ECG-Based Screening of Chagas Disease Using Deep Residual Networks and Feature-Based Machine Learning

Marion Taconné*, Stefano Magni, Cristian Drudi, Sara Maria Pagotto, Valentina D. A. Corino, Pietro Cerveri, Anna Maria Bianchi, Riccardo Barbieri and Luca Mainardi

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

Abstract

Chagas disease affects millions worldwide and can progress to chronic cardiomyopathy. Early detection is essential, but access to serological tests remains limited in endemic regions. Since Chagas-related abnormalities can be detectable on electrocardiograms (ECGs), we developed two automated approaches for disease screening in the framework of the George B. Moody PhysioNet Challenge 2025. The first is a deep learning method based on a residual neural network (ResNet) applied directly to ECG waveforms. The second is a machine learning pipeline that extracts fiducial features (intervals, slopes, and amplitude) and classifies them using Gradient Boosting. On internal validation, the ResNet achieved a score of 0.723 and the machine learning pipeline 0.486. In the unofficial phase, the ResNet reached 0.566 and ranked 20th. In the official phase, performance dropped to 0.061 for the ResNet, while the ML pipeline achieved 0.088. These results illustrate the challenges of robust ECG-based Chagas screening, as the official phase evaluates true generalization from diverse clinical sources.

1. Introduction

Chagas disease, caused by Trypanosoma cruzi, affects an estimated 6–7 million individuals globally, with the majority residing in South America but cases increasingly reported worldwide due to migration. Chronic Chagas cardiomyopathy represents the most severe form of the disease and is a major cause of morbidity and mortality in endemic areas. Between 20% and 40% of chronically infected individuals eventually develop cardiac involvement, which may manifest as heart failure, arrhythmias, or thromboembolic events [1]. Early diagnosis is essential to prevent progression, yet access to confirmatory serological testing remains limited in resource-constrained environments.

Electrocardiography offers a simple, inexpensive, and widely available diagnostic tool. Chagas patients often present conduction abnormalities, ventricular arrhythmias, or nonspecific ST–T changes, which may serve as early markers of disease [2, 3]. Automated ECG interpretation supported by artificial intelligence (AI) has shown increasing promise in cardiovascular medicine [4, 5].

Two main AI paradigms are commonly explored for ECG analysis. Deep learning (DL) methods learn directly from raw waveforms, capturing complex morphological and temporal patterns without requiring explicit feature engineering [6]. In contrast, feature-based machine learning (ML) approaches rely on manually engineered descriptors derived from the ECG signal, which are then used by conventional classifiers. While DL models can automatically learn complex representations and often achieve superior accuracy when trained on large datasets, ML approaches retain the advantage of interpretability and can offer robustness when based on physiologically meaningful features.

The George B. Moody PhysioNet Challenge 2025 provides a large, curated dataset of annotated ECGs for the development of algorithms targeting Chagas disease detection. In this work, we present and compare two complementary approaches: a residual neural network (ResNet) tailored for end-to-end ECG classification, and a feature-based ML pipeline that extracts fiducial ECG features and classifies them using Gradient Boosting. We report results from internal validation as well as from the unofficial and official phases of the Challenge, highlighting the strengths and limitations of each method.

2. Methods

2.1. Study Population

The dataset was provided by the George B. Moody PhysioNet Challenge 2025 and included standard 12-lead ECGs from multiple public and private sources. The largest portion came from the CODE-15% dataset (Brazil, >300,000 ECGs, 400 Hz, 7–10 s) with weak labels based on self-reported Chagas status. The SaMi-Trop dataset (Brazil, 1,631 ECGs, 400 Hz, 7–10 s) consisted entirely

of serologically confirmed positive cases, providing strong labels. The PTB-XL dataset (Germany, 21,799 ECGs, 500 Hz, 10 s) contained presumed negative cases, adding diversity in morphology and acquisition protocols.

In addition, several smaller private datasets from endemic regions with strong serological labels were reserved for hidden validation and testing. All data were distributed in WFDB format with metadata on demographics, acquisition parameters, and binary Chagas labels. This heterogeneous mix of weakly and strongly labeled data was intended to reflect real-world screening conditions.

