Wavelet-Derived Entropy and Complexity Biomarkers for ECG-Based Detection of Chagasic Cardiomyopathy

G V Clemente[1,2,3], L R Andrini[2,3], M Llamedo Soria[1]

[1]Universidad Tecnológica Nacional, Buenos Aires, Argentina
[2]CONICET, Argentina
[3]CMaLP, Universidad Nacional de La Plata, La Plata, Argentina

Abstract

As part of the Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025, our team ***CompleXformers*** *developed an interpretable ECG-based approach to identify Chagasic cardiomyopathy (CCM) from standard 12-lead ECGs. The method constructs a principal-component beat (PCB) through R-peak alignment and singular-value decomposition, followed by a wavelet-based entropy–complexity analysis using the Daubechies-6 wavelet. From the relative wavelet energy distribution, two biomarkers are computed: the normalized Shannon entropy (H) and the wavelet statistical complexity (C), which together characterize morphological and spectral variability in the PCB. These features are used to train a Random-Forest classifier for binary CCM detection.*

Our approach achieved an official Challenge score of ***0.119 (rank 36)*** *on the hidden test set. Five-fold cross-validation on the public training data confirmed stable entropy–complexity patterns across folds, demonstrating low sensitivity to alignment errors and noise. The proposed pipeline emphasizes physiological interpretability, reproducibility, and low computational cost, facilitating its integration into diagnostic workflows in resource-limited environments.*

1. Introduction

We participated in the 2025 George B. Moody PhysioNet Challenge, which invited teams to develop open-source algorithms for detecting Chagas disease from ECGs [1, 2]. The Challenge leveraged several large, annotated ECG databases, including CODE-15, SaMi-Trop, PTB-XL, REDS-II, and ELSA-Brasil [3–7].

Our team, **CompleXformers**, proposed an interpretable ECG-based approach that prioritizes transparency and physiological meaning over model complexity. Instead of relying on deep learning, we characterize each subject by two wavelet-derived biomarkers—entropy (H) and complexity (C)—computed from the relative wavelet energy of a principal-component beat (PCB). These measures quantify morphological disorder and temporal organization, offering a compact representation of cardiac heterogeneity associated with Chagasic cardiomyopathy (CCM).

This framework explores whether intrinsic wavelet entropy–complexity signatures can distinguish CCM-related electrical remodeling, aiming to provide clinically interpretable biomarkers for low-resource diagnostic settings.

2. Methods

2.1. Challenge data and labeling

We used exclusively the official training datasets released for the 2025 George B. Moody PhysioNet Challenge [1, 2], namely CODE-15%, SaMi-Trop, and PTB-XL (12-lead ECGs with harmonized metadata). Following the organizers' labels, we framed a binary task: *CCM* (confirmed Chagasic cardiomyopathy) vs. *Non-CCM* (healthy/other cardiopathies). No external data were introduced, and we did not relabel the training data. Records were retained unless failing strict quality checks (below).

2.2. Signal conditioning and quality controls

All ECG leads were resampled to $f_s = 1000\,\mathrm{Hz}$ to enable a homogeneous time–scale analysis around the QRS complexes. A zero-phase band-pass filtered the signals in the range of 0.5–$40\,\mathrm{Hz}$ to mitigate baseline wander and high-frequency noise; an adaptive notch at $50/60\,\mathrm{Hz}$ was applied if narrowband interference was detected. Baseline was estimated on decimated knots and removed by cubic-spline interpolation [8]. Brief artifacts (NaNs, saturated samples, or flatlines) were corrected via linear interpolation, whereas extended artifacts prompted exclusion of the affected beats while preserving the remainder of the

 ISSN: 2325-887X DOI: 10.22489/CinC.2025.061

record. All code paths are deterministic with fixed seeds and double-precision arithmetic. The full preprocessing configuration is summarized in Table 1.

Step	Setting (units)
Resampling	$f_s = 1000$ Hz
Band-pass	0.5–40 Hz (zero-phase IIR)
Notch	50/60 Hz (adaptive)
Baseline	Cubic-spline removal (decimated knots)
Sanity checks	NaNs / saturations / flatlines (beat-level exclusion)
Precision	float64; fixed random seeds

Table 1. Preprocessing settings. Parameters applied per lead and per record before segmentation.

