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

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

[1] Departamento de Ingeniería, Universidad Tecnológica Nacional, Facultad Regional Buenos Aires, Argentina
[2] CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), Argentina
[3] CMaLP (Centro de Matemática de La Plata), La Plata, Argentina

Abstract

A compact, interpretable feature-extraction pipeline for ECG-based detection of Chagasic cardiomyopathy (CCM) is presented. A principal component beat is constructed from multi-lead ECG through robust alignment and SVD, and it is analyzed with a continuous wavelet transform (CWT) based on a Daubechies-6 wavelet treated as continuous. From the resulting scale-wise energy distribution, two biomarkers are derived: normalized Shannon entropy $\mathcal{H}$ and wavelet statistical complexity $\mathcal{C}$. A Random Forest classifier is trained on $(\mathcal{H}, \mathcal{C})$. The implementation, developed for the **George B. Moody PhysioNet Challenge**, *is equipped with robust fallbacks to guarantee feature extraction for every record. Significant differences between CCM and non-CCM groups are observed, and competitive leaderboard performance is achieved, highlighting the potential of interpretable biomarkers.*

1. Introduction

Chagas disease, caused by *Trypanosoma cruzi*, affects millions in Latin America and continues to be a major cause of sudden cardiac death. The early detection of CCM is regarded as crucial. The use of black-box deep models has been associated with high computational cost and large data requirements, which limit their clinical translation.

As an alternative, a biophysically meaningful and low-dimensional representation is proposed, based on wavelet-derived entropy and complexity. Two biomarkers are defined: the entropy $\mathcal{H}$, obtained from normalized Shannon entropy, and the complexity $\mathcal{C}$, obtained from wavelet statistical complexity. From a biological perspective, $\mathcal{H}$ quantifies the degree of disorder in the distribution of wavelet energy across scales, which can be related to the heterogeneity of cardiac electrical activity. Conversely, $\mathcal{C}$ measures the balance between order and disorder, thus capturing structured deviations from uniformity that reflect the presence of organized but altered conduction patterns in the myocardium. Both biomarkers are extracted from a principal component beat, which is robustly derived from multi-lead ECG.

2. Methods

2.1. Datasets and preprocessing

The George B. Moody PhysioNet Challenge datasets of 12-lead ECGs was used. Records labeled as CCM correspond to patients with a reported diagnosis of Chagasic cardiomyopathy, while those labeled as Non-CCM correspond to individuals without such a diagnosis, including both healthy subjects and patients with other cardiovascular conditions.

The Challenge data are derived from three main sources:

- **CODE-15% dataset:** a large Brazilian cohort from which approximately 15% of the CODE study records were released, including demographic information and specific Chagas labels.
- **SaMi-Trop dataset:** a cohort study of patients with confirmed Chagas disease from endemic regions in Brazil, with ECGs and associated demographic data.
- **PTB-XL dataset:** a large German dataset of clinical 12-lead ECGs, which includes a wide range of cardiac and non-cardiac pathologies.

Signals from these datasets were converted to WFDB format by the Challenge organizers, including available demographics and Chagas labels. All signals were subsequently resampled to 1000 Hz. In particular, filters from `NeuroKit2`, a Python toolbox for neurophysiological signal processing, were applied to each lead, by which bandpass filtering, baseline wander suppression, and adaptive notch filtering at the specified powerline frequency were performed, resulting in artifact-reduced ECG traces suitable for delineation. Baseline wander was further cor-

rected through cubic spline interpolation.

2.2. QRS segmentation and alignment

R-peaks were first detected in each beat of each lead of the ECG from each patient. For each lead, 256 ms windows (35 % pre-R, 65 % post-R) were then extracted. In cases where R-peaks were missing, overlapping or tiled windows were employed as a fallback. Beat alignment was performed by cross-correlation and by the Woody algorithm against a median reference. The Woody algorithm, widely used in biomedical signal processing, iteratively aligns signals by estimating relative delays and averaging them, thus improving robustness in the presence of noise or jitter. Beats with normalized correlation < 0.85 were discarded. When alignment failed, the median or a deterministic central segment was substituted.

2.3. Principal Component Beat

Valid per-lead beats were stacked, centered, and reduced via singular value decomposition (SVD). The first right singular vector defined the temporal pattern, anchored using the highest-energy lead. This produced a robust principal component beat.

