Ensemble Learning with Early Fusion of Kernel-Transformed and Classical Electrocardiogram Features for Chagas Disease Detection

Victor Li, Runze Yan, Alex Fedorov, Jiaying Lu

Emory University, Atlanta, GA, USA

Abstract

The electrocardiogram (ECG) offers an accessible and non-invasive assessment of human health. Chagas disease, which affects nearly 6.5 million people across Central and South America, is known to have symptoms that appear in ECGs. Using time-series machine learning techniques, critical information can be extracted from these ECGs to detect Chagas disease as opposed to serological tests. As part of the George B. Moody PhysioNet Challenge 2025, we developed a classification approach consisting of two components: (1) a multi-view representation of 12-lead ECGs; (2) ensemble classification. Our team, GAIN-ECG, developed a novel approach that combines kernel-based feature extraction through MiniRocket with classical signal features, such as Heart Rate Variability (HRV), Discrete Wavelet Transform (DWT), and Fast Fourier Transform (FFT) features, through early fusion. We then employ an ensemble framework to classify the onset of Chagas disease. Testing against a held-out subset of the public training set, our model achieved a challenge score of 0.481, AUROC of 0.880, and F1 of 0.113. On the hidden validation set, our model received a challenge score of 0.090.

1. Introduction

As team GAIN-ECG, we participated in the 2025 George B. Moody PhysioNet Challenge. This challenge invited teams to develop automated and open-sourced algorithms for classifying cases of Chagas from electrocardiograms (ECG) [1, 2]. Although ECG-based diagnoses of Chagas disease can often be inaccurate, they can inform the use of limited and invasive serological tests.

Our team's Challenge entry tackles this classification task through a novel two-part approach utilizing multi-view representations of 12-lead ECGs and ensemble learning. The core of our approach lies in feature engineering and extraction. Assuming high performance can be achieved by our downstream classifier, our model's performance hinges on the richness and reliability of the features.

Training and validating on the Challenge's public training set and hidden validation set [3–7], we found that cross-ECG machine batch effects significantly undermined the reliability of our feature extraction process. To address this, we utilize classical ECG features that directly capture signal-specific characteristics, rather than relying on pretrained foundation models, which risk encoding batch-dependent features that worsen performance when evaluated on unseen sources. Also, we implement a convolutional kernel-based encoder to extract additional features, which are incorporated with early fusion.

2. Methods

Our two-stage multi-view ensemble classification framework consist of two key components: (1) multi-view ECG representation learning, and (2) an ensemble classifier. The multi-view representation learning module [8] is designed to capture the characteristics of input ECG signals from multiple, complementary perspectives. The ensemble classifier [9] follows a similar principle, leveraging diverse base classifiers with different inductive biases to enhance predictive performance. As shown in Figure 1, our framework begins by extracting classical features from three different views: Heart Rate Variability [10] (HRV), Discrete Wavelet Transform [11] (DWT), Fast Fourier Transform [12] (FFT), and convolutional kernel-based representation [13] (ConvK). The multi-view features are then concatenated to be feed into the ensemble classifier [14], which utilizes base classifiers with different induction bias.

2.1. Multi-view ECG Representation Learning

We denote one raw input 12-lead 400Hz ECG record as $\mathbf{x} \in \mathbb{R}^{\mathbb{L}\times\mathbb{T}}$, where $L = 12$ and T is originally variable depending on the record length (*e.g.*, 7 or 10 seconds). To maintain consistency across samples, each signal is either truncated or zero-padded to a fixed length of $T = 4096$.

Heart Rate Variability Representation. We opt for HRV as one view of ECG representation, which quantifies the variation in time intervals between consecutive heartbeats. To extract the HRV features, we first detect R-peaks in

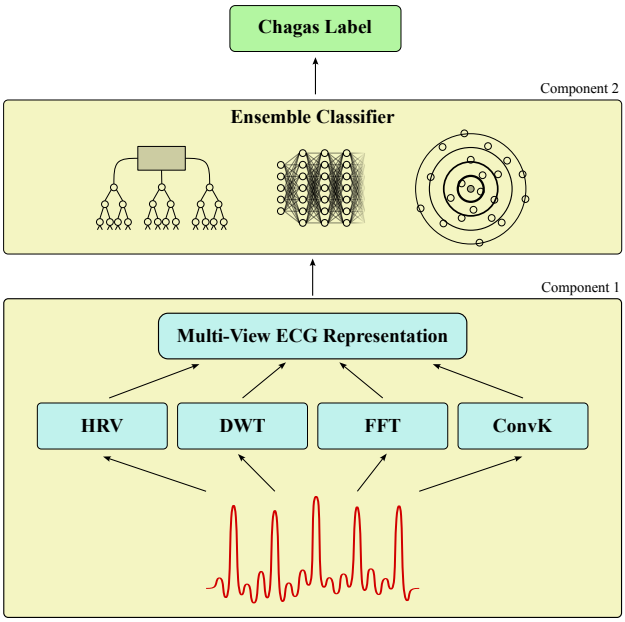

Figure 1. **Overview of proposed framework**. Every ECG signal is first processed through the feature extractors in the first component to create multi-view ECG representations. Then, Chagas diagnosis is determined by an ensemble classifier in the second component.

