Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 3 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i03.58309

PAPER

Polynomial Characterization Clustering Method for Bayès Syndrome Detection

Lorena G. Franco[1,2] (✉), Luis A. Escobar Robledo[3], Rubén Wainschenker[2], Antoni Bayès de Luna[4], Hugo Curti[2], José M. Massa[2]

[1]Universidad Tecnológica Nacional FRD, Campana, Argentina

[2]UNCPBA, Tandil, Argentina

[3]Pittsburgh University, Pittsburgh, PA, USA

[4]Institut de Recerca del Hospital de la Santa Creu I Sant, Barcelona, Spain

lorenafranco@intia.exa.unicen.edu.ar

ABSTRACT

Several studies have established a strong association between Bayès Syndrome and multiple cardiovascular and neurological conditions. As it represents a potential risk for patients, its detection at an early stage is considered relevant. The morphology of the electrocardiogram (EKG) P wave is related to the detection of this syndrome. There are numerous works that detect and characterize the P wave. However, they are grounded in pure mathematical techniques. While effective, such methods often entail substantial costs in terms of computational time, particularly for screening applications in big data contexts. Our approach combines the robustness of wave characterization through the use of polynomial regression to identify relevant aspects of morphology. Clustering methods have been analyzed to classify the P wave: K-Means++ and fast autonomous unsupervised multidimensional (FAUM) using the polynomial coefficients as features. The obtained clustering and classification metrics for 49 P-waves show that the F1 Score = 1.00 with k = 3 and polynomial degree = 3. On the other hand, FAUM significantly enhances temporal efficiency compared to other implementations of K-Means++, making it particularly suitable for applications involving large sample sizes.

KEYWORDS

Syndrome de Bayès, electrocardiogram (EKG), polynomial, K-Means++, fast autonomous unsupervised multidimensional (FAUM)

1 INTRODUCTION

Bayès syndrome was discovered and studied by its namesake, Dr. Antoni Bayès de Luna. It became known in a series of his articles published in the last decades [1–4]. During the last few years, studies have revealed the association of Bayès syndrome to different conditions that do not necessarily involve the circulatory system. The concept of interatrial block (IAB) is the most frequent and relevant at the atrial level. IAB was divided in the same way as at the ventricular, sinoatrial, and atrioventricular levels into first-degree or partial, third-degree or advanced, and second-degree or intermittent [2, 4–12]. Bayès de Luna et al. [1] analyzed EKGs, demonstrating a

Franco, L. G., Robledo, L. A. E., Wainschenker, R., Luna, A. B. D., Curti, H., Massa, J. M. (2025). Polynomial Characterization Clustering Method for Bayès Syndrome Detection. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(3), pp. 170–186. https://doi.org/10.3991/ijoe.v22i03.58309

Article submitted 2025-08-19. Revision uploaded 2025-12-13. Final acceptance 2025-12-13.

prevalence of Advanced IAB (A-IAB) of 1%, whereas when only patients with structural heart disease were selected, the prevalence was 2%.

Interatrial block has been associated with medical disorders such as atrial fibrillation (AF), myocardial ischemia, left atrial enlargement, and systemic emboli [13]. IAB is considered a risk factor for cardioembolic stroke [11–12, 14]. In middle-aged individuals, advanced IAB is associated with a threefold increase in the risk of AF and almost a twofold increase in the risk of stroke. Additionally, P wave duration has been linked to cardiovascular mortality and sudden cardiac death [12]. At very advanced ages, the presence of IAB is further related to total mortality. Some studies show that the prevalence of dementia progressively increased when passing from normal P wave to partial IAB, advanced IAB, and AF [15]. Considering the aforementioned, early recognition of this condition is crucial. The diagnosis of partial or advanced blocks can be made through an EKG analysis.

With regard to EKGs, although modern electrocardiographs are digital, many centers still have analog paper-based equipment or lack storage infrastructure. In light of the above and considering the need for have EKGs monitor patients with IAB over the years, keeping those results is not an easy task. The processed EKGs were originally in paper format, therefore requiring their digitization. This was done considering that the P wave must be preserved. In previous works [16], the authors explored digitalization and segmentation techniques aimed at preserving it. Clustering methods were also applied to automatically group the samples [17–18].