2.4. Wavelet energy distribution

For each principal component $s(t)$, with t denoting time, the CWT was applied using the Daubechies wavelet of order 6 ($\psi_{j,k} = \mathrm{db6}$) as the mother wavelet. The wavelet coefficients $c_{j,k}(t)$ were obtained as the inner product between the signal and the scaled and shifted wavelet:

$$c_{j,k}(t) = \frac{1}{\sqrt{j}} \int_{-\infty}^{\infty} s(t)\, \psi^*\left(\frac{t-k}{j}\right) dt, \qquad j = 1, \ldots, 16.$$

The relationship between the wavelet scale j and the equivalent frequency f_j in Hz for the continuous wavelet transform is given by

$$f_j = \frac{f_c}{j\,\Delta t},$$

where f_c is the center frequency of the mother wavelet selected and Δt is the sampling interval in seconds.

In particular, when the `db6` wavelet is used with scales $j = 1, \ldots, J = 16$, the resulting frequency coverage extends approximately from 31.28 Hz up to the Nyquist frequency. The following figure illustrates the scalogram obtained from the Continuous Wavelet Transform (CWT) of $s(t)$:

Then, the energy at each scale was computed as

$$E_j = \sum_{k=1}^{K} |c_{j,k}|^2, \qquad j = 1, \ldots, J.$$

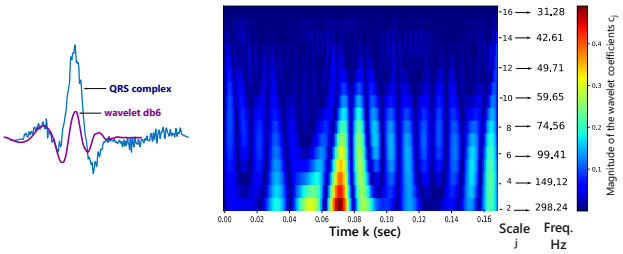

Figure 1. Scalogram of the signal $s(t)$ obtained using the CWT with the db6 mother wavelet and scales $j = 1, \ldots, 16$.

From these values, a normalized probability distribution across scales was defined:

$$\rho_j = \frac{E_j}{\sum_{j=1}^{J} E_j}, \qquad j = 1, \ldots, J.$$

By construction, the vector $p = (p_1, \ldots, p_J)$ belongs to the probability simplex

$$P = \left\{ \rho \in R^J \mid \rho_j \geq 0, \sum_{j=1}^{J} \rho_j = 1 \right\}.$$

2.5. Shannon Entropy and Statistical Wavelet Complexity biomarkers

The normalized Shannon entropy is defined as

$$\boxed{H[P] \;=\; \frac{-\sum_{j=1}^{J} \rho_j \,\ln \rho_j}{\ln J}}$$

The normalized Statistical Wavelet Complexity is defined as

$$\boxed{C[P] \;=\; Q_0\, H[P]\, \mathrm{JS}(\mathrm{p}, \mathrm{p_e})}$$

where $\mathrm{JS}(\mathrm{p}, \mathrm{p_e})$ denotes the Jensen Shannon divergence between the distribution p and the uniform distribution p_e; and Q_0 is a closed-form normalization factor ensuring that $C \in [0, 1]$, given by

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

2.6. Training and inference pipeline

The routine was designed as a compact, fail-safe pipeline centered on two interpretable features: normalized Shannon entropy $\mathcal{H}$ and statistical wavelet complexity $\mathcal{C}$.

For training, the Challenge records were discovered from the provided data directory and were processed independently. Multilead ECG signals were loaded together with metadata (in particular, the sampling frequency). Features were then extracted, yielding the

two-dimensional vector $(H[P], C[P])$ computed from a principal-component beat derived from the multilead input. Ground-truth labels were obtained via the Challenge helpers and were stored as boolean targets. After extraction, the design matrix had shape $N \times 2$, and training was conducted only if both classes were represented so as to avoid degenerate classifiers.

A Random Forest classifier was then fitted to the features. A small number of trees and a bounded number of leaf nodes were selected ($n_{\text{estimators}} = 12$, max_leaf_nodes $= 34$, random_state $= 56$) to promote fast and stable training with controlled variance. A reload was performed immediately to validate that the artifact could be reopened in the same environment.

At inference, the saved dictionary was loaded and the estimator was retrieved. A binary decision was returned via *predict* and a positive-class probability via *predict_proba*. If no model was available, a conservative default $(0,\ 0.0)$ was produced to prevent downstream failures. If features could not be computed for a given record, *None* values were returned to signal the failure explicitly.

