Using Dynamic Time Warping and Agglomerative Clustering of ECG Data to Group Distinct PQRST Morphologies in Patients with Chagas Disease

Carlos M D F Belém[1], João P V Madeiro[1], Roberto C Pedrosa[2]

[1]Universidade Federal do Ceará (UFC), Fortaleza/CE, Brazil
[1]Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro/RJ, Brazil

Abstract

Chagas disease is a significant public health issue in Latin America, affecting millions at risk of severe cardiac abnormalities. We aimed to capture distinct clusters of PQRST morphologies in individuals diagnosed with Chagas cardiomyopathy using unsupervised learning techniques. From 207 Holter recordings, we derived representative average heartbeats through ECG segmentation and K-Means clustering as pre-processing steps. We then computed a Dynamic Time Warping (DTW) distance matrix to perform agglomerative hierarchical clustering, assessing various cluster numbers (K) against silhouette scores for three linkage methods. Using complete linkage and K = 19 yielded a silhouette score of 0.182. Visual inspection confirmed separation of PQRST morphologies, capturing relevant patterns—including prevalent T-wave inversions, acute T-waves, and net positive or negative QRS complexes. We computed clinical metrics within each cluster to explore potential links between ECG morphologies and patient health. Averaging heartbeats made the process feasible due to the quadratic complexity of DTW. This study highlights the potential of this clustering approach in stratifying ECG data for high-level analysis, segregating ECGs with notable noise levels, and providing a tool for rapid morphological categorization.

1. Introduction

Chagas disease (CD), caused by the protozoan *Trypanosoma cruzi*, remains a major public health concern in Latin America, where it is endemic in 21 countries and affects an estimated 6 to 7 million people globally [1]. About 30–40% of asymptomatic individuals eventually develop significant clinical manifestations after 10 to 30 years, with cardiac involvement being the most severe—leading to myocarditis, fibrosis, heart failure, and sudden cardiac death [2][3].

Grouping similar unlabeled time series in large biomedical datasets like Holter ECG recordings can aid in diagnosis. This study developed an automated ECG analysis pipeline for Chagas patients, enabling data sampling, segmentation, and clustering. We also averaged the Rassi score for each cluster to explore potential associations between ECG morphology and patient risk [4].

2. Methods

2.1. Clinical data

The study utilized clinical and laboratory data from 310 patients diagnosed with Chagas Disease at the University Hospital Clementino Fraga Filho (HUCFF), Federal University of Rio de Janeiro, collected over 26 years (1990–2016). The dataset acquisition was approved by the HUCFF-UFRJ ethics committee.

For each patient, the database includes one or more 24-hour ECG recordings (DAT files) from lead D-II, sampled at 128 Hz using a Holter monitor. Additionally, it contains comprehensive clinical data such as echocardiogram results, medication usage, and outcomes like Sudden Cardiac Death—a severe consequence of Chagas cardiomyopathy.

2.2. Sampling the database and processing the data

We selected the most recent Holter recording from each patient and discarded those with excessive noise, retaining 207 DAT files (115 female, 92 male). Non-cardiac signals at the beginning and end of each recording were removed. From each Holter, we randomly extracted a 30-minute window from each third of the recording, automatically retesting and re-randomizing noisy sections if necessary. For 15 recordings that were too short, we used the entire signal.

Accurate heartbeat segmentation is crucial for clustering. We used NeuroKit's default method [5] for R-peak detection and segmentation, which involves filtering to remove low-frequency and power line noise,

computing the signal's derivative, smoothing to highlight significant changes, and applying thresholds to detect QRS complexes based on relative prominence and minimum intervals between R-peaks. This method was validated using an online ECG database [6]. For recordings where NeuroKit failed, Hamilton's algorithm [7] was applied to 13 recordings and Rodrigues' algorithm [8] to 5 recordings. Eleven recordings required custom parameters, while 183 were segmented using the standard pipeline.

After segmentation, we computed a representative heartbeat for each recording by averaging R-peak-aligned heartbeats of the same length. To handle morphological variations like ventricular extrasystoles, we grouped beats by morphology using z-score normalization and K-Means clustering [9] with two clusters, employing Euclidean distance as the metric. The heartbeats from the largest cluster were averaged to create the representative heartbeat. Figure 1 illustrates examples where abnormal beats were successfully segregated before averaging.

