On Spurious Features When Predicting Chagas

Matthew Dupont[1], Richard J Povinelli[2]

[1]Milwaukee, Wisconsin, US
[2]Marquette University, Milwaukee, Wisconsin, U.S.

Abstract

The paper explores ECG screening for Chagas, a parasitic disease found primarily in South America, as part of 'Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025'. Our approach is twofold: identify and suppress artifacts that are not indicative of Chagas and develop detection models based on power spectra and hand-engineered features. Across three datasets, SaMi-Trop (BR), CODE-15% (BR), and PTB-XL (DE), we demonstrate that known spurious features can identify Chagas status at an above chance rate using a single-feature logistic regression classifier. Additionally, we show that such features are picked up by more sophisticated machine learning methods. We identify and mask spurious features by evaluating single-feature models on a set of engineered features and frequency spectra. We compare the model error on samples of same disease state patients across datasets with a two-sample Kolmogorov-Smirnov (KS) test and reject features which fail to disprove the null hypothesis. We then mask identified spurious features and train a final decision tree from the remaining features. As team Blue and Gold, our submitted unmasked model achieved a challenge score of 0.112 on the official test set, for a rank of 37 out of 41.

1. Introduction

Chagas disease, caused by Trypanosoma cruzi, remains a major, underdiagnosed cause of cardiomyopathy across South America and increasingly into North America. An estimated 7 million people are infected, with approximately 10,000 deaths per year [1]. The single-cell parasite injects itself into muscle cells, including the cardiac muscle, where it multiplies and then bursts the cell, releasing the parasite. It is the change in the cardiac tissue caused by the parasite and thus also the propagation of the electromagnetic waves across it that we wish to detect via an ECG.

ECGs are attractive for Chagas disease screening: low-cost, non-invasive, and deployable at scale. The George B. Moody PhysioNet Challenge 2025 [2], [3] is to detect Chagas from the ECGs. We (team Blue and Gold) approach the challenge in two ways. First, we show that there are artifacts in the Challenge dataset indicative of the region from which the signals were collected, i.e., Brazil (SaMi-Trop [4] and CODE-15% [5]) vs. Germany (PTB-XL [6]). Second, we suppress the artifacts to build more robust models. This is important as all cases in the training dataset with Chagas are from Brazil, while the controls are mostly from Europe.

2. Related Work

Prior research has identified multiple ECG rhythm changes associated with Chagas disease. Right bundle branch block and left anterior fascicular block are associated with the transition of Chagas disease from the latent to chronic phase. Other common ECG features include atrioventricular blocks, preventricular contractions, atrial fibrillation, and ST and T Wave abnormalities [5]. These features are not specific to Chagas disease; multiple ECG abnormalities and comorbid conditions are often present, especially in older patients, complicating diagnosis [5], [7], [8].

Machine learning algorithms analyzing ECG signals generally fall into two categories. The first group uses discrete wavelet or Fourier transform to analyze spectral data and use that spectral data in traditional machine learning (ML) models, especially support vector machines. The second group applies neural networks to the signals, especially convolutional neural networks (CNNs) and long short-term memory (LSTM) models. Jidling et al demonstrated strong performance across Chagas disease datasets when applying ResNet to the CODE and SaMi-Trop datasets [9], [10].

3. Data

We used the three public 12-lead ECG datasets provided by the challenge organizers with distinct geographies: SaMi-Trop (Brazil; longitudinal Chagas cohort, positives) [4], CODE-15% (Brazil; stratified 15% subset of the Telehealth Network of Minas Gerais with weak/self-

reported Chagas labels; 345,779 exams from 233,770 patients, 2010–2016) [5], and PTB-XL (Germany; clinical 10s ECGs, used here as negatives) [6]. Meta data is also provided including gender, age, and information about the characteristics of the ECG leads.

4. Methods

Model training was performed in two phases; initial training on a subset of data to identify spurious features, followed by model training on input features masked to omit spurious features.

