www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(2), April 2025, 147-157. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i2.528

Detection of cardiac arrhythmias in a 12-lead ECG dataset of more than 10,000 patients: a preliminary study using clustering algorithms

Jessica Alvariño-Durán, José Hernández-Torruco, Oscar Chávez-Bosquez, Betania Hernández-Ocaña
Universidad Juárez Autónoma de Tabasco

Abstract. The groupings of cardiac arrhythmias allow the identification of common patterns, distinctive characteristics, and similarities between different cases. In datasets where less common types of arrhythmias are identified; these grouping tools can better classify each subtype. This research was carried out on electrocardiogram records from a data set with more than 10,000 patients, previously labeled by cardiology specialists into 11 heart rhythms and grouped according to medical guidelines into four groups. A preliminary analysis of an ongoing project for detecting cardiac arrhythmias using unsupervised learning tools: clustering is presented. Feature selection was performed using filter tools, and the RR interval was extracted from the ECG records to be incorporated into the dataset under analysis as a new attribute. Internal validation metrics are used to check the quality of the selected clustering methods.
Keywords: arrhythmia; electrocardiogram; clustering.

Article Info
Received Jan 26, 2025
Accepted Mar 11, 2025

1 Introduction

Cardiac arrhythmias are abnormalities or alterations in the activation or normal beating of the cardiac myocardium. The causes of these anomalies can be diverse, as can their severity and clinical consequences (Fu, 2015). According to reports obtained from the Global Burden of Disease (GBD) Study, cardiovascular diseases (CVD) are collectively considered the leading cause of death in the world and contribute substantially to health loss and excessive health system costs (Vaduganathan et al., 2022).

Electrocardiography (ECG) is the main method for diagnosing heart diseases (Alipour et al., 2022). It is a non-invasive approach that graphically visualizes the heart's electrical activity. Analyzing the electrocardiographic signals makes it possible to diagnose a particular patient's health status (Lara Prado, 2016). In an electrocardiogram recording a normal heartbeat, the typical tracing consists of a P wave, a QRS complex, and a T wave. Both the duration and beat-to-beat variations of these waves are physiologically important; therefore, the time interval defined by the peaks and limits of the waves is clinically relevant (Alipour et al., 2022).

Visual analysis of ECG signals is an empirical science requiring much of knowledge and clinical training. According to the current practice of screening and diagnosing arrhythmias, cardiology specialists review ECG data, establish the diagnosis, and implement treatment plans (Zheng et al., 2020). However, the demand for highly accurate automated heart disease diagnoses has recently increased (Cárcamo Morales et al., 2020), in parallel with public health policy to implement broader screening procedures (Zheng et al., 2020). Artificial Intelligence has intervened to support medical specialists, reducing the time required to read data sets by mapping these readings.

To analyze variables and detect cardiac arrhythmias, a data set with 11 rhythms previously labeled by a group of experienced cardiologists was selected based on medical experience (Zheng et al., 2020). After this first label of rhythms and other cardiac conditions, the 11 main rhythms are merged into four groups, also defined by medical experience, in such a way that a balance is achieved between the less common rhythms and, therefore, with the least number of samples (patients) and higher level arrhythmia types (Zheng et al., 2020).

The starting point of this research was the need for a detailed quantitative exploration of the dataset collected by Chapman University and Shaoxing People's Hospital (Zheng et al., 2020). An unsupervised methodology was developed through clustering, considering an adaptation of the original dataset from extracting new features defined by mathematical formulations. To obtain

the possible groups of arrhythmia classes, defined in an automated way, and their possible comparison with the results suggested by cardiology specialists. Clustering performance is evaluated in terms of internal metrics.

2 Dataset Description

The data set used comprises 12-lead ECG signals of 10 seconds duration for 10,646 recordings, with a frequency of 500 Hz, collected by Chapman University and Shaoxing People's Hospital (Zheng et al., 2020). 11 types of common rhythms are presented, of which 83% correspond to arrhythmic recordings and 17% correspond to normal sinus rhythm. In addition, 56 additional conditions were identified. Data are labeled considering the following methodology:

1. Each subject underwent a resting-state 12-lead ECG test performed over 10 seconds.
2. A licensed medical specialist labeled the rhythm and other heart conditions.
3. A second medical specialist performed secondary validation.
4. Given the disagreement between the first two medical opinions, a high-level medical specialist intervenes and makes the final decision (Zheng et al., 2020).

Each of these heart rhythms is merged for subsequent analysis and classification, considering that they are not completely balanced, following a suggestion from cardiologist specialists. Several records with fewer instances of arrhythmia types and a more significant number of samples are hierarchically grouped. As a result, eleven rhythms were converted into four groups (SB, AFIB, GSVT, SR) as shown in Table 1. The division into four classes, as reported in the literature, was carried out because of the similar effects of the diseases in a heartbeat. The guidelines (January et al., 2014; Ben-Tal et al., 2012) recommend that AFIB and FA often coexist. Supraventricular tachycardia is a general term used in daily ECG examinations. In addition, because of the small number of instances of some rhythms, the practice of merging all tachycardias originating in supraventricular locations to the GSVT group was adopted. Merging sinus rhythm and sinus irregularity in the SR group allows this combination to be distinguished from that in the GSVT group. Sinus irregularities are mostly benign rhythms that in healthy conditions, especially in young people (MUSE, 2024). Sinus irregularity can be easily separated from sinus rhythm later by a single criterion: the variation of the RR interval (Zheng et al., 2020). Working details must be concise; well-known operations should not be described in detail.

