Feature-Reduced Ensemble Model for ECG-Based Chagas Disease Diagnosis

Eunseo Choi, Sanghyun Ham

Department of Biomedical Engineering, Kyung Hee University, Yongin, Republic of Korea

Abstract

Team KHU_BME developed a machine learning (ML) model for early detection of Chagas disease using large-scale 12-lead ECG data. Chagas disease, caused by Trypanosoma cruzi, is often underdiagnosed in endemic regions where molecular or serological testing is limited. ECG offers a low-cost, non-invasive alternative that can capture conduction abnormalities of chronic Chagas cardiomyopathy.

From an initial 109 features, morphological, temporal, and spectral descriptors were extracted and reduced to 44 clinically relevant features, such as QRS duration and rsR' patterns. This feature reduction improved generalizability, efficiency, and interpretability. Class imbalance was addressed with SMOTE, and hyperparameters were tuned for Random Forest, XGBoost, and Logistic Regression classifiers.

The ensemble model achieved a Challenge score of 0.139, AUROC 0.817, AUPRC 0.718, Accuracy 0.746, and F-measure 0.645 on our held-out test set, and a Challenge score of 0.094 on the official test set. These results demonstrate the feasibility of ECG-based ML with feature reduction as an efficient screening tool for Chagas disease in resource-limited settings.

1. Introduction

Chagas disease, caused by Trypanosoma cruzi, remains prevalent in Latin America, with many patients progressing to chronic stages without timely diagnosis. Chronic Chagas cardiomyopathy is a major cause of heart failure, arrhythmias, and sudden cardiac death, highlighting the need for early detection. Definitive serological or molecular tests are costly and infrastructure-dependent, limiting use in resource-limited settings, whereas electrocardiography (ECG) is a low-cost, non-invasive tool for detecting conduction abnormalities. Here, we developed a machine learning model using large-scale 12-lead ECG data, extracting morphological and spectral features, optimizing hyperparameters, and proposing an ensemble classifier. Comparative analyses with baselines demonstrated the feasibility of ECG-based ML as a scalable screening strategy.

1.1. Clinical background

Chagas disease, caused by Trypanosoma cruzi, remains a major endemic condition in Latin America. The infection often progresses silently, but once chronic cardiomyopathy develops, it leads to heart failure, arrhythmias, and sudden death, driving most of the disease's morbidity and mortality.

1.2. Limitations of current diagnostic approaches

Definitive diagnosis depends on serology or PCR, which are accurate but costly and infrastructure-dependent, limiting early large-scale screening in endemic low-resource areas and contributing to delayed detection and poor outcomes.

1.3. Motivation for ECG-based machine learning

Electrocardiography (ECG) is a non-invasive, inexpensive, and widely available tool capable of detecting conduction abnormalities characteristic of Chagas cardiomyopathy, making it well-suited for early screening in endemic regions. Recent advances in machine learning enable automated recognition of disease-specific morphological and spectral ECG patterns; however, high-dimensional feature sets in imbalanced datasets risk overfitting and poor generalizability. To address this, we implemented feature reduction, retaining clinically meaningful and statistically robust descriptors while eliminating redundancy, thereby improving computational efficiency, enhancing interpretability, and aligning the model with established pathophysiology.

2. Methods

We developed a machine learning–based diagnostic model for early detection of Chagas disease using large-scale 12-lead ECG datasets (CODE-15 and SaMi-Trop). Morphological and spectral features were extracted, multiple algorithms were benchmarked, and hyperparameters optimized(details of the optimization methods will be further discussed in Section 2.8). An ensemble strategy was employed, yielding superior performance compared to individual models.

2.1. Data sources

Two datasets were used: CODE-15, comprising a heterogeneous population with diverse cardiovascular conditions, and SaMi-Trop, containing exclusively Chagas-positive cases to address class imbalance. Records were included only if headers and signals were valid; corrupted channels or ambiguous labels were excluded.