Within the existing literature, numerous methods and tools for EKG analysis have been presented [19–20]. In particular, there are publications aimed at identifying the P wave [21–24], but there are no methods that identify the IAB or A-IAB. Regarding the wave analysis of the cardiac cycle, a significant number of methods were applied based on frequency analysis, such as wavelets and Fourier [25–27], among others. Another approach consists of exploring the problem from a classification point of view.

Unsupervised classification techniques do not involve any type of supervised information, which is beneficial for their application, but they must be carefully used in diagnostic-related problems. On the other hand, supervised classification techniques require a large amount of labeled data to train the models that will be used for classification. This poses a disadvantage since labeling large amounts of data is a time-consuming and expensive task. In the last decades, there has been a growing interest in the use of semi-supervised clustering, which has emerged as a solution to the challenges faced by both supervised and unsupervised classification techniques [28]. As previously mentioned, clustering techniques were used in a semi-supervised manner. Regarding clustering methods, K-Means (and its more modern variants such as K-Means++) [29–30] have proven to be successful in a large number of problems [31]. This method clusters data into a fixed number of classes, k. The other one used in this work is based on a novel technique called Fast Autonomous Unsupervised Multidimensional (FAUM) [32], which incorporates a heuristic algorithm based on an analysis of the entropy and the balance of cardinality between classes. FAUM automatically finds several classes in an unsupervised manner. Additionally, a fixed number of classes can be established, similar to K-Means, although internally it uses a hierarchical algorithm based on the cardinality of classes and the use of non-Euclidean distance functions in cases of dimensions greater than 3. The choice of the distance function is in principle relevant because it determines how the cluster will work. The methods belong to the category of unsupervised clustering methods, but it is possible to use them as classifiers, applying them in a semi-supervised way. The clustering technique, as in previous works

[17, 18], was evaluated using the confusion matrix analysis indicators: precision, accuracy, recall, and F1 Score. In this work, it was also decided that the results would be evaluated with the performance indicators of the clustering methods, such as the Dunn index and the Silhouette index [33, 34]. In this way, it is possible to observe whether there are discrepancies or if they are similar.

To identify the morphology of the EKG P wave, this work presents a combination of polynomial approximation and the use of clustering methods. The signals, from which the polynomial coefficients were obtained, have been normalized concerning their amplitude as well as the duration of time. The clustering methods K-Means++ and FAUM were applied to the polynomial coefficients. These methods were tested in a semi-supervised and predictive way to subsequently evaluate their application in the detection of Bayès syndrome. Recently, clustering methods have been applied as classifiers, a technique known as classification by clustering [35–36]. In this context, various approaches have been employed, including weakly supervision [37], semi-supervision [28, 38], and others [39].

Although, in general, EKG availability is widespread, in the case of Bayès syndrome, a disease with relatively low incidence, there are no labeled datasets. In this work, a dataset provided by the research group that first identified the syndrome was selected. This choice was not based on the limited number of samples, but rather on the fact that these data have become a reference for the syndrome and were used in the original publications [2].

Given the importance of detecting the presence of Bayès syndrome and the motivation to contribute to this using computational techniques, the main objective of this work is to develop a method capable of recognizing different EKG-P wave morphologies, particularly those related to an early stage of Bayès syndrome. Additionally, the potential for efficiently applying this method to large volumes of data, such as screening purposes, was also considered.

Many techniques have been proposed for signal reconstruction and compression in this field, such as Fourier transforms (FT), discrete cosine transforms (DCT), and wavelet transforms. The primary distinction between the aforementioned methods lies in the functions employed. The most commonly used is the polynomial basis. Polynomials are mathematical functions that require only multiplication and addition for their evaluation. They also offer the advantage of flexibility, enabling the representation of a wide range of nonlinear relationships. Polynomials are frequently used in signal processing [40]. Considering that polynomial approximation has been shown to be efficient in EKG modeling [41], it was decided to model P-wave morphology using a polynomial approach, exploring various polynomial degrees to identify the most suitable one. In the existing literature, polynomials of degree 2, 3, and higher degrees have been explored for EKGs [42–46].

