those with more than 12 leads had the extra chan-

nels removed. This preprocessing ensured all inputs fed into the model had the same shape and structure, facilitating consistent learning across varied data sources.

Each lead was independently normalized by per-lead z-scoring (mean subtraction and division by standard deviation with $\epsilon = 10^{-6}$) to have a zero mean and unit variance to mitigate non-diagnostic discrepancies in amplitude and baseline shift. This per-lead standardization ensures the model learns clinically relevant morphologies rather than being biased by inter-lead variations that arise from the different electrical viewpoints of the heart.

No relabeling was performed. Records that failed to load or produced non-finite values after preprocessing were excluded from training. At inference time, if loading failed (e.g., missing file, corrupted header, truncated signal), the system returned a deterministic fallback (negative label with low probability) to avoid pipeline interruption. The metadata was not included in the final model. Ablations indicated only minor benefits, which we attribute to incomplete metadata.

2.2. Data Augmentation

To mitigate overfitting and simulate real-world acquisition variability, we applied stochastic and physiologically plausible augmentations during training [9].

Specifically, we used a random time shift of ± 0.5 s to mimic onset misalignment; amplitude scaling with a multiplicative factor sampled uniformly from 0.9 to 1.1; lead dropout by randomly removing one channel per sample to simulate missing leads; additive zero-mean Gaussian noise with $\sigma = 0.01$; and Mixup regularization, linearly interpolating pairs of samples using a $\beta(0.4, 0.4)$ distribution at the batch level [5].

We used data augmentation to improve model generalization. Our methods simulated realistic signal variations (amplitude scaling, temporal shifts, lead failure) and included Mixup to improve regularization and calibration.

2.3. Model Architecture

We designed a compact CNN–BiLSTM–attention network to capture local ECG morphology and long-range temporal structure while keeping computational cost low. The network (Figure 1) takes each ECG as a fixed 12×4000 tensor and passes it through a three-layer 1D convolutional stem (kernel$=5$, padding$=2$, channels $12 \to 32 \to 64 \to 128$). Each convolution is followed by batch normalization, ReLU, maxpooling (stride$=2$), and dropout ($p = 0.4$). After the three conv–pool blocks, the temporal length is $4000/2^3 = 500$; we then apply adaptive average pooling to obtain 100 steps, which gives a fixed sequence length for the recurrent encoder and reduces computation. The resulting tensor has shape $[100, 128]$ per

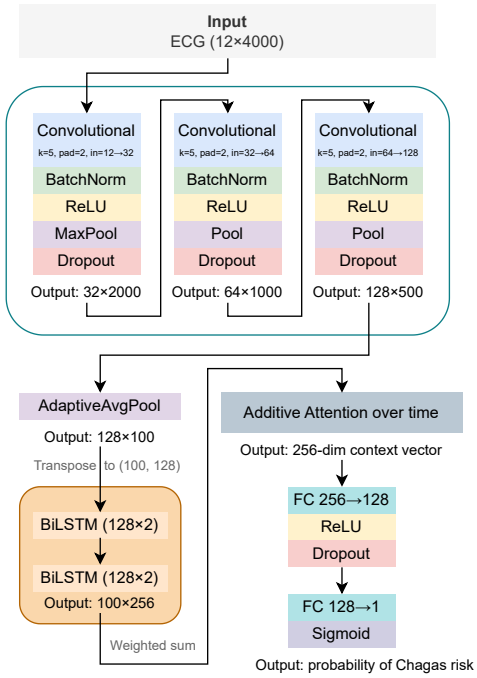

Figure 1. End-to-end model architecture for ECG-based Chagas screening

record (time $\times$ features).

A two-layer bidirectional LSTM (hidden size $= 128$ per direction; dropout$= 0.4$ between layers) encodes this sequence to capture long-range morphology and rhythm. We use two stacked LSTM layers to increase the depth of temporal representations; deeper LSTMs are a standard and effective choice for sequence classification and speech/time-series tasks [10], while bidirectional provides both past and future context within the 10 s window [11]. Concretely, the BiLSTM outputs $[100, 256]$ (forward+backward).

An additive attention module aggregates the 100 time steps into a single context vector [12]. Following, we project each step from 256 to 128, apply $\tanh$, score with a scalar linear layer, normalize scores with a softmax over time, and take the weighted sum, yielding a 256-dimensional context that emphasizes diagnostically salient intervals (e.g., QRS complexes, P–R segments).