2.2. Preprocessing

All ECGs were resampled to 400 Hz and bandpass filtered (0.5–40 Hz) to suppress baseline drift, muscle noise, and powerline interference. Signals were normalized per lead to minimize inter-patient amplitude variability and ensure robust feature extraction.

2.3. Feature extraction

We extracted 109 ECG features spanning HRV, morphology, axis, spectral, and statistical measures, as summarized in Table 1, and applied importance-based pruning to retain only the most informative variables.

Table 1. Extracted ECG features grouped by category.

Group	Number of features
HRV (Heart Rate Variability)	7
Morphological	10
Axis	3
PSD (Power Spectral Density)	48
Spectral	9
ZCR (Zero-Crossing Rate)	8
EVM(Energy/Variance/Median Absolute Deviation(MAD))	24
Total	109

2.4. Feature reduction

To reduce overfitting and enhance generalizability, we applied a systematic feature reduction pipeline. Starting from 109 features, permutation analysis and model-based importance ranking with Random Forest, XGBoost, and Logistic Regression guided the elimination of low-importance and redundant variables. The final 44-feature set preserved clinically meaningful descriptors, including QRS duration and rsR′ morphology, thereby improving dimensionality, computational efficiency, and interpretability without loss of performance.

2.5. Model development

Three base classifiers were constructed: Random Forest (RF), Extreme Gradient Boosting (XGB), and Logistic Regression (LR). Each model was trained on the reduced feature set, and their probability outputs were combined via weighted averaging, leveraging the complementary advantages of tree-based and linear models.

2.6. Training protocol

Model training employed stratified k-fold cross-validation to ensure balanced representation of positive and negative cases across folds. Hyperparameter optimization was performed for each classifier, and oversampling with SMOTE was applied to address residual class imbalance. This protocol maximized robustness and reproducibility while mitigating bias from data heterogeneity.

2.7. Evaluation metrics

Performance was assessed using the official Challenge metrics: Challenge score (primary), AUROC, AUPRC, Accuracy, and F1. Cross-validation results were reported as mean ± SD, and final evaluation on the held-out and official test sets determined leaderboard ranking.

2.8. Hyperparameter optimization

Hyperparameters were optimized via grid search with F1 as objective. Only non-default values are reported in Table 2, as they consistently improved cross-validation performance.

Table 2. Final optimized hyperparameters: Random Forest used 800 trees with constrained depth for stability, XGBoost applied a low learning rate and class imbalance correction for rare positives, and Logistic Regression with 2000 iterations ensured a stable baseline.

Model	Parameter	Value
Random Forest (RF)	n_estimators	800
	max_depth	12
	min_samples_split	10
	min_samples_leaf	4
XGBoost (XGB)	n_estimators	400
	learning_rate	0.05
	max_depth	6
	colsample_bytree	0.6
	reg_lambda	0.5

	scale_pos_weight	neg/pos ratio
Logistic Regression (LR)	C	10.0
	max_iter	2000

3. Result

Detailed performance metrics for the ensemble model under different evaluation protocols (cross-validation, validation, internal held-out test, and official Challenge test) are collectively presented in Table 3.

3.1. Cross-validation results (CV)

In 10-fold cross-validation on the training set, the ensemble achieved a Challenge score of 0.1285 ± 0.0062, AUROC of 0.8211 ± 0.0134, AUPRC of 0.7010 ± 0.0214, Accuracy of 0.7254 ± 0.0134, and F-measure of 0.6595 ± 0.0164.

3.2. Validation results (V)

On the validation subset, the ensemble achieved a Challenge score of 0.139, AUROC of 0.817, AUPRC of 0.718, Accuracy of 0.746, and F-measure of 0.645.

3.3. Internal validation (IV)

Evaluation on an internal held-out subset of the training data yielded slightly higher performance, with a Challenge score of 0.139, AUROC of 0.830, AUPRC of 0.732, Accuracy of 0.755, and F-measure of 0.656. These results were consistent with the cross-validation and validation subsets, indicating stable generalization.