The scientific gap that this paper attempts to bridge is the use of efficient methods to describe the P-wave using polynomials together with advanced computational classification methods such as FAUM, which is not covered in the literature.

A synthesis of the materials and the proposed clustering method is provided in Section 2. The results obtained are presented in Section 3. Section 4 contains discussions, and finally, Section 5 presents the conclusions.

2 MATERIALS AND METHODS

This section presents the pipeline composed of a series of successive steps described in each of the sections (data input and output between '{' and '}', processes

between '[' and ']'): {paper EKG} –> [Capture, section 2.2] –> {digitized EKG} –> [Standardization, section 2.2.1] –> {Floating point and integer fixed point presentation EKG} –> [Polynomial characterization, section 2.2.2] –> {polynomial coefficients} –> [Clustering methods, sections 2.2.3, 2.2.4, 2.2.5] –> {results}.

In the following section, the rest of the pipeline is presented: {results} –> [Evaluation, section 3] –> {metrics}.

2.1 Materials

Forty-nine samples from a total of 600 EKGs from the research of Dr. Bayès and his working group were used in the context of a joint collaborative project. Since the most relevant wave (biphasic) occurs at a very low rate in older adults, using all samples would lead to a significant imbalance between this class and the rest, especially the normal class. Although the sample size may seem small, studies with similar sample sizes have been reported [47]. These samples consist of paper-supported EKG images, which were scanned at a sufficient resolution to allow the digitization of amplitude levels and wave duration with an acceptable error. Unlike other investigations where the same time baseline was used in each image following the guidelines of [48], in this work it was decided to take into account the time baseline of the P wave morphology; therefore, depending on the characteristics of the wave, its duration varies in time. Then a digitization process based on previous work [16] was applied to these samples to preserve the P wave. Examples of P waves are shown in Figure 1.

The images obtained through the digitization (1200 dpi) process constitute the materials of this work. At first, binarization and thresholding (Otsu method) were applied to obtain a curve of the smallest possible thickness and therefore improve accuracy. A skeletonization was applied to the results of images with an iterative multiple erosion method. This process is illustrated in Figure 2a–2d. To establish a reference for the positive and negative values of the wave, a method based on the manual technique used by clinicians was applied [49]. Finally, a list of intensity values was obtained for each column of the image, corresponding to each temporal sampling element of the EKG. This list of values was initially calculated with high floating-point precision and then normalized between –1 and 1.