A fully connected head (256$\to$128$\to$1) with ReLU and dropout produces the final logit, and a sigmoid maps it to the predicted probability of Chagas disease.

2.4. Addressing Class Imbalance

To address the severe class imbalance inherent in the Chagas screening dataset, we employed the focal loss function, originally proposed by Lin et al. [3]. This loss function reshapes the standard binary cross-entropy (BCE)

loss by adding a modulating factor that reduces the loss for well-classified examples, thereby focusing the model's training on harder, misclassified samples.

The formal, α-balanced variant of the focal loss is defined in the original paper as:

$$\mathcal{L}_{\text{focal}} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

where p_t is the probability of the ground-truth class. This is commonly implemented in practice using the BCE function for a given logit z:

$$\mathcal{L}_{\text{focal}} = \alpha(1-\hat{p})^{\gamma}\text{BCE}(\hat{\text{p}},\text{y}) \tag{2}$$

where $\hat{p} = \sigma(z)$ is the estimated probability for the positive class. Following the empirical results of the original study, we use the authors' recommended hyperparameters: a focusing parameter of $\gamma = 2.0$ and a class-balancing parameter of $\alpha = 0.25$ [3]. We use the Adam optimizer [13] (learning rate = 1e-3, weight decay = 1e-5), batch size = 32, and train for up to 40 epochs. A `ReduceLROnPlateau` scheduler decreases the learning rate by $\times 0.1$ if validation loss does not improve for 3 epochs. Early stopping halts training after 7 epochs without improvement. Dropout ($p = 0.4$) is applied in the convolutional stem and the penultimate fully connected layer.

2.5. Sampling Strategy and Training/Validation Split

To mitigate class imbalance and varied label quality across datasets, we designed a two-part sampling strategy. The primary mechanism for countering class imbalance was a `WeightedRandomSampler`. This strategy modifies the data distribution seen by the model by setting the sampling weight (w_c) for each class (c) to be inversely proportional to its frequency, thus over-sampling the minority class [14]:

$$w_c \propto \frac{1}{\text{count}(c)} \tag{3}$$

Additionally, to prioritize learning from the most reliable data, we applied a secondary weighting based on the data source. Samples from the SaMi-Trop and PTB-XL datasets, which contain validated or higher-quality labels, received an additional weight multiplier of 2.0. This combined approach ensures that each training batch is not only balanced with respect to the Chagas label but also enriched with high-quality examples.

For this study, we adopted a single stratified train/validation split. The validation set was reserved exclusively for hyperparameter tuning and early stopping, and was never used for model retraining. During training, we recorded a comprehensive set of performance metrics per epoch, including loss, F1 score, AUROC, sensitivity, precision, and accuracy. In addition, the decision threshold was optimized at each epoch by sweeping thresholds from 0.1 to 0.9 in increments of 0.1 and selecting the value that maximized the F1 score. The threshold corresponding to the best validation epoch was stored in the model checkpoint and consistently reused for binary predictions at inference.

3. Results

For the Challenge metric, only probabilities are used to compute TPR@5%, the expected proportion of positives among the top 5% ranked by predicted probability [1, 2]. In the official validation set and the hidden test set provided by the Challenge organizers, our algorithm achieved a TPR@5% of **0.343** and **0.238**, respectively, ranking **10th** on the public leaderboard (Table 1).

Training	Validation	Test	Ranking
0.2941	0.343	0.238	10/41

Table 1. The table shows the challenge scores for our selected entry (Chagas_UBT). Scores are reported on a local holdout dataset (200 ECGs: 100 positives / 100 negatives) excluded from training and model selection, as well as the official validation set and hidden test set evaluated by the challenge organizers.

4. Discussion and Conclusions

We present an open-source CNN–BiLSTM–attention model that processes 10 s, 12-lead ECGs and achieves competitive screening performance for Chagas disease (Table 1). A strength of the pipeline is its integration of multiple regularization strategies (mixup augmentation, lead dropout, focal loss, and weighted sampling) for severe class imbalance and an attention mechanism that concentrates on physiologically meaningful intervals such as the P–R and QRS complexes, which aligns with known manifestations of Chagas cardiomyopathy.