Initial training selected two disjoint sets of records; the first to train single-feature models and the second to evaluate whether the feature each single-feature model was trained on was spurious. The training set consisted of 500 records from each of the SaMi-Trop dataset, the CODE15% dataset (Chagas positive cases only), the CODE15% dataset (Chagas negative cases only), and the PTB-XL dataset for 2000 total records. The evaluation set consisted of 100 records selected from each of the same datasets.

439 single-feature LightGBM models were trained on the single-feature training dataset. Each model was trained using only one feature as a predictor for disease state and tested on the evaluation set. Probability estimates were used to calculate absolute model error. A two-sample Kolmogorov-Smirnov test was performed between the PTB-XL and CODE15% (Negative) subsets, and the SaMi-Trop and CODE15% (Positive) subsets to calculate the KS statistic and p value. An alpha of .05 was chosen, and a Holm-Bonferroni adjustment was applied to account for multiple tests. An activation mask was generated to omit features with a p value in either test below the Holm-Bonferroni-adjusted threshold value. Decision trees were trained on both the full feature set as well as the pruned feature set.

4.1. Artifact Identification

Since Brazil (SaMi-Trop) has primarily 60Hz line frequency, whereas Germany (PTB-XL) has 50Hz. These powerline effects can be observed when many record spectral power estimates are overlaid, as in Figure 1. We hypothesized that the power ratio of 50/60Hz would differentiate the datasets. Power spectral features were calculated using Welch's method at 1 Hz intervals on lead I. A simple one-feature logistic regression was analyzed using 10-fold cross-validation on the public training dataset, resulting in an accuracy of 65% and a challenge score of 0.088. To verify that a more sophisticated learner, a decision tree, with power features from 0 to 170 Hz, was trained.

This decision tree had an accuracy of 96.7% and challenge score of 0.179. However, the 60Hz feature was found in the second level of the tree, with other high-level decisions made based on frequencies close to 50 and 60Hz, indicating that a more sophisticated learner is misled by this artifact. This vulnerability informed the subsequent feature extraction and masking processes.

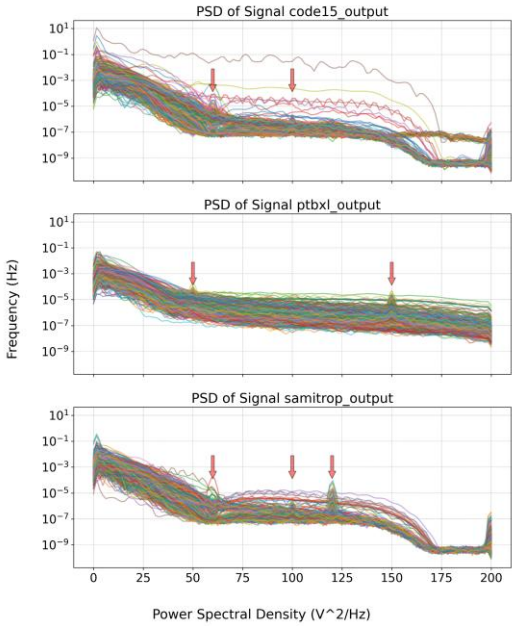


Figure 1. Overlaid welch frequency spectra for each dataset. Differences in power at 50Hz, 60Hz, 100Hz, 120Hz and 150Hz are highlighted.

4.2. Feature Extraction

The final model examined 429 total features split into frequency features and engineered features. Frequency features consisted of the spectral power of frequencies binned into 1Hz bins. The engineered features can further be broken down into duration-based features, amplitude-based features, and other features.

Duration and Amplitude based features were derived from peak identification using the neurokit package. Each ECG lead was cleaned via neurokit's ecg_clean function using the neurokit strategy. Peak detection was performed via neurokit's ecg_peaks and ecg_delineate to split out waveforms and identify salient peaks, onsets, and offsets. Incomplete waveforms at the start or end of each signal containing nans were removed. Isoelectric voltage was calculated for each waveform from the average of the PQ interval. J points were identified by smoothing the waveform with a gaussian filter, taking its derivative, and selecting the first point where the absolute value of the

derivative was below a threshold of .01, or the lowest value of the derivative if the threshold was never subceeded. Amplitude features were calculated on every waveform for every lead and each of the P, Q, R, S, and T peaks and ST elevation. The mean, max, min, and standard deviation (std) of these amplitudes were calculated and used as features.