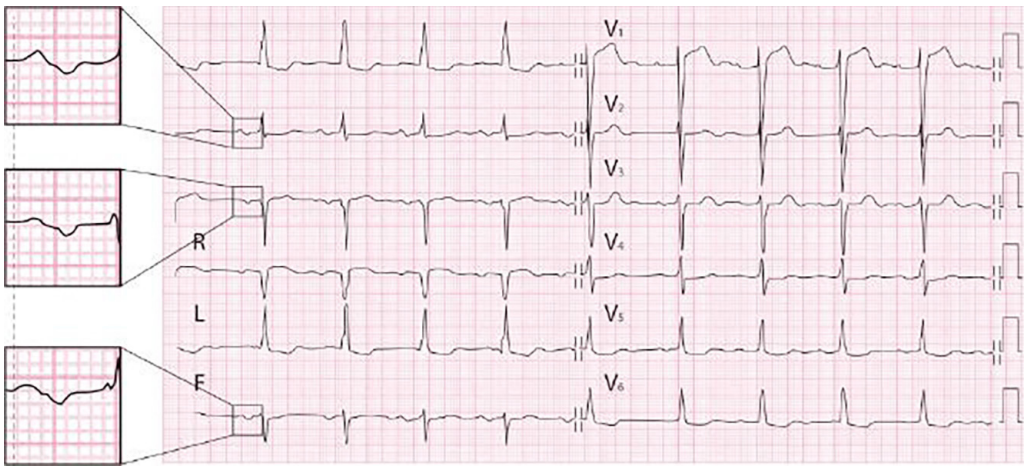

Fig. 1. EKG with AIAB with P waves highlighted with rectangles at the left

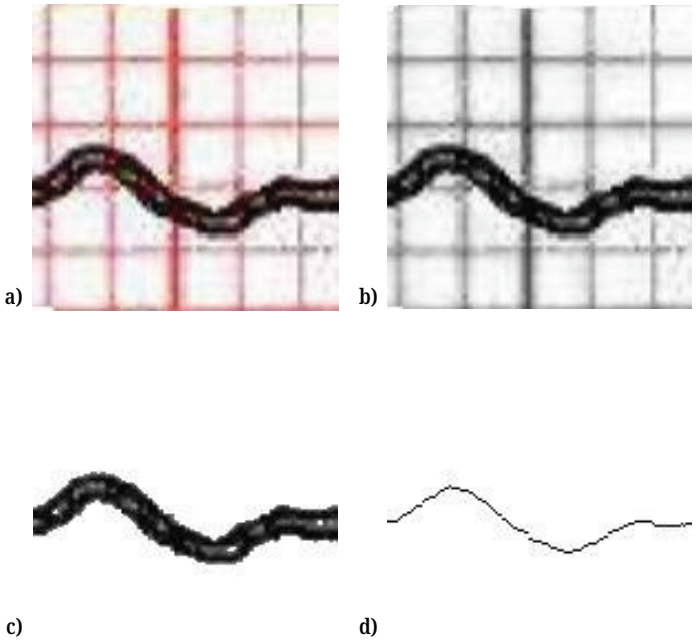

Fig. 2. Image digitization: a) Original image, b) Binarized, c) Thresholding, and d) Skeletonization

Figure 3 shows the different P wave morphologies considered in this work; extracted from the dataset and skeletonized.

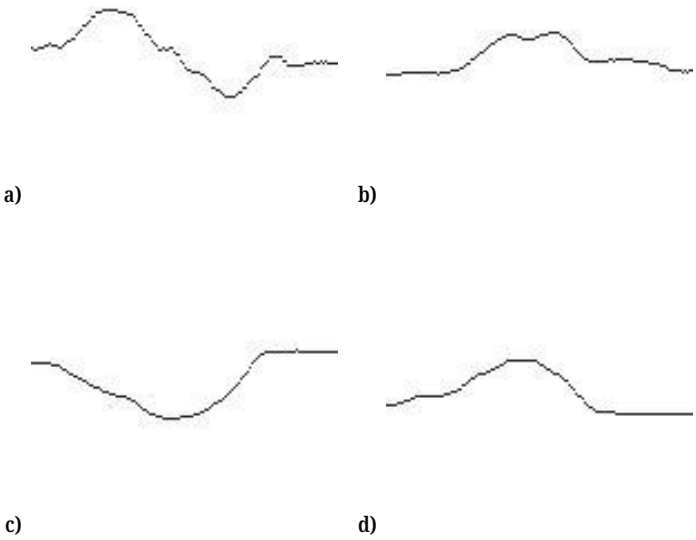

Fig. 3. P wave morphology: a) Biphasic P wave, b) Bimodal P wave, c) Negative P wave, and d) Normal P wave

The composition of the dataset used in the tests is shown in Table 1. The number of biphasic P waves corresponds to that indicated in [1], considering that they can be found in three leads of an EKG, which contains a few cycles.

Table 1. Relevant sample

P Wave Morphology	Number of Samples
Biphasic P wave	23
Bimodal P wave	3
Negative P wave	8
Normal P wave	15
Total	49

Regarding the FAUM clustering software, the latest updated version, 1.0–2 (2021), was used.