Table 1. Information on the eleven common rhythms and four groupings defined in the dataset

Heart rate	Instances	Groups	No conditions	Groups
Sinus Bradycardia (SB)	3889	SB (3889)	2200	SB (2200)
Sinus Rhythm (SR)	399	SR (2225)	1366	SR (1665)
Sinus Irregularity (SA)	1826		299	
Atrial Fibrillation (AFIB)	1780	AFIB (2225)	413	AFIB (500)
Atrial Flutter (AF)	445		87	
Sinus Tachycardia (ST)	1568	GSVT (2307)	643	GSVT (1062)
Supraventricular Tachycardia (SVT)	587		379	
Atrial Tachycardia (AT)	121		24	
Atrioventricular Node Reentrant Tachycardia (AVNRT)	16		7	
Wandering Atrial Pacemaker (SAAWRT)	8		5	
Atrioventricular Reentrant Tachycardia (AVRT)	7		4	

The dataset is composed of 16 attributes and 10,648 instances. The variables were acquired using the GE MUSE system (Murat et al., 2021). These include diagnostic information for each subject, file name, heart rate, other additional conditions, patient age, gender, ventricular rate, atrial rate, QRS complex duration, QT interval, corrected QT interval, R axis, T axis, QRS counting, start of Q, end of Q, and end of T.

3 Related studies

Various researchers have shown interest in classifying cardiac arrhythmias based on this dataset, mainly because of its number of records. Each of these investigations aims to provide an efficient classification of common rhythm types previously labeled by

cardiology specialists. To achieve this, they conducted a methodology for grouping the rhythms due to the characteristics of the data set, following different guidelines. In most cases, they coincide in four and seven groups. Both groupings present high values of evaluation metrics; however, it is essential to highlight that with the four groups, better results are obtained in studies where both groupings are considered. New attributes are defined in each investigation, or new characteristics are extracted from the dataset.

In the state-of-the-art dataset, a methodology is applied through supervised algorithms, as shown in Table 2. Novel classification methods stand out because of their high-quality indices. Despite the prevalence of research (Mastoi et al., 2022; Rieg et al., 2020; Yoon & Kang, 2023) that adopts the methodology of grouping into four groups of arrhythmia classes defined in the pilot study (Zheng et al., 2020), several authors define even groupings (SB, SR, SI, AFIB, AF, ST, SVT) determined by eliminating the classes with the lowest number of instances (Meqdad et al., 2022; Faust et al., 2021). Due to the evidence of the high risk that atrial fibrillation implies and its involvement in the world population (January et al., 2014), in Yang et al. (2022), only three groups are defined (AFIB, AFL, NSR), which allows a better classification performance of this type. AF or non-AF is determined in (Faust & Acharya, 2021). On the other hand, in (Yoon & Kang, 2023; Dash & Liu, 2000), six groups are defined (SVT, ST, SB, AFIB, AFL, NSR). Overall, these studies provide valuable insight into the diversity of arrhythmic responses, highlighting the importance of adapting the clustering methodology according to the particularities of this dataset.

Table 2. Study of the state-of-the-art

STUDY	METHOD	EXTRACTED FEATURES	GROUPS	METRICS
J. (Zheng et al., 2020)	Extreme Gradient Boosting Tree	39,830 all the morphological origin of the signal. Includes the 11 reported in the dataset	4	F1=0,98(whitout conditions) F1=0,97(with conditions)
(Baygin et al., 2021)	Support Vector Machines (SVM)	16,384 features with homomorphically irreducible tree technique (HIT)	7	Accuracy=92,95%
			4	Accuracy=97,18%
(Yildirim et al., 2020)	DNN Model	DNN model	7	Accuracy=92,94%
			4	Accuracy=96,13%
(Faust & Acharya, 2021)	Residual Neural Network Deep Learning	RR Interval	2	Accuracy=98,55%, Sensitivity=99.40%, Specificity=94.30%
			6	Overall Accuracy=98,37%
(Murat et al., 2021)	Model DNN	Feature vectors are obtained from the DNN and fused with 11 features from the ECG	4	Accuracy=98%
(Yoon & Kang, 2023)	ResNet-50 and logistic regression achieved	396 features: first-order features, GLCM features (second-order features), and GLRLM features (higher-order features)	7	AUC=0.995 Accuracy=93.97% Sensitivity=0.940 Precision =0.937 F1score =0.936
(Mastoi et al., 2022)	fusion technique, Dual Event-Related Moving Average (DERMA) with Factional Fourier-Transform algorithm (FrlFT).	ECG wave detection	4	Mean Accuracy=78.25% Sensitivity=79% Positive Predictivity=80%
(Meqdad et al., 2022)	Deep Convolutional Neural Network (CNN) models	Wavelet Transform	7	Mean Accuracy = 97,60%
	Component-Aware Transformer (CAT),	Segmented an ECG through the 1D U-Net-based segmentation model	2	F1score = 0.85 AUC = 0,9823.
(Rieg et al., 2020)	White-box machine learning approach. C5.0	24 features based on statistical values and morphology signal	4	balanced accuracy of 95.35%.
			6	