each ECG record (the tallest spikes in the QRS complex). For each individual lead $\mathbf{x}^{(i)} \in \mathbb{R}^{\mathbb{T}}$, this is achieved by finding a set $\mathbf{C}$ of all local-maxima in the signal:

$$\mathbf{C} = \{t \in \{2, \ldots, T-1\} : \mathbf{x}^{(i)}[t\text{-}1] < \mathbf{x}^{(i)}[t] > \mathbf{x}^{(i)}[t+1]\}. \quad (1)$$

To ensure the plausibility of our peaks, we enforce a minimum distance of 200 ms between consecutive peaks, which corresponds to a maximum heart rate of 300 bpm. R-peaks are then greedily selected from C under this constraint. Based on the obtained R-peaks, we derive the following time-domain HRV features: (1) number of peaks ($\bar{\mathrm{R}}$), (2) mean RR intervals ($\bar{\mathrm{R}}\bar{\mathrm{R}}$) that capture the beat-to-beat timing; (3) Standard Deviation of RR intervals (SDNN) that reflects overall heart rate variability; (4) Root Mean Square of Successive Differences (RMSSD) that emphasizes short-term variability between consecutive beats. The sequence of RR intervals are obtained by $RR_i = \mathbf{C}_{R_{i+1}} - \mathbf{C}_{R_i}$, where $\mathbf{C}_{R_{i+1}}$ denotes the time of the i-th R-peak in $\mathbf{C}$. After concatenating HRV features from all 12 leads, we obtain the HRV view representation as $\mathbf{h}_{\text{HRV}} = \bar{\mathrm{R}} \oplus \bar{\mathrm{R}}\bar{\mathrm{R}} \oplus \text{SDNN} \oplus \text{RMSSD}$.

Discrete Wavelet Transform Representation. Applying a 4-level DWT decomposition to each individual lead using Daubechies-4 (db4) wavelet, we get detailed coefficients d_ℓ at levels $\ell = 1, 2, 3, 4$ (ignoring approximation coefficients). The energy at level ℓ is

$$E_\ell(\mathbf{x}^{(i)}) = \sum_{k=1}^{n_\ell} d_\ell[k]^2, \ \ell = 1, 2, 3, 4, \quad (2)$$

where d_ℓ are the detail coefficients of the signal $\mathbf{x}^{(i)}$ at level ℓ after DWT, n_ℓ is the number of coefficients, and k is the index of the coefficient within the level. These energy features capture the power of oscillations across 4 different frequency scales. Processing each lead individually and concatenating, the DWT representation can be denoted as $\mathbf{h}_{\text{DWT}} = \mathbf{E_1}(\mathbf{x}) \oplus \mathbf{E_2}(\mathbf{x}) \oplus \mathbf{E_3}(\mathbf{x}) \oplus \mathbf{E_4}(\mathbf{x})$.

Fast Fourier Transform Representation. To extract the FFT features, we first subsample signals down to 200Hz to reduce computational complexity while preserving the dominant oscillatory components. This allows the model to capture global spectral patterns that may be overlooked by R-peak statistics or wavelet energies. Applying the FFT [12] to each subsampled signal of length N yields a set of complex coefficients $\mathbf{B}$ and corresponding frequency bins $\mathbf{F}$. Since ECG signals are real-valued, their FFTs are symmetric about zero, and we retain only the non-negative frequencies. The amplitudes $\mathbf{A}$ are computed by

$$\mathbf{A}_i = \frac{2}{N} \cdot |\mathbf{B}_i|, \quad \text{for } \mathbf{F}_i \geq \mathbf{0}. \quad (3)$$

The resulting representation is formed by concatenation across all leads as $\mathbf{h}_{\text{FFT}} = \mathbf{F} \oplus \mathbf{A}$.