2.2. Deep Learning Approach

2.2.1. Preprocessing

All ECGs were resampled to a uniform frequency of 500 Hz. Baseline wander was removed using a high-pass filter with a 0.5 Hz cutoff. To ensure consistent input dimensionality, signals were normalized to a fixed length of 10 s: shorter recordings were zero-padded while longer recordings were cropped centrally. Each lead was z-score standardized.

A signal quality index (SQI) was implemented to detect corrupted or unreliable ECGs [7]. The SQI combined measures of frequency content, amplitude distribution, and recording duration. ECGs failing the SQI threshold were automatically assigned to the non-Chagas class during inference.

2.2.2. Model Architecture

We employed a deep residual convolutional neural network (ResNet), consisting of stacked residual blocks with batch normalization and ReLU activations. Each block incorporated skip connections to enable gradient flow in deeper networks. The network was designed to capture both local waveform morphology and long-range temporal dependencies across leads. Global average pooling was applied before the final dense layer with sigmoid activation for binary classification (Figure 1).

2.2.3. Training Strategy

Training was performed using the Adam optimizer with an initial learning rate of 1e-4 and binary cross-entropy loss. We applied early stopping based on validation performance. Class imbalance was addressed through weighted loss functions. Data augmentation included random cropping and amplitude scaling to improve generalization.

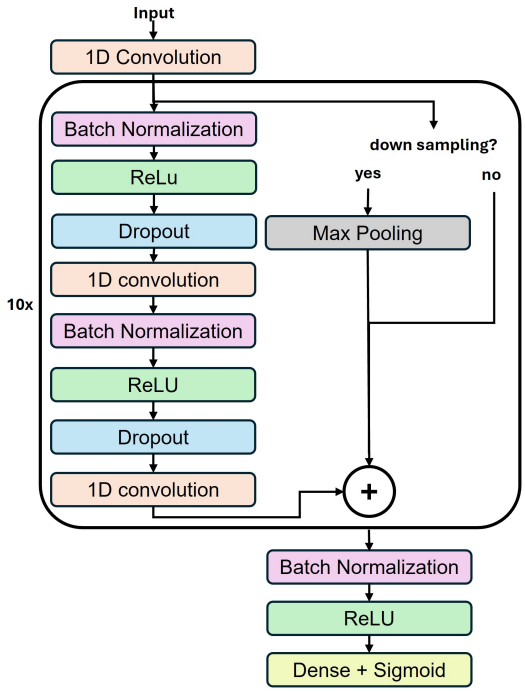

Figure 1: ResNet model architecture.

2.3. Feature-Based Machine Learning Approach

2.3.1. Preprocessing

In parallel, we implemented a feature-based ML pipeline. ECGs were band-pass filtered (1–40 Hz), resampled to 400 Hz.

2.3.2. Feature extraction

Fiducial points were detected using the neurokit2 library [8]. From these fiducial points, we extracted a set of interpretable features per lead (Figure 2), similarly to [9]:

- Amplitudes: median amplitude of P, Q, R, S, J and T points.
- Ratios: R/P and R/T amplitude ratios.
- Slopes: ascending and descending slopes of P and T waves, as well as QR, RS, and SJ segments.
- Intervals: median duration of PR, PS, PT, QT, QRS, RS, ToTp and TpTe intervals.
- Negative percentage of QRS area.

For each feature, the median value across all beats of a recording was computed, resulting in 24 features per lead. Since all 12 leads were processed, the final patient-level feature vector consisted of 288 features (24×12).

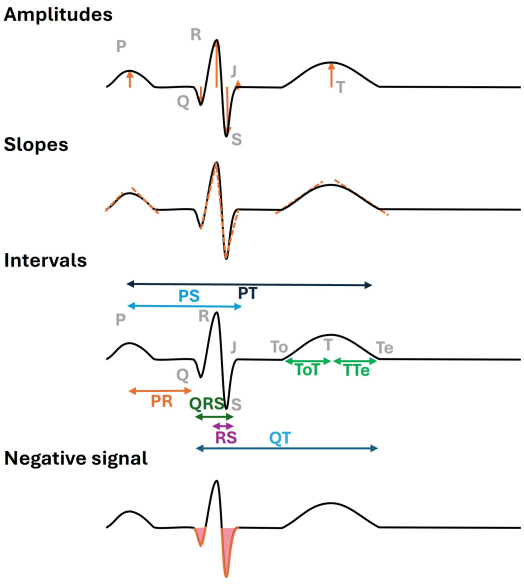

Figure 2: Morphological features extraction: amplitude of the fiducial points, slopes of the waves, interval width, and negative % of the signal (To: T offset, Te: T end).