(Andayeshgar et al., 2022)	Novel Graph Convolutional Network (GCN)	Benefitting from Mutual Information (MI) indices	7	Sensitivity=98.45% Precision=97.98% Specificity=99.85% Accuracy=99.71%
(Meqdad et al., 2022)	Convolutional Neural Network (CNN) models	Models are encoded as the GP algorithm's evolutionary trees to create the CNN models' interpretability feature.	7	Accuracy=98%
(Faust et al., 2021)	Deep Learning algorithm	Sequence of RR intervals	3	Accuracy=99.98% Sensitivity= 100.00% Specificity=99.94%
(Oliveira et al., 2022)	Voting Ensemble, used a combination of the XGradient Boost, Random Forest, and Gradient Boost models	11 characteristics defined in the dataset and characteristics determined by TW	4	F1-score= 0,93
(Lee et al., 2021)	A novel method to generate the Gray-Level Co-occurrence Matrix (GLCM) and Gray-Level Run-length Matrix (GLRLM) from one-dimensional signals combined with XGBoost.	396 features: first-order features, GLCM features (second-order features), and GLRLM features (higher-order features)	4	Accuracy=90.46% AUC=0.982 Sensitivity=0.892 Precision=0.900 F1score=0.895
(Ozpolat & Karabatak, 2023)	Quantum Support Vector Machine (QSVM) algorithm	11 features defined in the dataset	4	Accuracy =84.64%
(Dhananjay & Sivaraman, 2021)	Extremely Randomized Tree Classifier. CatBoost machine learning algorithm. CB classifier minimizes overfitting of the model.	9 features based on morphology signal	3	Accuracy=99% Sensitivity=99.17% Precision=99.25%
(Guo et al., 2023)	Support Vector Machines (SVM), K-vecinos más cercanos (KNN), perceptrón multicapa (MLP) y Naïve Bayes	Statistical parameters and autoregressive coefficients	4	Precision=92.02%

4 Data preprocessing

In (Zheng et al., 2020), the authors of the dataset mention that the ECG signal is filtered, free of artifacts and noise. The recordings were provided with initial preprocessing to smooth the ECG signals using the Butterworth filter and nonlocal averaging technique (Zheng et al., 2020). A search for missing data is conducted, which is essential to ensure the quality and reliability of the data used in the analysis and the construction of machine learning models. The result of the R base function “is.na” is that the dataset does not have missing data. Besides, a study of the presence of outliers by variables is carried out using boxplot graphics, a function of the “graphics” base library. Because of the graphs, the outliers are observed in the variables that make up the dataset.

5 Feature Selection

The authors of the dataset mention that the ECG signal is filtered and free of artifacts and noise (Zheng et al., 2020). The recordings were provided with initial preprocessing to smooth the ECG signals using the Butterworth filter and nonlocal averaging technique (Zheng et al., 2020). A search for missing data was conducted using the R base function “is.na”. It was determined that the dataset does not contain any missing data.

Feature Selection is based on filtering out irrelevant or redundant features from the dataset. By reducing dimensionality, improving model performance, and facilitating interpretation and analysis, feature selection significantly contributes to the efficiency of the modeling process (Dash & Liu, 2000). This study presents the analysis resulting from the methods defined in the FSelector package: CFS, Consistency, Chi-square, Information gain, and OneR (Romanski et al., 2023). The Consistency and CFS filters

list a subset of the most relevant selected features. In contrast, Information Gain, OneR, and Chi-squared assign scores to each feature, indicating the order of relevance (Romanski et al., 2023).

6 Feature Extraction

In cardiac arrhythmia analysis, it is essential to derive attributes that effectively capture the patterns and information in the original data. Several researchers, based on the randomness of the morphology of the intrapatient and interpatient ECG signals, have defined the relevance of the extraction of new features derived from those previously defined in the data set (Zheng et al., 2020). At this point, the authors agree that the RR interval, defined as the time between two consecutive R peaks in the ECG signal, can represent a promising exponent in terms of defining new patterns that provide relevant information to the dataset and then to the classification of the different types of arrhythmias (Faust et al., 2021).

According to specialized medical literature (The Texas Heart Institute, n.d.), abnormal variations in the RR interval may indicate cardiac arrhythmias. The authors present the entire state of the art using new variables not included in the dataset, which allows a better classification of the types of arrhythmias. In the dataset proposed for this study, 16 attributes were acquired from the GE MUSE system (MUSE, 2024). The GE MUSE system user manual defines the calculation of the corrected QT interval (QTc) following the Bazzet formula (Bazett, 1997) as in (1).

$$QT_c = \frac{QT\ interval}{\sqrt{RR}} \tag{1}$$

Considering this equation, the RR variable is cleared to obtain the RR interval from the data provided in the dataset. It is visually confirmed that the intervals obtained are between the parameters defined in the specialized literature (Medwave, n.d.), and the final value should not exceed 0.44 seconds; Above this level, we speak of a prolonged QT interval.