Convolutional Kernel-based Representation. We further apply convolutional kernels to ECG signals to capture meaningful patterns such as shape, frequency, and variance. This is achieved through the MiniRocket encoder [13], which uses a set of fixed kernels applied at multiple temporal scales. Each kernel response is summarized

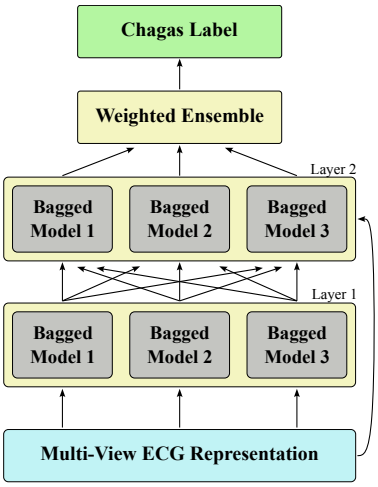

Figure 2. **Overview of AutoGluon Ensemble Classifier**. Base models are first trained with bagging, then organized into stacked layers.

using the Proportion of Positive Values (PPV) statistic,

$$\text{PPV}_{k,d} = \frac{1}{T}\sum_{t=1}^{T} 1((\mathbf{x}^{(i)} * k_d)[t] > q_{k,d}). \quad (4)$$

where $(\mathbf{x}^{(i)} * k_d)[t]$ is the convolution of the signal $\mathbf{x}^{(\mathbf{i})}$ with kernel k at dilation d, and $q_{k,d}$ is a bias threshold. In MiniRocket, dilations d are selected to evenly span a range of temporal scales, allowing each kernel to capture patterns across temporal scales. Bias thresholds $q_{k,d}$ are computed as quantiles of the convolution outputs on the training data, ensuring that each kernel-dilation pair produces a balanced distribution of activations.

We configure the encoder with 504 kernels, which correspond to MiniRocket's base set of 84 kernels applied across 6 different dilations. Since ECGs are multivariate, kernel outputs are aggregated before computing their PPVs. These PPV values then form the convolutional kernel-based view representations as $\mathbf{h}_{\text{ConvK}} = \mathbf{Q}$.

2.2. Ensemble Classifier

From the multi-view ECG representation learning, we obtain the final multi-view vector representation $\mathbf{h} = \mathbf{h}_{\text{HRV}} \oplus \mathbf{h}_{\text{DWT}} \oplus \mathbf{h}_{\text{FFT}} \oplus \mathbf{h}_{\text{ConvK}} \oplus \mathbf{h}_{\text{Stat}}$, where we further calculate the means and standard deviations of each individual ECG lead to obtain $\mathbf{h}_{\text{Stat}}$. For our classification task, we employ the state-of-the-art ensemble tabular predictor f_Θ via *AutoGluon-V1.3.0* (AG) [14], thus $\hat{y} = f_\theta(\mathbf{h})$. In a nutshell, the ensemble classifier f_θ incorporate a set of M weak classifiers to make distinct predictions $\hat{y}_i = f_{\theta,i}(\mathbf{h})$, and the final ensemble prediction is obtained through a greedy weighted combination:

$$\hat{y} = \sum_{i=1}^{M} w_i f_{\theta,i}(\mathbf{h}), \quad (5)$$

where w_i are non-negative weights optimized by the ensemble algorithm [15]. A wide range of weak classifiers is included: KNN, RF, boosting models (*e.g.*, CatBoost, XGBoost, GBM), and tabular neural networks. Specifically, we configured AG to the "best quality" preset and impose a 24-hour time limit on training. This further introduce advanced ensemble techniques over Eq. (5) by multi-layer stack ensembling and repeated k-fold bagging techniques [14], as can be seen in Figure 2. The multi-layer stack extend the traditional ensemble framework by allowing adding more layers over the first layer of base models, sharing the same principle as a multi-layer neural network. A skip connection [16] of multi-view ECG representation is added to further augment the second layer input. The repeated k-fold bagging [17] further improve the prediction performance of each base model by fitting *k* copies of each base model with a different data chunk held-out from each copy.

Model Name	Training			Validation
	Score	AUROC	AUPRC	Score
RF	0.266	0.802	0.087	0.062
ConvK+RF	0.354	0.821	0.124	N/A
ConvK+AG	0.399	0.836	0.168	N/A
Multi-View+AG	**0.410**	**0.853**	**0.177**	**0.090**
Multi-View+AG+BE	0.426	0.861	0.205	0.085
ECGFounder+AG+BE	0.481	0.880	0.242	0.054

Table 1. **Ablation study illustrating incremental improvements** in performance as kernel-based features, the AutoGluon ensemble, and additional signal features are introduced into the model.

Training	Validation	Test	Ranking
0.410	0.090	TBA	267/346

Table 2. **Challenge scores for our selected entry** (team GAIN-ECG), including the ranking of our team on the hidden **validation** set. We used an internal 80%/20% split on the public training set, repeated scoring on the hidden validation set, and placeholder scoring on the hidden test set.