2.3. Beat detection, windowing, and alignment

R-peaks were detected per lead using adaptive thresholds on the cleaned trace (NeuroKit2-style detector). Around each R, we extracted a fixed window of duration $T_w = 0.256\,\mathrm{s}$ with 35% pre-R and 65% post-R to cover QRS complexes and early ST. Initial alignment used cross-correlation against a running-median template; residual jitter was corrected with Woody's iterative method. Beats with normalized correlation < 0.85 to the template were discarded to avoid smearing the dominant morphology. If a lead lacked enough valid beats, we fell back to fixed, non-aligned central windows to preserve pipeline continuity.

Let L be the number of valid leads and T the window length (samples). We build a matrix $X \in R^{L\times T}$ by stacking aligned beat waveforms aggregated per lead with trimmed means.

2.4. Principal-component beat

After column-centering X, we compute the SVD

$$X = U\,\Sigma\,V^{\top}, \qquad U \in R^{L\times L},\ V \in R^{T\times T}, \quad (1)$$

and define the *Principal-component beat (PCB)* as $s(t) \equiv v_1$, the first right singular vector (dominant temporal morphology). We monitor the explained-variance ratio $\sigma_1^2/\sum_i \sigma_i^2$ to ensure a clear first mode. If SVD is ill-conditioned or too few beats remain, we fall back to the central window of the lead with maximal variance (deterministic).

2.5. Continuous wavelet analysis and scale energy

We apply the Continuous Wavelet Transform to the PCB using a Daubechies-6 mother wavelet [8]. Sixteen analysis scales $\{a_j\}_{j=1}^{16}$ are selected to span pseudo-frequencies from approximately 31.3 Hz up to the Nyquist limit, emphasizing higher-frequency activity related to the steep QRS slopes. For a sampling interval $\Delta t = 1/f_s$, pseudo-frequencies map as $f_j \approx f_c/(a_j\,\Delta t)$, where f_c denotes the db6 center frequency.

The CWT-like coefficients, defined for each scale a_j and time shift k, are

$$c_{j,k} = \frac{1}{\sqrt{a_j}} \int s(t)\,\psi^*\Big(\frac{t-k}{a_j}\Big)\,dt,$$

where k indexes the translation of the wavelet along time. The per-scale energies $E_j = \sum_k |c_{j,k}|^2$ define a normalized distribution over scales,

$$\rho_j = \frac{E_j}{\sum_{m=1}^{J=16} E_m}, \qquad \sum_{j=1}^{J} \rho_j = 1,$$

which summarizes the relative energy contribution of each scale (units: normalized units, n.u.). The $1/\sqrt{a_j}$ factor ensures the standard ℓ_2-norm energy normalization across scales. This wavelet-energy formalism follows the early definitions of wavelet entropy proposed by Rosso *et al.* [9, 10].

2.6. Wavelet entropy and statistical complexity

Let $\rho = (\rho_1, \rho_2, \ldots, \rho_J)^{\top}$ denote the scale–energy probability vector, constrained to the J-dimensional probability simplex

$$\mathcal{P} = \left\{ \rho \in \mathbf{R}^{\mathbf{J}}_{\geq \mathbf{0}} \mid \sum_{\mathbf{j=1}}^{\mathbf{J}} \rho_{\mathbf{j}} = \mathbf{1} \right\}.$$

The normalized Shannon entropy is defined as

$$H(\rho) = -\frac{\sum_{j=1}^{J} \rho_j \ln(\rho_j)}{\ln(J)}.$$

The uniform reference distribution, representing the barycentric point of $\mathcal{P}$, is

$$\rho_e = \left(\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J}\right)^{\top} \in \mathcal{P}.$$

The Jensen–Shannon divergence between two elements $p, q \in \mathcal{P}$ is

$$\mathrm{JS(p,q)} = \mathrm{H}\big(\tfrac{\mathrm{p+q}}{2}\big) - \tfrac{1}{2}\mathrm{H(p)} - \tfrac{1}{2}\mathrm{H(q)}.$$