3.4. Official Challenge score (CS)

On the official hidden test set, our submission achieved a Challenge score of **0.094**, which determined our final leaderboard ranking.

Table 3. Performance results across different evaluation subsets. (Abbreviations: CS = Challenge score; ACC = Accuracy; F1 = F-measure.)

set	CS	AUROC	AUPRC	ACC	F1
CV	0.1285± 0.0062	0.8211± 0.0134	0.7010± 0.0214	0.7254± 0.0134	0.6595± 0.0164
V	0.139	0.817	0.718	0.746	0.645
IV	0.139	0.830	0.732	0.755	0.656
CS	0.094	-	-	-	-

3.5. Feature importance and ROC analysis

Figure 1. Normalized feature importance identified SDNN and QRS duration in lead V2 as the top predictors, followed by axis, ST/T, and spectral features. These reflect key pathophysiological patterns of Chagas cardiomyopathy, including reduced HRV, conduction delay, and repolarization abnormalities.

Figure 2. The out-of-fold ROC curve yielded AUROC=0.8209, indicating robust discrimination with high sensitivity at low false positive rates, suitable for screening use.

Figure 1.

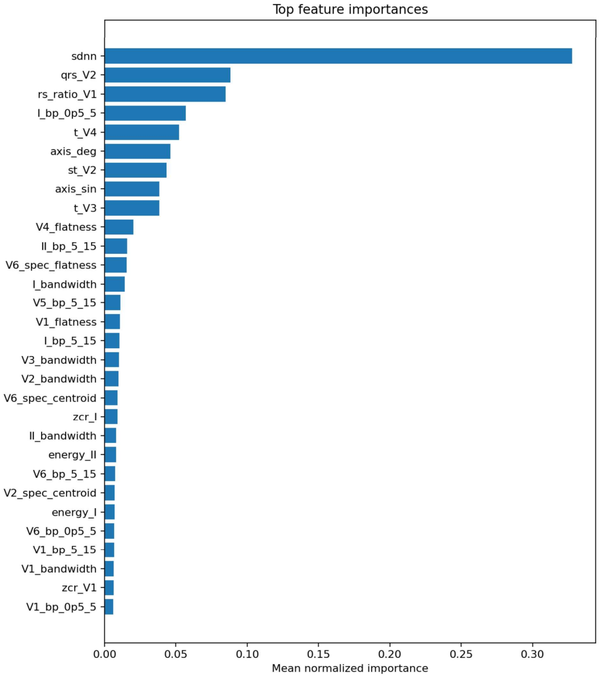
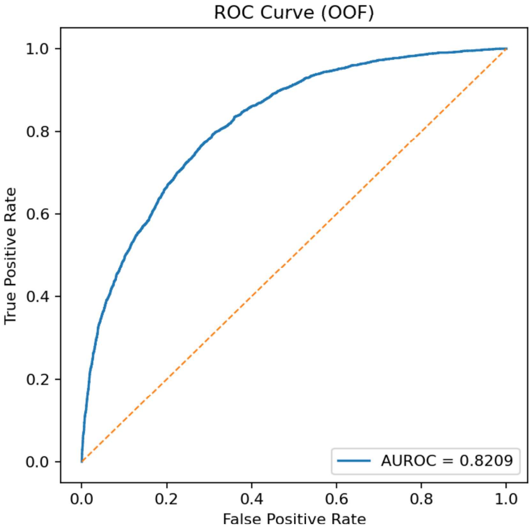


Figure 2.



4. Discussion

4.1. Interpretation of key findings

The feature importance analysis demonstrated that HRV (e.g., SDNN) and QRS duration in V2 were the most influential predictors. rsR′ in right precordial leads, axis deviation, and ST/T abnormalities also ranked highly, corroborating known conduction and repolarization changes in chronic Chagas disease. These findings highlight recognized ECG hallmarks of Chagas disease, supporting the model's pathophysiological plausibility.