3. Results

Using the publicly available dataset, we performed an ablation study with an 80%/20% train/test split to track the benefit of each component we added to our model. For each model, we compute the training challenge score, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). First, we began with a basic RF on signal means and standard deviations. Second, we incorporated the convolutional kernel-transformed features (ConvK) from MiniRocket. Third, we switched the classifier from RF to AG's ensemble framework. Fourth, we fused the additional signal features with the ConvK features. In addition, we experimented with statistical batch-effect elimination techniques (BE), such as z-score normalization and data weighting, and explored replacing the MiniRocket encoder with an ECG foundation model (ECGFounder [18]).

Based on the ablation study, we finalized our approach by fusing features from multiple ECG representations and classifying the resulting multi-view feature vectors with an AG-based ensemble classifier. During the unofficial phase of the challenge, this model achieved a score of 0.564. During the official phase, our model was then validated and ranked with repeated scoring on the hidden validation set, as shown in Table 2.

4. Discussion and Conclusions

By combining multi-view ECG representations with ensemble modeling, our approach was able to achieve a higher challenge score than the baseline random forest model and ECG foundation model. The integration of clas-

sical signal features with kernel-based features provided a richer representation of ECG to a robust ensemble framework. However, a comparison of the training and validation scores highlights that our model struggles to generalize across datasets. First, the high AUROC values are misleading in the context of the strong class imbalance in Chagas disease. While the elevated training challenge score suggests the model is accurate at its top 5% most confident predictions, the low AUPRC reveals that it generates many false positives at lower thresholds.

One possible reason for this lack of generalization is the composition of the training set. The public training set comprises three distinct datasets: two small datasets containing only positives or negatives, and a larger mixed set. This structure likely led our model to exploit dataset-specific artifacts rather than features that are truly indicative of Chagas disease. Consequently, our model was most confident and accurate when classifying samples from the two smaller homogeneous datasets. Because of the 5% threshold, lower-confidence predictions from the larger mixed dataset were excluded from the challenge score, inflating the training score. However, when evaluated on the hidden validation set, this reliance on dataset-specific features lead to a sharp drop in performance since the validation data came from entirely different sources. To address the batch effect, we experimented with data weighting, z-score normalizations, and foundation model embeddings. While these adjustments improved performance on the training set, they did not translate into better generalization on the held out validation set. Future work should investigate more robust approaches, such as adversarial training, to mitigate these dataset-specific biases and improve generalization for classifying highly imbalanced classes.

Acknowledgments

This work was supported in part by the Emory Undergraduate Research Programs and Emory School of Nursing's Center for Data Science.

References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. 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, et al. Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025. Computing in Cardiology 2025;52:1–4.

[3] 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.

[4] 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.

[5] 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.

[6] Nunes MCP, Buss LF, Silva JLP, Martins LNA, Oliveira CDL, Cardoso CS, 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.

[7] 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.

[8] Kataria S, Xiao R, Ruchti T, Clark MT, Lu J, Lee R, et al. Continuous cardiac arrest prediction in icu using ppg foundation model. In EMBC. 2025; .

[9] Xu Y, Wang X, Lu J, Ding S, Defu Cao HY, Liu Y, et al. Enecg: Efficient ensemble learning for electrocardiogram multi-task foundation model. In AMIA. 2025; .

[10] Malik M. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. Circulation 03 1996;93:1043–1065.

[11] Mallat S. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 1989;11(7):674–693.

[12] Brigham EO, Morrow RE. The fast fourier transform. IEEE Spectrum 1967;4(12):63–70.

[13] Dempster A, Schmidt DF, Webb GI. Minirocket: A very fast (almost) deterministic transform for time series classification. In SIGKDD, KDD '21. August 2021; 248–257.

[14] Erickson N, Mueller J, Shirkov A, Zhang H, Larroy P, Li M, et al. Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint arXiv200306505 2020;.

[15] Caruana R, Niculescu-Mizil A, Crew G, Ksikes A. Ensemble selection from libraries of models. In ICML. New York, NY, USA, 2004; 18.

[16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In CVPR. 2016; 770–778.

[17] Parmanto B, Munro PW, Doyle HR. Reducing variance of committee prediction with resampling techniques. Connection Science 1996;8(3-4):405–426.

[18] Li J, Aguirre A, Moura J, Liu C, Zhong L, Sun C, et al. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. arXiv preprint arXiv241004133 2024;.

Address for correspondence:

Jiaying Lu, PhD
1520 Clifton Road, Suite 245, Atlanta, GA 30322, USA
jiaying.lu@emory.edu