A deterministic rescue path was implemented to ensure per-record feature availability, activated whenever the primary extractor failed or yielded non-finite outputs. The lead with the largest variance was selected, and a central window of approximately 0.256 s at 1000 Hz was obtained. A continuous-style wavelet analysis was then performed using a Daubechies-6 wavelet with $J = 16$ scales, and coefficients were normalized with a $1/\sqrt{\text{scale}}$ gain. Per-scale energies were summed over time, and a probability mass function over scales was formed. The entropy $H[P]$ was computed as the Shannon entropy of this distribution, normalized by $\ln J$. The complexity $C[P]$ was obtained by modulating $H[P]$ with the Jensen–Shannon divergence between the empirical distribution and the uniform distribution over scales.

Overall, the system was conceived to be robust by design: signals were sanitized at load time, features were strictly validated, a principled fallback reproduced the intended wavelet-based measurements from raw data, and model input/output was kept minimal and reproducible. This combination ensured that every record could be processed end-to-end while preserving the interpretability provided by $(H[P], C[P])$.

3. Results

Compared with the CCM group, the Non-CCM cohort exhibited, on average, higher entropy (H) and lower complexity (C). This combination is consistent with a broader dispersion of wavelet energy across scales and with weaker, less organized departures from uniformity, i.e., reduced structural organization in QRS morphology. In practical terms, larger H reflects greater variability in the scale-wise energy distribution, whereas smaller C indicates diminished patterned structure that would otherwise arise from organized conduction abnormalities.

On the public validation leaderboard for the official phase of the 2025 George B. Moody PhysioNet/Computing in Cardiology Challenge, our team **CompleXformers** achieved a score of 0.187.

Fig. 2 shows the feature distribution across groups:

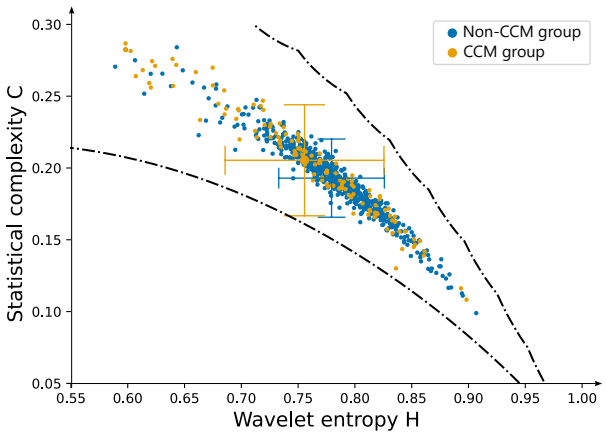

Figure 2. Distribution of entropy H and complexity C between CCM and Non-CCM groups.

4. Discussion

The proposed biomarkers are interpretable: H reflects the spread of wavelet energy, C captures structured deviations from uniformity. This method is computationally light, robust to missing data, and does not rely on deep networks.

4.1. Limitations and Future Work

The current approach captures only QRS-based scale distributions. Possible extensions include T-wave descriptors, rhythm statistics, and learned wavelets. In addition, analysis of QRS complexes in low-frequency bands may be incorporated, capturing wavelet scales complementary to those already considered, with the aim of revealing slower conduction components and morphological modulations that are not visible at the previously analyzed scales. Integrating multi-beat or temporal features may further improve detection.

5. Conclusions

We implemented and validated a wavelet-based entropy/complexity pipeline for Chagas detection. Despite

using only two features, (H, C) achieved group separability and a competitive Challenge score. This demonstrates the promise of interpretable biomarkers for deployment in low-resource environments.

References

[1] L. Sörnmo and P. Laguna, *Bioelectrical Signal Processing in Cardiac and Neurological Applications*, Academic Press, 2005.

[2] O. A. Rosso et al., "Wavelet entropy: a new tool for analysis of short duration brain electrical signals," J. Neurosci. Methods, 2001.

[3] A. M. Kowalski et al., "Distances in probability space and the statistical complexity setup," Entropy, 2011.

[4] E. R. Valverde et al., "Wavelet-based entropy and complexity to identify cardiac electrical instability post-MI," Biomed. Signal Process. Control, 2021.

[5] O. A. Rosso and M. Mairal, "Characterization of EEG epileptic records," Physica A, 2002.

[6] L. Breiman, "Random Forests," Machine Learning, 2001.

Address for correspondence:

G V Clemente
Departamento de Ingeniería, Universidad Tecnológica Nacional, Facultad Regional Buenos Aires, Argentina
gclemente@mate.unlp.edu.ar