7 Clustering Method

By grouping arrhythmias, it is possible to identify common patterns, distinctive characteristics, and similarities between different cases. It allows a better understanding of the clinical characteristics, triggering factors, and possible underlying causes of arrhythmias. Identifying subgroups may have implications for prognosis and specific treatment for each subtype (The Texas Heart Institute, n.d.).

A modification of the grouping of K-medoids called PAM (Partitioning Around Medoids) is proposed as a first approach to obtaining groups of arrhythmia classes from clustering. This algorithm operates on the dataset's dissimilarity matrix. It minimizes the objective function by exchanging all non-medoid and medoid points iteratively until convergence (Aggarwal & Reddy, 2014). Medoid corresponds to the most central element of the cluster.

It is defined as an element within the cluster whose average distance (difference) between it and all other elements of the same cluster is the smallest possible (Joaquin, n.d.). The steps are based on selecting k initial random observations as medoids (initialization). Calculate the distance matrix between observations; each observation is assigned to the nearest medoid, forming k clusters. For each cluster, check if selecting another observation from the current cluster that is not a medoid reduces the average distance of the cluster (swap); the cost of the swap is calculated, and if it reduces the objective function (sum of dissimilarities), the swap is performed. The swap and reassignment are repeated iteratively until no further changes occur or convergence is reached (Joaquin, n.d.).

On the other hand, the performance of the CLARA (Clustering Large Applications) algorithm is evaluated. Its working principle is determined by considering a small sample of the data, with a fixed size (*sampsize*), instead of finding medoids for the entire data set. It then applies the PAM algorithm to generate an optimal set of medoids (Joaquin, n.d.). A benefit of k-medoid algorithms is handling outliers because the prototypes are medoids, which are dataset members. Therefore, no centroids are created artificially, allowing outliers to be addressed (Laurinec, 2024).

Once the clustering algorithms are defined, calculating the distance and selecting the number of clusters must be considered. Depending on the characteristics of the data set, the Gower distance allows the handling of mixed data types. A similarity coefficient is determined based on the different types of variable information. Once the distance matrix is determined, the number

of groups (k) is selected. In this research, 4 clusters are defined to validate the suggestion of cardiology specialists (Zheng et al., 2020).

8 Internal Evaluation Metrics

It is essential to evaluate the result of clustering algorithms; however, it is difficult to define when the result is acceptable (Hassani & Seidl, 2015). For this reason, there are techniques and indices for validating a grouping made. External and internal validation are the two most important categories for clustering validation. Internal validation metrics measure clustering solely based on information from the data (Hennig, 2023). In this project, the Silhouette coefficient (ASW), Pearson gamma correlation coefficient (Psgm), Dunn index (dunn), Dunn 2 index (dunn2), Entropy (Entropy), Within-Between ratio (WR), Calinski-Harabasz index (Ch) are explored internally. The values of the internal measurements are generated using the "cluster.stats" function of the fpc package (Baygin et al., 2021) in the R programming environment.

9 Experimental Design

This experimental design is structured for the analysis of cardiac arrhythmias using clustering in a large dataset (Figure 1), starting from selecting relevant features using the FSelector package in R, to reduce the dimensionality of the dataset. The RR interval of the ECG signal was extracted from mathematical formulations, given its clinical value in detecting arrhythmias. Two clustering algorithms, PAM and CLARA, were used to group the data because of their robustness in the presence of outliers and ability to handle large amounts of data. The clustering quality was assessed by comparing the groups of rhythms with and without additional cardiac conditions.

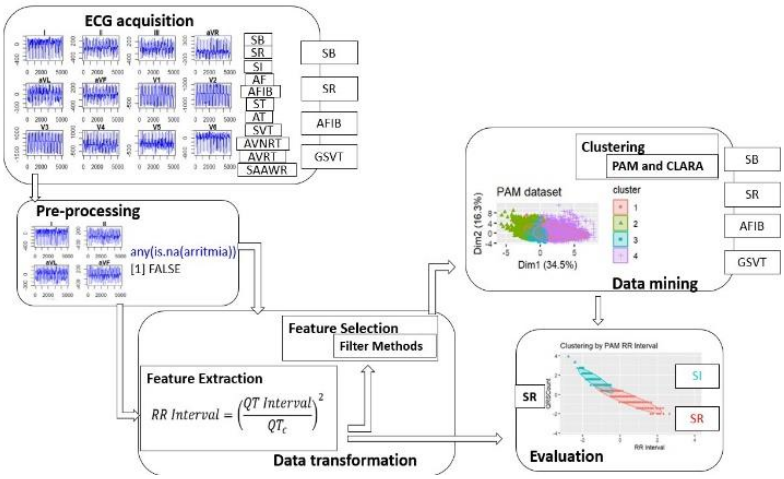

Fig. 1. Proposed methodology.

9.1 Feature Selection

After data preprocessing, as previously defined in (Zheng et al., 2020). Feature selection is performed to obtain the relevant variables that describe the dataset. The prevalence of five attributes was evident in the five methods used: VentricularRate, AtrialRate, QRSCount, QTInterval, and TOffset. This result coincides with what was reported in the state of the art (Yoon & Kang, 2023). From these obtained characteristics, cluttering is defined using the PAM and CLARA methods, which are evaluated in terms of the quality of the identified metrics.