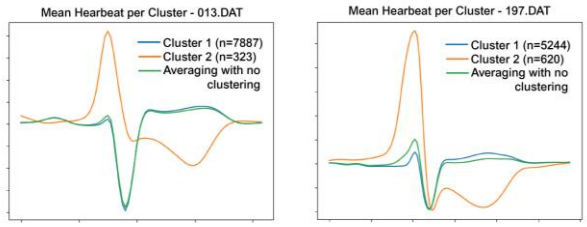

Figure 1. Average heartbeats from the two clusters and the overall set, illustrating how clustering distinguishes morphologies before final averaging.

2.3. Hierarchical Clustering and DTW

We applied agglomerative hierarchical clustering [10], which generates a hierarchy of clusters without requiring a predefined number of clusters. This method uses a pairwise distance matrix and produces a dendrogram for visualization. To measure dissimilarity, we used Dynamic Time Warping (DTW) [11], which aligns time series based on sequence and timing, making it suitable for varying-length heartbeats (warping examples are shown on Figure 2). For a reduction in computational load, we used averaged heartbeats instead of the entire signal.

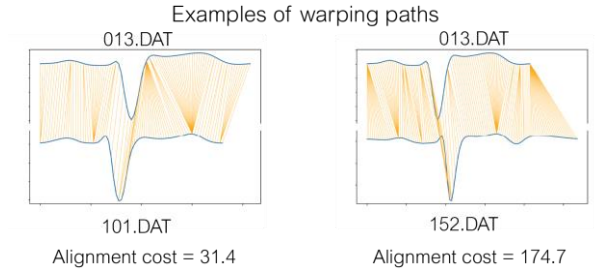

Figure 2. Dataset examples comparing heartbeat pairs, showing alignment costs computed.

2.4. Linkage methods and silhouette score

We evaluated single, complete, and average linkage methods to determine how distances between clusters are calculated. To choose the best method and number of clusters (K), we computed the silhouette score [12] for each method across various K values, as seen in Figure 3. The silhouette score, ranging from -1 to 1, assesses how well a sample fits within its cluster compared to others. Although average linkage produced higher scores, it led to more singleton clusters and greater clinical value dispersion. We ultimately selected 19 clusters using the complete linkage method, which provided a good balance between cluster cohesion and silhouette score quality.

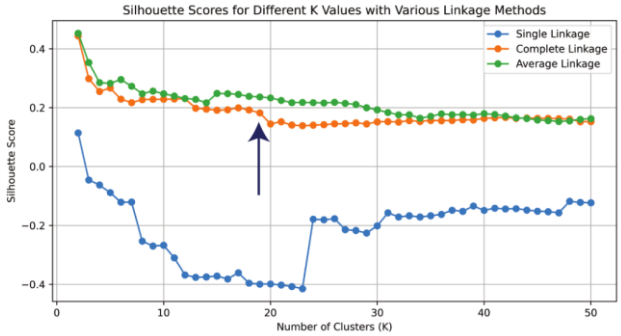

Figure 3. Silhouette score for each value of K and each linkage method.

2.5. Cluster evaluation

We evaluated the coherence of cluster morphologies by consulting cardiology experts. To explore associations between grouped PQRST morphologies and clinical presentations in Chagas disease, we analyzed two key clinical metrics within each cluster: the average Rassi score and the prevalence of Sudden Cardiac Death (SCD), a critical outcome in chagasic cardiomyopathy. The Rassi score predicts mortality risk in Chagas disease patients.

3. Results

3.1. Dendrogram

We generated a dendrogram to visualize the hierarchical relationships among clustered heartbeats, aiding in understanding data structure and identifying meaningful morphological subgroupings.

Figure 4 illustrates a dendrogram segment, highlighting waveform patterns and distinctions between clusters, demonstrating the effectiveness of hierarchical clustering in capturing complex heartbeat features.

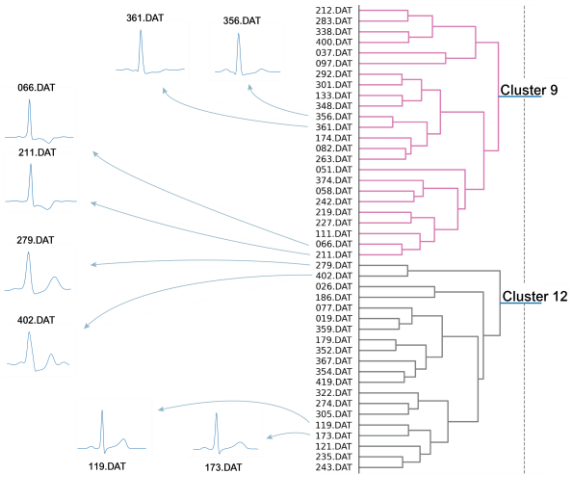

Figure 4. Section of the dendrogram with k=19 clusters, showing sample heartbeats. Reducing the agglomeration threshold would separate 66.DAT and 256.DAT into different groups.