The *wavelet statistical complexity*, defined over $\mathcal{P}$, is

$$C(\rho) = Q_0\,H(\rho)\,\mathrm{JS}(\rho, \rho_{\mathrm{e}}),$$

where the closed-form normalizer Q_0 ensures $C \in [0, 1]$ for J states:

$$Q_0 = \frac{-2}{\left(\frac{J+1}{J}\right)\ln(J+1) - 2\ln(2J) + \ln J}.$$

This formalism follows the statistical complexity framework of Kowalski *et al.* [11] and its biomedical adaptation by Valverde *et al.* [12]. **Variable definitions.** ρ_j (n.u.) is the probability associated with the wavelet energy at scale a_j; $H(\rho)$ quantifies the dispersion of energy across scales; $C(\rho)$ increases when this dispersion coexists with structured deviations from ρ_e, reflecting organized heterogeneity in the signal's multiscale structure.

2.7. Classifier, hyperparameters, and operating characteristics

A Random-Forest (RF) classifier receives the two-dimensional feature vector (H, C). Given the low feature dimensionality and the goal of minimizing variance across folds, a compact RF configuration was adopted with 12 estimators, a maximum of 34 leaf nodes, and a fixed random seed of 56, without class reweighting. This setup provided stable performance under small preprocessing variations and avoided overfitting to dataset-specific artifacts.

2.8. Validation, robustness, and reproducibility

We conducted stratified 5-fold cross-validation on the public training data with subject-wise grouping to avoid patient leakage. Each fold executed the full pipeline (preprocessing $\rightarrow$ detection/alignment $\rightarrow$ PCB $\rightarrow$ CWT $\rightarrow$ (H, C) $\rightarrow$ RF) to assess stability and select qualitative defaults (e.g., correlation threshold 0.85, $J = 16$). To probe robustness, low-amplitude white noise (std $= 0.01$) and temporal jitter of ± 0.004 s were injected before alignment; correlations of H and C with baseline values confirmed stability under perturbations. All randomness used fixed seeds, and artifacts (model and configuration) were reloaded prior to inference. The implementation avoids GPU dependencies, achieving per-record run-times of a few seconds on a standard laptop. Deterministic fallbacks ensure feature extraction continuity in all cases, including insufficient alignment or missing data.

3. Results

We summarize our results using both quantitative Challenge scores (Table 2) and qualitative visualizations of the extracted biomarkers (Figs. 1–2). Figure 1 shows the scalogram of the principal-component beat (PCB) computed with a db6 wavelet and J=16 scales, where time–frequency localization highlights the dominant QRS content and defines the scale-energy probabilities ρ.

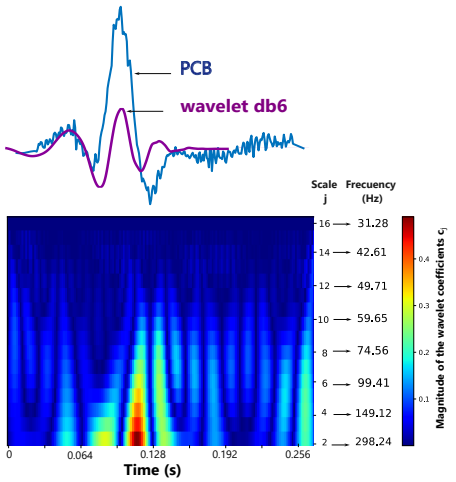

Figure 1. Scalogram of the principal-component beat using db6 with J=16 scales. Axes: Time (s) and Frequency (Hz); magnitude in arbitrary units. Scale energies define the probability distribution ρ for (H, C) computation.

Figure 2 displays the distribution of records in the entropy–complexity plane (H, C), derived from the PCB wavelet energy, with mean and standard-deviation markers summarizing inter-group variability.

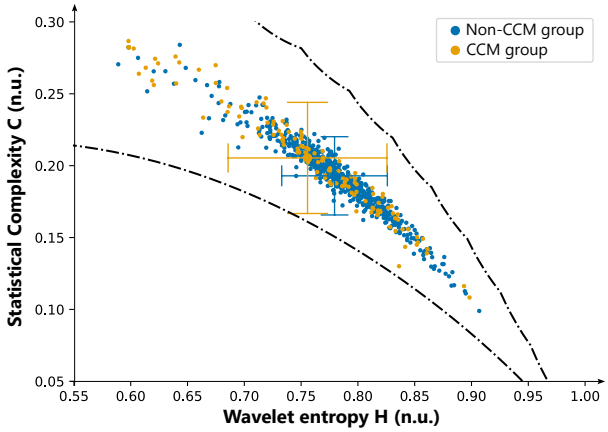

Figure 2. Entropy–complexity plane (H, C) (n.u.) computed from PCB wavelet energies. Each point corresponds to one record; summary markers indicate group means and standard deviations (self-contained legend).