Duration features were calculated from the mean, max, min, and std of the P wave, QRS complex, and T-wave durations, the PR interval, QT interval, and ST interval.

Other features calculated include a mix of demographic features (age, sex) overall signal mean and std, heart rate, heart rate variability (HRV) and heart axis. Heart axis was calculated by applying a vectorcardiogram transform using Kors coefficients [11] and taking the X, Y, and Z numbers from the amplitude at the R peak.

5. Results

Three models were considered for the challenge; the first training a frequency-only LightGBM decision tree trained on the frequency features only, a full-feature decision tree trained on both frequency features and engineered features, and a feature-masked model trained on all features not masked by the KS test. The full-feature model was selected for challenge evaluation, due to errors identified in the KS masking model shortly before the submission deadline. The second model achieved a final challenge score of .112, and challenge rank of 37 out of 41. The feature-masked model was evaluated on a stratified subset of 20% of the training data, used as a local test set. The locally trained feature elimination model achieved a challenge score of .234 on the held-out subset of the training data.

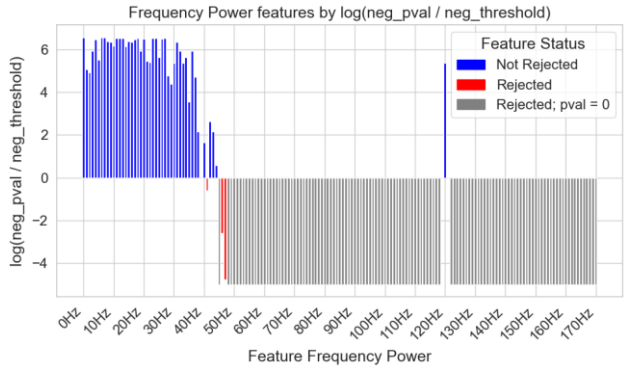


Figure 2. Log ratio of *p* value over Holm-Bonferroni adjusted threshold between Chagas Negative Code15 and PTBXL datasets.

p values and rejection status were recorded for each feature. No features were rejected on both the test of positive and negative disease state. Only the 8Hz power frequency feature failed the KS test between Chagas positive datasets. Frequency power features were evaluated by taking the log ratio of the negative disease state p-val over the Holm-Bonferroni adjusted threshold. Figure 2 shows these log ratios arranged by frequency, demonstrating high, non-reject p-values for features from 0HZ to 44Hz, sharply dropping off to 0 above 44Hz.

Only 13 engineered features were rejected by the negative disease state KS test, summarized in Table 1.

Feature category	Rejected Features
Frequency Power	41Hz, all features above 44Hz except 120Hz;
Duration	T Wave mean and min, P wave mean, QT interval mean
Lead I Amplitude	R Amplitude mean
Lead II Amplitude	P Amplitude mean
Lead AVL Amplitude	R Amplitude min
Lead V1 Amplitude	Q Amplitude mean
Lead V3 Amplitude	P Amplitude mean
Lead V6 Amplitude	Q Amplitude min, S Amplitude min
Other	signal_mean, signal_std

Table 1. Summary of features rejected by KS testing.

6. Discussion and Conclusions

The PTB-XL and CODE15% datasets are geographically and temporally distant, with larger differences in the distributions of demographic features than the SaMi-Trop and CODE15% datasets. As such, we expect to find fewer features able to robustly distinguish the Chagas-positive patient cohorts than the Chagas-negative cohorts. This is borne out by the data, suggesting the KS test can identify spurious features.