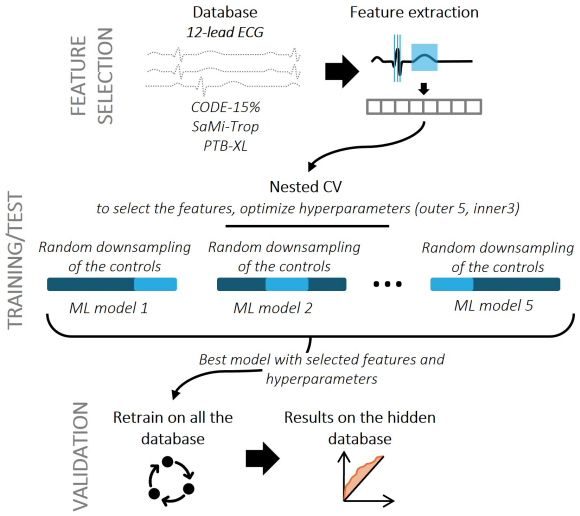

Figure 3: Methodological steps of the machine learning pipeline.

2.3.3. ML pipeline

In the ML pipeline illustrated in Figure 3, we first applied median imputation to handle missing values, followed by z-score standardization. Class imbalance was addressed by random undersampling of the majority class. Feature selection was then performed using the same model as estimator, with the selection threshold optimized through grid search.

Different classifiers were tested within this pipeline, including Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machines (SVM). Their hyperparameters were tuned inside nested cross-validation (5 outer and 3 inter loops), and the model achieving the best performance across folds was selected for submission to the official phase. This final model was then retrained on the full training dataset before submission.

3. Results

The different performances of the two approaches during the different phases of the challenge are summarized in Table 1.

3.1. Internal Validation

On internal hold-out validation derived from the public training data, the ResNet achieved a Challenge score of 0.723. Among the machine learning classifiers tested, Gradient Boosting provided the most stable performance with a score of 0.486, while Random Forest, Logistic Regression, and SVM performed slightly lower.

3.2. Unofficial Phase

In the unofficial phase, only the ResNet was submitted. It achieved a score of 0.566, ranking 20th on the public leaderboard. This decrease compared to internal validation reflected a first indication of dataset heterogeneity between the public training set and the hidden validation cohorts.

3.3. Official Phase

In the official phase, both methods were evaluated on the hidden test set consisting of additional strongly labeled clinical data. The ResNet's performance dropped to 0.061, while the Gradient Boosting pipeline achieved 0.088. Although both scores were low, the slightly higher performance of the feature-based method suggests that interpretable biomarkers may provide some resilience against dataset variability, even if overall generalization remained limited.

Table 1: Performance of deep learning (ResNet) and feature-based ML (Gradient Boosting) across internal, unofficial, and official phases of the Challenge.

Method	Internal	Unofficial	Official
DL (ResNet)	0.723	0.566	0.061
ML (Gradient Boosting)	0.486	–	0.088

4. Discussion

Both approaches demonstrated encouraging performance on internal validation data, but results deteriorated in the official phase. The ResNet, competitive during the unofficial phase (0.566), fell to 0.061 on the official test set. The ML pipeline, though not tested in the unofficial phase, achieved a slightly higher score of 0.088.

This gap underscores the fundamental difference between challenge phases. The unofficial leaderboard reflects performance on a hidden validation set similar in distribution to the public training data, whereas the official phase evaluates true generalization: the ability to handle unseen, heterogeneous datasets with strong serological labels collected in different clinical contexts. Both our models, like many others, struggled to adapt to this distribution shift.