The out-of-fold AUROC was 0.8209, showing strong discrimination with sensitivity at low false positive rates. This suggests the model identifies physiologically meaningful features while achieving robust accuracy for screening in imbalanced populations.

4.2. Advantages of feature reduction

By discarding redundant or noisy variables, the model achieved improved stability and generalizability compared to the full feature set.

4.3. Limitations

The primary limitation of this study lies in the class imbalance and potential sampling bias across datasets. Although SMOTE was applied, oversampling may introduce synthetic artifacts. Additionally, the external generalizability to non-Brazilian or non-Latin American populations remains untested.

4.4. Future work

Future directions include the integration of deep learning approaches to capture raw waveform representations, external validation on independent cohorts, and deployment in real-world clinical screening scenarios. Combining lightweight ML with embedded hardware could further support point-of-care applications. Moreover, we aim to explore novel physiological correlations that may enable ECG-based screening of early-stage Chagas patients, reflecting our commitment as biomedical engineering students to move beyond code optimization and engage with the fundamental challenges of Chagas disease.

5. Conclusion

We proposed a machine learning–based ECG model for the early detection of Chagas disease, with feature reduction as the central contribution. The ensemble classifier achieved competitive performance on both cross-validation and the official Challenge test set. Findings support ECG-based feature reduction as an effective screening tool in resource-limited settings.

Acknowledgments

We thank our colleagues and mentors at Kyung Hee University for their support and guidance. This work was supported by the Biomedical Engineering Department.

References

[1] M. F. Braggion-Santos, H. T. Moreira, G. J. Volpe, M. Koenigkam-Santos, J. A. Marin-Neto, and A. Schmidt, "Electrocardiogram abnormalities in chronic Chagas cardiomyopathy correlate with scar mass and left ventricular dysfunction as assessed by cardiac magnetic resonance imaging," J. Electrocardiol., vol. 72, pp. 66–71, May–Jun. 2022.

[2] A. L. Ribeiro, M. P. Nunes, M. M. Teixeira, and M. O. Rocha, "Diagnosis and management of Chagas disease and cardiomyopathy," Nat. Rev. Cardiol., vol. 9, no. 10, pp. 576–589, Oct. 2012.

[3] C. Jidling, D. Gedon, T. B. Schön, C. D. L. Oliveira, C. S. Cardoso, A. M. Ferreira, L. Giatti, S. M. Barreto, E. C. Sabino, A. L. P. Ribeiro, and A. H. Ribeiro, "Screening for Chagas disease from the electrocardiogram using a deep neural network," PLoS Negl. Trop. Dis., vol. 17, no. 7, p. e0011118, Jul. 2023.

[4] Y. Ha, S. Lee, J. Lim, K. Lee, Y. E. Chon, J. H. Lee, and H. C. Lee, "A machine learning model to predict de novo hepatocellular carcinoma beyond year 5 of antiviral therapy in patients with chronic hepatitis B," Liver Int., Dec. 2024, doi: 10.1111/liv.16139.

[5] Y. Choi, J. Park, H. Kim, et al., "Artificial intelligence models predicting abnormal uterine bleeding after COVID-19 vaccination," Sci. Rep., vol. 15, p. 7081, 2025.

Address for correspondence:

Eunseo Choi
Department of Biomedical Engineering
Kyung Hee University
1732 Deogyeong-daero, Giheung-gu
Yongin-si, Gyeonggi-do, 17104
Republic of Korea
E-mail: choicerecord@khu.ac.kr

Sanghyun Ham
Department of Biomedical Engineering
Kyung Hee University
1732 Deogyeong-daero, Giheung-gu
Yongin-si, Gyeonggi-do, 17104
Republic of Korea
E-mail: ham5991@khu.ac.kr