The high confidence in rejecting frequency features above 44Hz, as well as signal_std and signal_mean, may be secondary to differences in noise present in the CODE15% and PTB-XL dataset. In Figure 1, the PTB-XL demonstrates a broader range of spectral power above 60Hz. This broader range seems to be sufficient to learn dataset-specific features and suggests denoising may be crucial to generalizing models leveraging high frequency signals. While the highlighted and known spurious features were rejected by the KS test as anticipated, the broad rejection of high-frequency power features fails to

distinguish those features.Clearer patterns are less obvious in the engineered features, as relatively fewer of these features were identified as spurious. Further, mean, min, max and standard deviation are all correlated. Reducing these features to only mean and std may prove more effective.

Our approach examines the effect of only single-feature models and does not consider interaction effects. Examining each interaction effect presents a combinatorial increase in required training and analysis, limiting the feasibility of this approach. We suggest a possible approach to address these limitations would be to identify interaction effects in a model trained on all features and filtering only those interaction effects above a threshold of feature importance.

We suggest the incorporation of the predictive power of convolutional networks as a possible future direction. One approach may be to train convolutional networks to recognize known ECG morphological features such as the characteristic R-R' pattern in lead V1 associated with right bundle branch block and incorporate the predictions of these models as features to a further model. A second approach could be in applying convolutional autoencoders to reduce ECG signals to a latent space, then leveraging the encoder to produce those latent values as features.

References

[1] Chagas disease. Accessed: Aug. 25, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)

[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 ALP, 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, Weigle J, Saghafi S, Gomes P, Elola A, Hassannia, MS, Campbell, K, Bahrami Rad A, Ribeiro AH, 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, DOI: 10.48550/arXiv.2510.02202.

[4] C. S. Cardoso *et al.* Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile; In *BMJ Open*, vol. 6, no. 5, p. e011181, May 2016, doi: 10.1136/bmjopen-2016-011181.

[5] M. C. P. Nunes *et al.* Chagas cardiomyopathy: An update of current clinical knowledge and management: A scientific statement from the American Heart Association; In *Circulation*, vol. 138, no. 12, Sept. 2018, doi: 10.1161/CIR.0000000000000599.

[6] P. Wagner *et al.* PTB-XL, a large publicly available electrocardiography dataset; In *Sci. Data*, vol. 7, no. 1, p. 154, May 2020, doi: 10.1038/s41597-020-0495-6.

[7] L. Z. Rojas *et al.* Electrocardiographic abnormalities in Chagas disease in the general population: A systematic review and meta-analysis; In *PLoS Negl. Trop. Dis.*, vol. 12, no. 6, p. e0006567, June 2018, doi: 10.1371/journal.pntd.0006567.

[8] J. Soverow *et al.* Progression of Baseline Electrocardiogram Abnormalities in Chagas Patients Undergoing Antitrypanosomal Treatment; In *Open Forum Infect. Dis.*, vol. 6, no. 2, p. ofz012, Feb. 2019, doi: 10.1093/ofid/ofz012.

[9] C. Jidling *et al.* Screening for Chagas disease from the electrocardiogram using a deep neural network; In *PLoS Negl. Trop. Dis.*, vol. 17, no. 7, p. e0011118, July 2023, doi: 10.1371/journal.pntd.0011118.

[10] P. Pantelidis, M. Bampa, E. Oikonomou, and P. Papapetrou. Machine learning models for automated interpretation of 12-lead electrocardiographic signals: a narrative review of techniques, challenges, achievements and clinical relevance; In *J. Med. Artif. Intell.*, vol. 6, pp. 6–6, May 2023, doi: 10.21037/jmai-22-94.

[11] J. A. Kors, G. Van Herpen, A. C. Sittig, and J. H. Van Bemmel. Reconstruction of the Frank vectorcardiogram from standard electrocardiographic leads: Diagnostic comparison of different methods; In *Eur. Heart J.*, vol. 11, no. 12, pp. 1083–1092, Dec. 1990, doi: 10.1093/oxfordjournals.eurheartj.a059647.

Address for correspondence:

Matthew Dupont
W172S7811 Lannon Drive, Muskego, Wisconsin, U.S.
matthewdupont11@gmail.com