Figure 2 compares the PAM and CLARA clustering methods using all the original features (16 attributes) versus those selected by FSelector (5 attributes). Table 4 presents the results of the calculation of the internal metrics. An improvement in clustering quality metrics is observed when only the features selected by FSelector are used, indicating the effectiveness of the feature selection process.

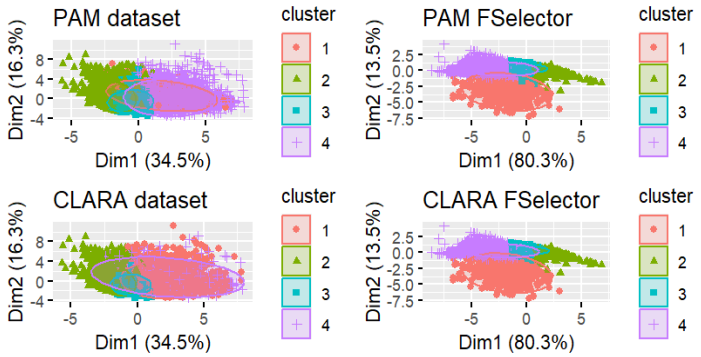

Fig. 2. Comparison of the PAM and CLARA clustering algorithms defined for cluster number 4 and clustering considering only the five variables defined by feature selection.

9.2 Feature Extraction

The RR interval is determined from equation 1, as defined in the previous sections. Figure 3 represents a synthesis of the dataset with the newly determined attribute.

```
> head(arritmia,n=3L)
                    FileName Rhythm     Beat PatientAge Gender VentricularRate
1 MUSE_20180113_171327_27000   AFIB RBBB TWC         85   MALE             117
2 MUSE_20180112_073319_29000     SB      TWC         59 FEMALE              52
3 MUSE_20180111_165520_97000     SA     NONE         20 FEMALE              67
  AtrialRate QRSDuration QTInterval QTCorrected RAxis TAxis QRSCount QOnset
1        234         114        356         496    81   -27       19    208
2         52          92        432         401    76    42        8    215
3         67          82        382         403    88    20       11    224
  QOffset TOffset RR_interval
1     265     386   0.5151535
2     261     431   1.1605898
3     265     415   0.8984970
```

Fig. 3. Example of a dataset with the RR attribute

By extracting this morphological characteristic of the signal, it is checked whether it constitutes relevance. Table 3 shows this new attribute among the first three when applying feature selection techniques. A total of five feature selection methods are applied.

Table 3. Results of applying the feature selection methods.

Feature	***Ranking***	***chi.sq***	***Ranking***	***I.G***	***Ranking***	***oneR***
RRInterval	1	0.5088	1	1.4578	3	0.6333
VentriRate	2	0.5084	2	1.4565	2	0.6334
AtrialRate	4	0.4610	3	1.3891	1	0.6346
QRSCount	3	0.4868	4	1.3316	4	0.6037
QTInterval	5	0.3375	5	0.7584	5	0.4677
TOffset	6	0.3256	6	0.7117	6	0.4490
QTCorrect	8	0.2564	7	0.2425	9	0.3061
PatientAge	7	0.2753	8	0.2214	7	0.3172
TAxis	9	0.2294	9	0.1778	8	0.3142
QRSDurat	11	0.1606	10	0.0830	10	0.2756
RAxis	14	0.1470	11	0.0579	12	0.2674
QOffset	13	0.1529	12	0.0484	11	0.2743
QOnset	10	0.1611	13	0.0297	13	0.2646
Gender	12	0.1124	14	0.0172	14	0.2610

The Consistency and CFS filters list a subset of the most relevant selected features. In contrast, Information Gain, OneR, and Chi-squared assign scores to each feature, indicating the order of relevance (Table 3).

Considering the groupings defined in the dataset (Zheng et al., 2020), the SR group comprises sinus rhythm and sinus irregularity because medical guidelines that define sinus irregularity can be easily separated from sinus rhythm using a single criterion: the variation of the RR interval. For computational verification, clustering was performed using the PAM algorithm (Figure 4). Visual inspection shows a relative separation of the clusters, despite the convergence of several records in both clusters.

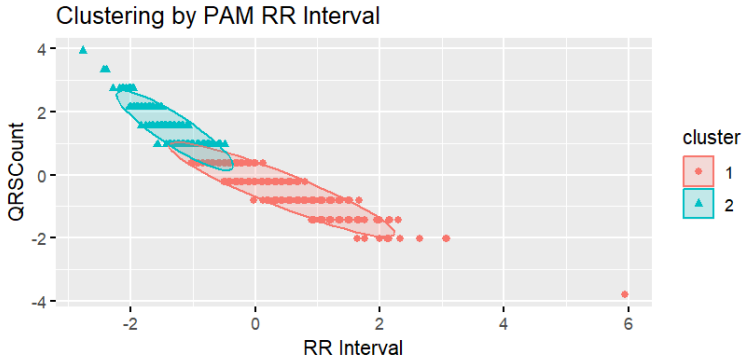

Fig. 4. Clustering by PAM for separation of the SR group.