2.2 Methods

To apply the clustering methods detailed in section I, it was necessary to build a set of features for each of the available signals. It is important to mention that the raw features (which represent the amplitude of the signal in each of the measurement periods) originated as a result of a physio-biological process that constitutes the patient's cardiovascular system, which is a somewhat integral phenomenon; therefore, the features have a strong temporal dependence on each other (in simple terms, it means that the amplitude of the wave at a certain time has a strong relationship with previous values). This means that the assumption of independence or orthogonality of the features that is desirable in the application of machine learning methods and, in particular, in clustering methods [50], is not met in this case. However, what happens is that although the behavior in this case of the P wave has a certain predictability as to what a P wave is expected to be, there are physiological phenomena, signal capture, and the Bayès Syndrome itself that alter the shape of the wave. The raw signals consist of a matrix of values, whose rows contain the different samples and the columns with the amplitude values for each sampled period.

Although in previous works [17–18] an attempt was made to cluster the curves using directly the distribution of the amplitudes, this had the disadvantage that the number of amplitudes was different in each curve. Taking into account that the shape of the curve determines which group belong to the signal in the sense of those shown in Table 1, Section 2.1, the polynomial-based method is explained.

Standardization. Since the importance of standardization lies in using a common scale without distorting the differences in the value of intervals or losing information, it was applied in this work.

The amplitude values in each column for each P wave sample, corresponding to each EKG time-sampling element, were initially calculated with high floating-point precision. Values were normalized between –1 and 1.

To subsequently calculate the polynomial coefficients, in addition to having normalized the amplitude values, the X values must be normalized between –1 and 1. Therefore, the values of the ordinate axis and those corresponding to the abscissa axis were normalized between –1 and 1.

Polynomial Characterization. One of the most used methods of curve fitting, which allows representation of the shape of curves from coefficients, is polynomial approximation. Having the same number of these coefficients for each curve allows clustering curves similar to each other by clustering the coefficients similar to each other.

Taking as a starting point the work of Chaturvedi [42], different polynomial degrees were considered: 2, 3, 4, 5, and 10, among others. Fitting was done using

the Matlab® 2020a polyfit function, discarding the independent coefficient since it is not related to the wave morphology. Once the coefficients of the polynomial of degree n have been obtained, in order to apply the FAUM clustering method, it was decided to digitize the values of coefficients with a precision of 2 bytes, which leads to 65536 possible values, since 1 byte allows to consider only 256 possible values, which is insufficient considering the resolution of the digitization. Therefore, all the coefficients of the 49-row and n+1-column matrix, with n being the polynomial degree, matrix were considered so that they can take values between 0 and 65535.

Specifically, two sets were constructed: the *Mfp* set, containing floating point values between –1 and 1 for the values of the polynomial coefficients, and the *Mip* set containing integer values between 0 and 65535. The values of the *Mip* set have been normalized because the FAUM method is based on magnitude scale shifting operations in fixed-point representations.

In Figure 4, an example of a 3-degree polynomial fitting of a biphasic P wave is shown. Since the goal of this fitting is to obtain the coefficients to describe the P wave instead of representing the waves, the evaluation error of the fitting is not needed for the clustering process.