Compared with REDS-II validation, performance on the REDS-II hidden test is broadly consistent (Table 2). We did class-balancing and then up-weighted the "strong" sources (SaMi-Trop, PTB-XL), which explains the good REDS-II generalization and the relatively small drop on SaMi-Trop. Several limitations remain, where ELSA-Brasil is the hardest cohort, pointing to remaining robustness gaps.

Future work will focus on two areas: **(i)** rank-aware optimization to directly target the TPR@5% metric using methods like LambdaRank [17] or differentiable sorting [18]; and **(ii)** uncertainty calibration, where techniques such as temperature scaling [19] and deep ensembles [20]

Metrics	REDS-II validation	REDS-II test	SaMi-Trop 3	ELSA-Brasil
TPR@5%	0.343	0.314	0.295	0.104
AUROC	0.734	0.730	0.722	0.555
AUPRC	0.150	0.099	0.105	0.036
Accuracy	0.980	0.967	0.974	0.979
F-measure	0.082	0.104	0.104	0.020

Table 2. Performance across different datasets. Metrics computed on the hidden validation set REDS-II [15] and the three hidden test sets: REDS-II, SaMi-Trop 3, and ELSA-Brasil[16]. TPR@5% is the true positive rate among the top 5% highest-risk predictions, which is the Challenge's primary ranking metric.

can improve model trustworthiness, especially amid shifts in data prevalence.

In conclusion, our study demonstrates that a moderately complex CNN–BiLSTM–Attention model, strengthened by targeted augmentation and regularization, can deliver competitive performance for Chagas disease screening while remaining computationally efficient. The implementation is practical for large-scale screening programs and adaptable to other rare ECG phenotypes. With rank-based learning and more rigorous calibration, we expect further advances where effective and reliable risk prioritization is most needed.

Acknowledgments

We acknowledge high performance computing resources of NHR@FAU (FAU), funded by federal and Bavarian state authorities; NHR@FAU hardware is partly funded by DFG (grant 440719683). Computations ran on *Alex* (NVIDIA A100).

References

[1] 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.

[2] 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.

[3] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (ICCV). 2017; 2999–3007.

[4] Lai J, Tan H, Wang J, Ji L, Guo J, Han B, et al. Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset. Nature Communications 2023;14(1):3741.

[5] Zhang H, Cisse M, Dauphin Y, Lopez-Paz D. mixup: Beyond empirical risk minimization, 2017.

[6] Ribeiro A, Ribeiro M, Paixão G, Oliveira D, Gomes P, Canazart J, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. Nature Communications 2020; 11(1):1760.

[7] Cardoso C, Sabino E, Oliveira C, de Oliveira L, Ferreira A, Cunha-Neto E, et al. Longitudinal study of patients with chronic chagas cardiomyopathy in brazil (SaMi-Trop project): a cohort profile. BMJ Open 2016;6(5):e0011181.

[8] 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;7:154.

[9] Raghu A, Shanmugam D, Pomerantsev E, Guttag J, Stultz C. Data augmentation for electrocardiograms, 2022.

[10] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization, 2015. URL https://arxiv.org/abs/1409.2329.

[11] Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 1997; 45(6):2673–2681.

[12] Jiang M, Gu J, Li Y, Wei B, Zhang J, Wang Z, et al. Hadln: Hybrid attention-based deep learning network for automated arrhythmia classification. Frontiers in Physiology 2021;12:683025.

[13] Kingma DP, Ba J. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[14] He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 2009; 21(9):1263–1284.

[15] Nunes M, Buss L, Silva J, Martins L, Oliveira C, Cardoso CS BB, et al. Incidence and predictors of progression to chagas cardiomyopathy: Long-term follow-up of trypanosoma cruzi-seropositive individuals. Circulation 2021; 144(19):1553–1566.

[16] Pinto-Filho M, Brant L, Dos Reis R, Giatti L, Duncan B, Lotufo P, et al. Prognostic value of electrocardiographic abnormalities in adults from the brazilian longitudinal study of adults' health. Heart 2021;107(19):1560–1566.

[17] Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, et al. Learning to rank using gradient descent. In ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning. 2005; 89–96.

[18] Blondel M, Teboul O, Berthet Q, Djolonga J. Fast differentiable sorting and ranking, 2020.

[19] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks, 2017.

[20] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

Address for correspondence:

Saber Jelodari Mamaghani
Nürnberger Str. 38, 95445 Bayreuth, Bayern, Germany
saber.jelodari@uni-bayreuth.de