To select the most representative variables concerning the RR_interval defined for the clustering application, the Pearson correlation matrix is calculated on the entire arrhythmia data set. From this matrix, pairs of variables with high correlation are identified. Figure 5 shows how the RR interval presents a strong correlation with the attributes VentricularRate, QRSCount, and QTInterval.

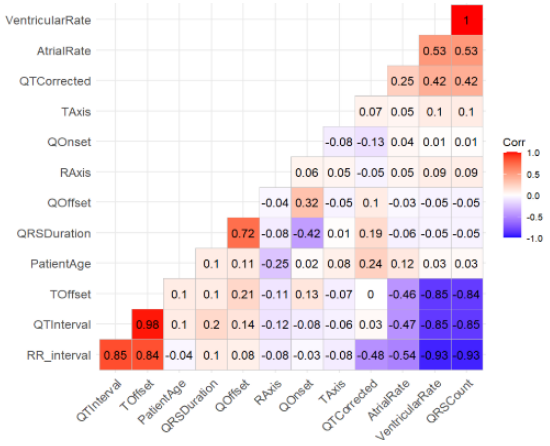

Fig. 5. Pearson correlation matrix for the dataset including the RR interval.

It is essential to highlight the SR group's data composition, as the rhythms that compose it present an imbalance between the classes. The normal sinus rhythm presents 1366 instances, and sinus irregularity presents 299.

9.3 Clustering

The value of each clustering quality metric is presented in Table 4. According to the interpretation of the data obtained, a prevalence of better results is observed for the PAM algorithm after applying feature selection and for records without additional conditions (shaded data in Table 4). These results are reflected in Figure 2 and Figure 4, where the dispersion of instances referring to each group is observed. ASW indicates better clustering with well-separated and cohesive clusters. Similarly, higher values of the Psgm reflect strong correlation and cohesion within clusters. Both Dunn and Dunn 2 Index Dunn2 also favor higher values, indicating well-separated and compact clusters.

In contrast, lower values of Entropy signify more homogeneous clusters. The WR is better when lower, suggesting lower variability within clusters than between clusters. Lastly, higher values of Ch suggest compact and well-separated clusters.

A particularity of the dataset under analysis is the presence of additional conditions in patients in addition to cardiac arrhythmias. However, in the current state of the art, only one study has been reported where the classification is taken into account, taking as reference the comparison between records with and without conditions. This analysis identified how additional cardiac conditions affected the clustering quality. A decrease in the clustering quality was observed when rhythms affected by other cardiac conditions were included, indicating more significant heterogeneity in these cases (Figure 6).

Table 4. Clustering quality metrics for the dataset with all its attributes and only for the variables defined by feature selection.

Cluster	NF	ASW	psgm	dunn	dunn2	Ch	WR	Entropy
Rango		[-1,1]		[0, ∞)				[0,log(k))
PAM	15	0.145	0.360	0.012	0.616	2782.3	0.637	1.222
	5	0.400	0.495	0.002	0.581	11197.9	0.380	1.225
CLARA	15	0.128	0.349	0.009	0.709	1177.2	0.804	1.136
	5	0.402	0.493	0.002	0.560	11037.7	0.381	1.230
Without additional conditions								
PAM	15	0.448	0.709	0.016	1.122	3713.6	0.453	1.232
	5	0.684	0.797	0.004	2.037	7670.7	0.251	1.230
CLARA	15	0.324	0.698	0.010	1.245	3817.9	0.678	1.222
	5	0.425	0.713	0.009	1.577	7569.6	0.345	1.200

NF: number of features

The improvement in rhythm grouping is evident when they do not present additional conditions. The rhythms having fewer records (AVNRT and AVRT) are grouped into a single cluster. However, the AFIB group has records in all clusters, representing a negative grouping aspect. On the other hand, in (Andayeshgar et al., 2022), the classification reported that most incorrect recordings occur between AF and AFIB classes, as well as between sinus rhythm (SR) and sinus irregularity (SA) classes.

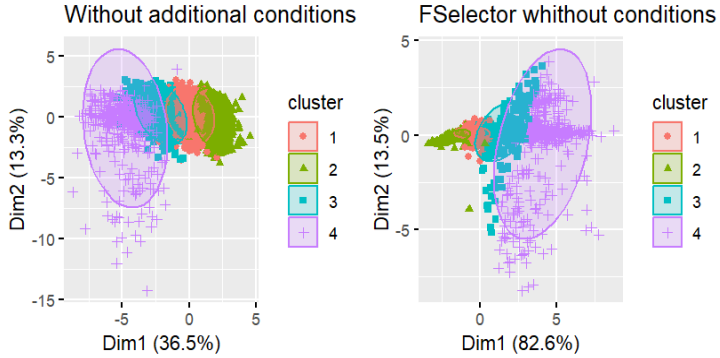

Fig. 6. Comparison of clustering using only records without additional conditions as reference and clustering considering only the five variables defined by feature selection.