Finally, Table 2 presents the official Challenge 2025 scores obtained by our team, summarizing validation and test performances across all datasets.

Set	Rank	Val	R-II	SaMi	ELSA	Test
CompleXformers	36	0.187	0.149	0.116	0.093	0.119

Table 2. Official Challenge 2025 scores for team *CompleXformers*. Val: REDS-II validation; R-II: REDS-II test; SaMi: SaMi-Trop 3; ELSA: ELSA-Brasil; Test: overall hidden test set.

4. Discussion and Conclusions

The proposed wavelet-based entropy–complexity pipeline produced interpretable biomarkers of multiscale cardiac dynamics in the Challenge 2025 datasets. Entropy (H) rises with the dispersion of wavelet energy across scales, whereas complexity (C) reflects organized departures from uniformity. In the non-Chagasic group—comprising both healthy and other cardiac conditions—higher H and lower C indicate a more heterogeneous but less organized energy distribution. By contrast, Chagasic cardiomyopathy shows more localized and structured energy concentration, consistent with fibrosis-related conduction delays, yielding lower H and higher C. Thus, (H, C) jointly capture the balance between dispersion and organization in ventricular activation.

Compared with conventional feature sets or deep-learning models, this approach is transparent, low-cost, and robust, relying on only two descriptors from the principal-component beat. Although compact, it favors interpretability, reproducibility, and physiological insight—key for translational and clinical use.

The method's main strengths are its simplicity, deterministic fallbacks, and resilience to moderate noise and timing jitter, ensuring stable performance across cross-validation folds. Future work will extend the analysis to the first three principal components (PC1–PC3) to incorporate additional spatial information while preserving interpretability.

The CompleXformers team achieved an official Challenge score of 0.119 (rank 36), with consistent stability across validation and test sets. Overall, the method offers a practical compromise between interpretability, physiological coherence, and computational efficiency, supporting its potential as a lightweight, explainable biomarker framework for large-scale ECG screening.

Acknowledgment

Competing interests: The authors declare no competing interests.
Ethical approval: Not required.
Funding: This research was supported by Universidad Tecnológica Nacional, Facultad Regional Buenos Aires (UTN-FRBA), under project BAICTC546 "Análisis electrocardiográfico mediante inteligencia artificial y complejidad wavelet", directed by Mariano Llamedo Soria, and by a CONICET doctoral fellowship.

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 Comput. Cardiol., 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] 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. Nat Commun 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. Sci Data 2020;7:154.

[6] 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. Circ 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] Sörnmo L, Laguna P. Bioelectrical Signal Processing in Cardiac and Neurological Applications. Burlington, MA: Elsevier Academic Press, 2005. ISBN 0-12-437552-9.

[9] Rosso OA, Blanco S, Yordanova J, Kolev V, Figliola A, Schürmann M, et al. Wavelet entropy: a new tool for analysis of short duration brain electrical signals. J Neurosci Methods 2001;105(1):65–75. ISSN 0165-0270.

[10] Rosso OA, Mairal M. Characterization of time dynamical evolution of electroencephalographic epileptic records. Physica A 2002;312(3–4):469–504. ISSN 0378-4371.

[11] Kowalski AM, Martin MT, Plastino A, Rosso OA, Casas M. Distances in probability space and the statistical complexity setup. Entropy 2011;13(6):1055–1075. ISSN 1099-4300.

[12] Valverde ER, Clemente GV, Arini PD, Vampa V. Wavelet-based entropy and complexity to identify cardiac electrical instability in patients post myocardial infarction. Biomed Signal Process Control 2021;69:102858. ISSN 1746-8094.

Address for correspondence:

Gisela Vanesa Clemente
[1]Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de La Plata (UNLP), La Plata, Argentina
[2]Universidad Tecnológica Nacional, Buenos Aires, Argentina
gclemente@mate.unlp.edu.ar