3.2. Resulting Clusters

We analyzed several of the 19 clusters generated. Cluster 0, the largest with 48 members, mostly showed enlarged QRS complexes with absent or small R-waves (Figure 5). Of these, 21 patients experienced SCD, and the cluster had a mean Rassi score of 11.23 (median 13), indicating high risk.

In cluster 7, all 8 members had net positive QRS complexes and prominent T-waves. The median Rassi score was 4.0 with a small IQR of 3.75, reflecting lower risk. Two patients (060.DAT and 363.DAT) experienced SCD approximately 4 and 3.5 years after their Holter recordings, respectively. Cluster 8 had a median Rassi score of 3.0 (IQR: 3.0) with 8 members, all showing positive P and T waves and net negative QRS complexes. Two patients (014.DAT and 383.DAT) experienced SCD about 4 and 5 years after their recordings, with Rassi scores of 8 and 11, respectively.

Cluster 16 had 7 members with net positive QRS complexes and negative T waves. The median Rassi score was 11 (IQR: 6). Five of the seven patients in this cluster experienced SCD, highlighting its high-risk nature. Figure 6 shows the representative beats from clusters 7, 8, and 16. Table 1 summarizes key metrics for each of the 19 clusters, including Rassi score statistics and SCD occurrences. Figure 7 presents a box plot of Rassi score distributions across clusters, highlighting variability in risk levels and allowing comparison of score dispersion.

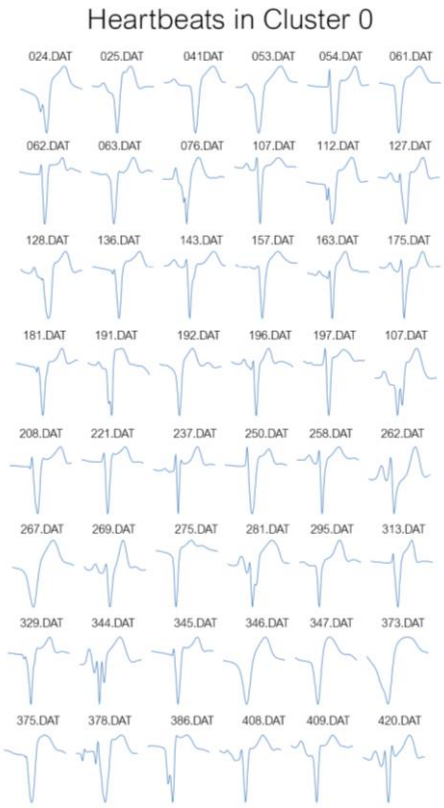

Figure 5. All representative heartbeats from cluster 0, indicating the original Holter reading file it came from.

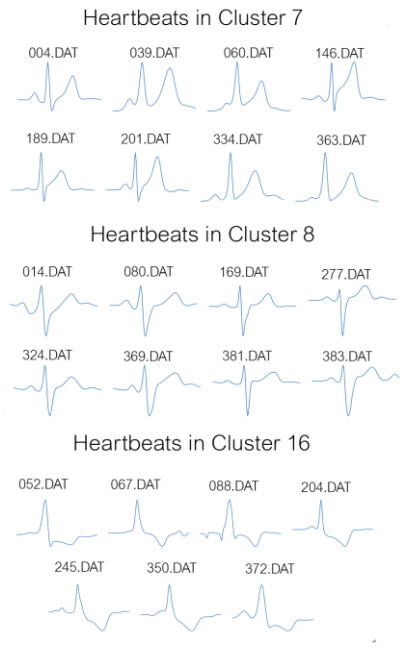

Figure 6. Representative heartbeats from clusters 7, 8, and 16, highlighting distinct morphological features.