10 Discussion and Results

This preliminary ongoing project analyzes a 12-lead ECG dataset from more than 10,000 patients to detect cardiac arrhythmias. Among the findings, a significant imbalance was observed between the classes and their treatment in the state of the art based on the grouping of the different types of arrhythmias, following medical guidelines. In the state-of-the-art review, only supervised

learning algorithms are considered for classification. Only one study (Zheng et al., 2020) has evaluated the influence of additional conditions on the classification of arrhythmias. New morphological features that are dependent on the ECG signal are extracted in most investigations.

The feature extraction process is analyzed using mathematical equations to obtain the RR interval, and this attribute's relevance is demonstrated in the dataset by applying feature selection. Based on the medical considerations that express the possible identification of sinus rhythm and sinus irregularities only considering the RR interval, a computerized grouping of the group defined as SR is carried out; even so, the data imbalance in the results is considered relevant. Considering that the dataset under analysis presents labels for cardiac conditions and arrhythmias, an analysis is performed to verify that the conditions do affect the classification.

11 Conclusions

The preliminary study addresses the problem of accurately detecting cardiac arrhythmias in the dataset. The complexity of the dataset was due primarily to data imbalance and the presence of additional cardiac conditions pose significant challenges to the effective classification of cardiac rhythms. To address this problem, a strategy was proposed based on feature selection and extraction of the RR interval, combined with applying PAM and CLARA clustering algorithms. In this approach, the FSelector tool is used to identify the most relevant attributes, allowing us to identify five particularly important attributes: RR_interval, QRSCount, QOffset, Q onset, and T offset. Subsequently, the PAM and CLARA clustering algorithms were applied, which allowed 4 groups to be identified in the dataset. The results were evaluated using clustering quality metrics and graphical visualizations. It was noted that additional cardiac conditions may affect the clustering quality, underscoring the importance of considering these variables in the analysis.

In future work, we will explore including more instances and applying other clustering algorithms and data preprocessing techniques. Furthermore, we will investigate the impact of different combinations of features and selection techniques to further optimize the detection of cardiac arrhythmias. On the other hand, the imbalance of the data and outliers will be treated using some existing machine-learning techniques.

References

Fu, D. (2015). Cardiac arrhythmias: Diagnosis, symptoms, and treatments. Cell Biochemistry and Biophysics, 73(2), 291-296. https://doi.org/10.1007/s12013-015-0626-4

Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V., & Roth, G. A. (2022). The global burden of cardiovascular diseases and risk. Journal of the American College of Cardiology, 80(25), 2361-2371. https://doi.org/10.1016/j.jacc.2022.11.005

Alipour, P., et al. (2022). Representation of women in atrial fibrillation clinical practice guidelines. Canadian Journal of Cardiology, 38(6), 729-735. https://doi.org/10.1016/j.cjca.2021.12.017

Lara Prado, J. I. (2016). El electrocardiograma: Una oportunidad de aprendizaje. Revista de la Facultad de Medicina (México), 59(6), 39-42. Recuperado de http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S0026-17422016000600039&lng=es&nrm=iso&tlng=es

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., & Rakovski, C. (2020). A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Scientific Data, 7(1), 48. https://doi.org/10.1038/s41597-020-0386-x

Cárcamo Morales, L. M., Osorio Iriarte, F. H., & Villegas García, J. D. J. (2020). Diseño de aplicación prototipo para la detección y clasificación de arritmia usando métodos de machine learning a partir de ECGs. Recuperado de https://manglar.uninorte.edu.co/handle/10584/9273

Zheng, J., et al. (2020). Optimal multi-stage arrhythmia classification approach. Scientific Reports, 10(1), 2898. https://doi.org/10.1038/s41598-020-59821-7

January, C. T., et al. (2014). 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: Executive summary: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. Circulation, 130(23), 2071-2104. https://doi.org/10.1161/CIR.0000000000000040

Ben-Tal, A., Shamailov, S. S., & Paton, J. F. R. (2012). Evaluating the physiological significance of respiratory sinus arrhythmia: Looking beyond ventilation-perfusion efficiency. The Journal of Physiology, 590(8), 1989-2008. https://doi.org/10.1113/jphysiol.2011.222422
MUSE v9 Devices and Interfaces Instruction Manual - SM - 2059568-013 - en | PDF | Gateway (Telecommunications) | Microsoft Windows. Recuperado de https://www.scribd.com/document/611556163/MUSE-v9-Devices-and-Interfaces-Instruction-Manual-SM-2059568-013-En
Murat, F., et al. (2021). Exploring deep features and ECG attributes to detect cardiac rhythm classes. Knowledge-Based Systems, 232, 107473. https://doi.org/10.1016/j.knosys.2021.107473

Mastoi, Q., et al. (2022). Novel DERMA fusion technique for ECG heartbeat classification. Life, 12(6), 842. https://doi.org/10.3390/life12060842

Rieg, T., Frick, J., Baumgartl, H., & Buettner, R. (2020). Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms. PLOS ONE, 15(12), e0243615. https://doi.org/10.1371/journal.pone.0243615

Yoon, T., & Kang, D. (2023). Multi-modal stacking ensemble for the diagnosis of cardiovascular diseases. Journal of Personalized Medicine, 13(2), 373. https://doi.org/10.3390/jpm13020373