Several factors likely contributed to the performance drop: reliance on weak labels in the large CODE dataset, which may have introduced noise during training; dataset heterogeneity in sampling rates, recording durations, and comorbidities; and the intrinsic difficulty of the task itself. In fact, some patients with serologically confirmed Chagas disease do not present detectable abnormalities on the ECG [1], making classification particularly challenging even for robust models.

The feature-based approach suggests a degree of robustness linked to clinically interpretable biomarkers such as intervals and slopes, whereas the ResNet can automatically learn complex waveform morphology. Both paradigms bring complementary advantages, but each is also limited by the heterogeneity of real-world data and the fact that not all positive cases manifest at the cardiac level.

Future work should focus on bridging these strengths by combining deep and feature-based models, integrating domain adaptation strategies, and leveraging semi-supervised approaches to mitigate the impact of weak labels. Representation learning may also enable relabeling and improved dataset curation, which are crucial for building models that truly generalize across sources.

5. Conclusion

We proposed two complementary approaches for ECG-based screening of Chagas disease: a deep residual network and a feature-based Gradient Boosting pipeline. Both performed well on internal validation, but their scores dropped in the official phase (0.061 and 0.088 respectively). This reflects the fact that the official phase tests true generalization, requiring robustness to unseen, heterogeneous, and strongly labeled datasets. Beyond methodological considerations, it is important to note that not all patients with serologically confirmed Chagas disease present with ECG abnormalities, which inherently limits the performance ceiling of ECG-based approaches. These findings therefore underline both the clinical and technical challenges of this task, and highlight the need for strategies resilient to domain shift and label noise, as well as for integration with complementary diagnostic modalities.

Acknowledgment

This work was developed in the context of the George B. Moody PhysioNet Challenge 2025.

References

[1] Ribeiro AL, Nunes MP, Teixeira MM, Rocha MO. Diagnosis and management of Chagas disease and cardiomyopathy. Nature Reviews Cardiology 2012;9(10):576–589. ISSN 17595002.

[2] Rojas LZ, Glisic M, Pletsch-Borba L, Echeverría LE, Bramer WM, Bano A, Stringa N, Zaciragic A, Kraja B, Asllanaj E, Chowdhury R, Morillo CA, Rueda-Ochoa OL, Franco OH, Muka T. Electrocardiographic abnormalities in Chagas disease in the general population: A systematic review and meta-analysis. PLoS Neglected Tropical Diseases 2018; 12(6):1–20. ISSN 19352735.

[3] Brito BOdF, Ribeiro ALP. Electrocardiogram in chagas disease. Revista da Sociedade Brasileira de Medicina Tropical 2018;51(5):570–577. ISSN 00378682.

[4] Siontis P, Noseworthy P, Attia Z, Friedman P. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nature Reviews Cardiology 2021; 18(7):465– 478.

[5] Moreno-Sánchez PA, García-Isla G, Corino VD, Vehkaoja A, Brukamp K, van Gils M, Mainardi L. ECG-based data-driven solutions for diagnosis and prognosis of cardiovascular diseases: A systematic review. Computers in Biology and Medicine 2024;172(February). ISSN 18790534.

[6] Liu X, Wang H, Li Z, Qin L. Deep learning in ECG diagnosis: A review. Knowledge Based Systems 2021;227:107187. ISSN 09507051.

[7] Zhao Z, Zhang Y. SQI quality evaluation mechanism of single-lead ECG signal based on simple heuristic fusion and fuzzy comprehensive evaluation. Frontiers in Physiology 2018;9(JUN):1–13. ISSN 1664042X.

[8] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SH. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods 2021;53(4):1689–1696. ISSN 15543528.

[9] Taconné M, Corino VD, Mainardi L. An ECG-Based Model for Left Ventricular Hypertrophy Detection: A Machine Learning Approach. IEEE Open Journal of Engineering in Medicine and Biology 2025;6:219–226. ISSN 26441276.

Address for correspondence:

Marion Taconné (marionhelene.taconne@polimi.it)
Politecnico di Milano,
Building 21, Via Camillo Golgi 39,
20133, Milan, Italy