Cluster	N	Mean ± SD	Median	IQR	Total SCD
10	1	18.0 ± 0	18.0	0.00	0
14	4	2.7 ± 2.0	3.0	1.25	0
11	4	5.0 ± 2.4	5.0	1.50	2
8	8	4.3 ± 3.5	3.0	3.00	2
7	8	5.0 ± 3.8	4.0	3.75	2
13	3	10 ± 4.6	13.0	4.0	2
2	9	4.8 ± 3.7	5.0	5.00	4
3	11	10 ± 5.5	11.0	5.00	5
5	2	11 ± 7.0	11.0	5.00	0
16	7	12 ± 5.6	11.0	6.00	5
15	4	3.5 ± 4.1	3.0	6.50	1
17	3	4.3 ± 7.5	0.0	6.50	1
6	7	6.4 ± 6.3	3.0	7.00	1
18	6	5.6 ± 4.2	4.0	7.25	2
9	24	7.1 ± 5.4	4.0	8.00	8
1	16	7.1 ± 4.4	7.0	8.00	7
12	20	5.2 ± 4.8	2.0	9.00	8
4	22	8.4 ± 6.4	9.5	10.7	7
0	48	11.2±5.7	13.0	11.0	21

Table 1. Rassi cluster metrics with K=19, including SCD prevalence, ordered by Rassi IQR (ascending order).

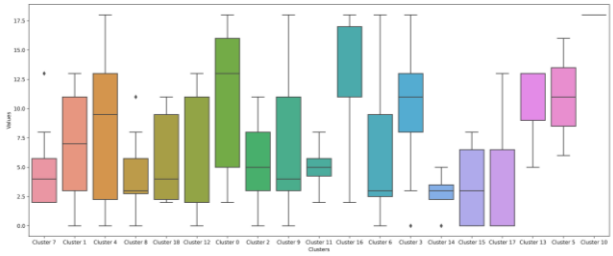

Figure 7. Box plots showing the distribution of Rassi scores across all clusters.

4. Discussion and Conclusions

This study demonstrates that hierarchical clustering effectively categorizes heartbeats based on morphological features, identifying clusters with varying risk profiles for sudden cardiac death (SCD). The correlation between ECG patterns, Rassi scores, and SCD occurrences underscores clinical relevance, although causation cannot be established without further studies. Relying solely on morphological features may overlook important factors like patient demographics and comorbidities.

Limitations include the potential for improved representative heartbeat accuracy using methods like DTW Barycenter Averaging and the consideration of different R-R intervals to capture more nuanced cardiac dynamics. Despite these, hierarchical clustering offers advantages in handling large ECG datasets and identifying patterns efficiently. The method is applicable to other cardiac conditions beyond Chagas disease, aiding in data labeling, annotation, and the development of automated ECG interpretation models.

References

[1] World Health Organization, "Chagas disease (American trypanosomiasis)," Fact Sheet No. 340, WHO, 2010.

[2] K. C. F. Lidani *et al.*, "Chagas disease: From discovery to a worldwide health problem," *Frontiers in Public Health*, vol. 7, p. 166, Jul. 2019, doi: 10.3389/fpubh.2019.00166.

[3] L. Z. Rojas *et al.*, "Electrocardiographic abnormalities in Chagas disease in the general population: A systematic review and meta-analysis," *PLoS Neglected Tropical Diseases*, vol. 12, no. 6, p. e0006567, Jun. 2018, doi: 10.1371/journal.pntd.0006567.

[4] A. Rassi *et al.*, "Development and validation of a risk score for predicting death in Chagas' heart disease," *The New England Journal of Medicine*, vol. 355, no. 8, pp. 799–808, Aug. 2006, doi: 10.1056/NEJMoa053241.

[5] D. Makowski *et al.*, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, Aug. 2021, doi: 10.3758/s13428-020-01516-y.

[6] J. Brammer, "Biopeaks: A graphical user interface for feature extraction from heart- and breathing biosignals," *Journal of Open Source Software*, vol. 5, no. 52, p. 2621, Aug. 2020, doi: 10.21105/joss.02621.

[7] P. Hamilton, "Open source ECG analysis," in *Computers in Cardiology*, Memphis, TN, USA, 2002, pp. 101–104, doi: 10.1109/CIC.2002.1166717.

[8] T. Rodrigues, S. Samoutphonh, H. P. da Silva, and A. L. N. Fred, "A low-complexity R-peak detection algorithm with adaptive thresholding for wearable devices," in *Proceedings of the 2021 International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 3021–3028, doi: 10.1109/ICPR48806.2021.9413245.

[9] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–136, Mar. 1982.

[10] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[11] P. Senin, "Dynamic time warping algorithm review," 2008. [Online].Available:http://seninp.github.io/assets/pubs/senin_dtw_litreview_2008.pdf

[12] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.

Address for correspondence:

Carlos Magno Dantas de Figueirêdo Belém
Rua Érico Mota, 66, apt. 205
Fortaleza-CE, Brazil