Meqdad, M. N., Abdali-Mohammadi, F., & Kadry, S. (2022). A new 12-lead ECG signals fusion method using evolutionary CNN trees for arrhythmia detection. Mathematics, 10(11), 1911. https://doi.org/10.3390/math10111911

Faust, O., Kareem, M., Ali, A., Ciaccio, E. J., & Acharya, U. R. (2021). Automated arrhythmia detection based on RR intervals. Diagnostics, 11(8), 1446. https://doi.org/10.3390/diagnostics11081446

Yang, M.-U., Lee, D.-I., & Park, S. (2022). Automated diagnosis of atrial fibrillation using ECG component-aware transformer. Computers in Biology and Medicine, 150, 106115. https://doi.org/10.1016/j.compbiomed.2022.106115

Faust, O., & Acharya, U. R. (2021). Automated classification of five arrhythmias and normal sinus rhythm based on RR interval signals. Expert Systems with Applications, 181, 115031. https://doi.org/10.1016/j.eswa.2021.115031

Dash, M., & Liu, H. (2000). Feature selection for clustering: 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2000. Knowledge Discovery and Data Mining, 110-121. https://doi.org/10.1007/3-540-45571-x_13

Romanski, P., Kotthoff, L., & Schratz, P. (2023). FSelector: Selecting attributes. Recuperado de https://cran.r-project.org/web/packages/FSelector/index.html

Arritmia. (2024). Recuperado de https://www.texasheart.org/heart-health/heart-information-center/topics/arritmia/

Bazett, H. C. (1997). An analysis of the time-relations of electrocardiograms. Noninvasive Electrocardiology, 2(2), 177-194. https://doi.org/10.1111/j.1542-474X.1997.tb00325.x

Trastornos del ritmo cardíaco: Electrofiología, arritmias benignas. (2024). Recuperado de https://www.medwave.cl/puestadia/aps/4077.html

Aggarwal, C. C., & Reddy, C. K. (Eds.). (2014). Data clustering: Algorithms and applications. Chapman and Hall/CRC.

Clustering y heatmaps: Aprendizaje no supervisado con R. (2024). Recuperado de https://rpubs.com/Joaquin_AR/310338

Laurinec, P. (2024). Overview of clustering methods in R. Recuperado de https://www.r-bloggers.com/2024/01/overview-of-clustering-methods-in-r/

Hassani, M., & Seidl, T. (2015). Internal clustering evaluation of data streams. En X.-L. Li, T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, & D. Cheung (Eds.), Trends and Applications in Knowledge Discovery and Data Mining (pp. 198-209). Springer International Publishing. https://doi.org/10.1007/978-3-319-25660-3_17

Hennig, C. (2023). Fpc: Flexible procedures for clustering. Recuperado de https://cran.r-project.org/web/packages/fpc/

Baygin, M., Tuncer, T., Dogan, S., Tan, R.-S., & Acharya, U. R. (2021). Automated arrhythmia detection with homeomorphically irreducible tree technique using more than 10,000 individual subject ECG records. Information Sciences, 575, 323-337. https://doi.org/10.1016/j.ins.2021.06.022

Yildirim, O., Talo, M., Ciaccio, E. J., Tan, R. S., & Acharya, U. R. (2020). Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ECG records. Computer Methods and Programs in Biomedicine, 197, 105740. https://doi.org/10.1016/j.cmpb.2020.105740

Dhananjay, B., & Sivaraman, J. (2021). Analysis and classification of heart rate using CatBoost feature ranking model. Biomedical Signal Processing and Control, 68, 102610. https://doi.org/10.1016/j.bspc.2021.102610

Oliveira, R. F., Ferreira, A. A., Moreira, G. J. P., & Luz, E. J. S. (2022). Um método ensemble para classificação de arritmias: Uma avaliação com mais de 10 mil registros de sinais de ECG. En Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS) (pp. 13-24). SBC. https://doi.org/10.5753/sbcas.2022.222429

Meqdad, M. N., Abdali-Mohammadi, F., & Kadry, S. (2022). Meta structural learning algorithm with interpretable convolutional neural networks for arrhythmia detection of multisession ECG. IEEE Access, 10, 61410-61425. https://doi.org/10.1109/ACCESS.2022.3181727

Andayeshgar, B., Abdali-Mohammadi, F., Sepahvand, M., Daneshkhah, A., Almasi, A., & Salari, N. (2022). Developing graph convolutional networks and mutual information for arrhythmic diagnosis based on multichannel ECG signals. International Journal of Environmental Research and Public Health, 19(17), 10707. https://doi.org/10.3390/ijerph191710707

Guo, C., Ahmed, S., & Alouini, M.-S. (2023). Machine learning-based automatic cardiovascular disease diagnosis using two ECG leads. arXiv. https://doi.org/10.48550/arXiv.2305.16055

Lee, H., et al. (2021). Cardiac arrhythmia classification based on one-dimensional morphological features. Applied Sciences, 11(20), 9460. https://doi.org/10.3390/app11209460

Ozpolat, Z., & Karabatak, M. (2023). Performance evaluation of quantum-based machine learning algorithms for cardiac arrhythmia classification. Diagnostics, 13(6), 1099. https://doi.org/10.3390/diagnostics13061099