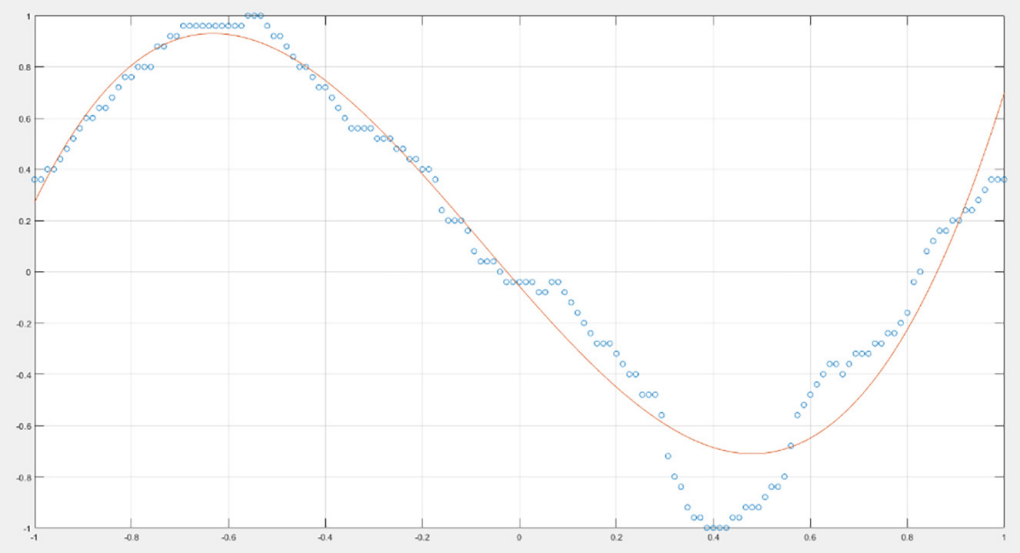

Fig. 4. P wave with A-IAB approximated with a polynomial of 3-degree

A chart of the coefficients of 3-degree polynomials without the independent term for each sample can be seen in Figure 5.

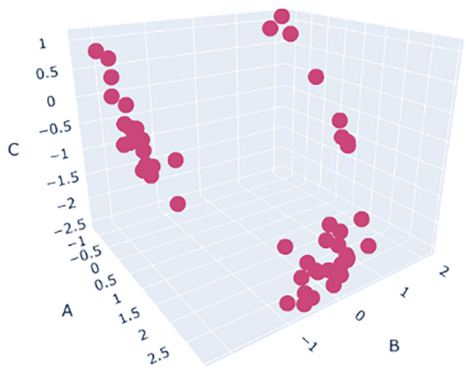

Fig. 5. 3-Degree polynomial coefficients for the 49 P wave samples

K-Means++ clustering. The K-Means++ method implemented in Matlab was applied to the *Mfp* set, and the K-Means++ implementation of FAUM (faum –1.0–2) was applied to the *Mip* set. As with the degrees, different values for k were used: 2, 3, 4, and 5 clusters. Since the number of morphologies is 4, as stated in section 2.1, the obvious setting of k is 4. However, there may be the possibility that 2 or 3 clusters may form. Regarding the initialization, Matlab K-Means++ uses the Mersenne Twister random number generator with a default seed = 0. Also, it is possible to consider the possibility of having subclusters inside any cluster or an additional cluster formed by outliers.

In the case of FAUM, the initialization method on which this implementation is based is derived from the same method used by the K-Means++ implementation of Matlab, but it has been modified to further restrict the randomness of the initial seeds (seconds from 00:00:00 UTC, 1 Jan 1970 to the start). As a consequence, performance is improved, achieving a good result with a smaller number of iterations in comparison with the Matlab implementation. The distance measurement function chosen at first was the Euclidean distance in both implementations, since this distance is suitable for the number of features related to the polynomial degrees. Additionally, other distances were considered. In the case of K-Means++ of FAUM, some tests performed with the Chebyshev distance are also presented later in Section 3. The reason for choosing Chebyshev is that in the case of having polynomials of higher degrees, this distance is more accurate than Euclidean. Another advantage of the FAUM implementation is the use of fixed-point representation for the values of the feature vectors. This allows to increase the time efficiency over other K-Means++ implementations by a significant factor, depending on the number of samples, features, and the precision (for example, it has been demonstrated that a sample size increase of 43.12 factor results in a 1.66 factor reduction in processing time). This represents a very interesting benefit when it must be applied to a significant number of samples.

In addition to the already mentioned Euclidean and Chebyshev distances, other distance functions such as correlation, cosine, and cityblock in the K-Means++ method of Matlab were used. Also, clustering discarding the independent term was explored.

FAUM clustering with fixed number of clusters. It is important to mention that FAUM [51] consists of a deterministic and heuristic method that allows the discovering of natural clusters in a dataset. From the input data, it generates multidimensional histograms in an iterative way, establishing different granularity sizes (hyperbins) from these histograms maximizing the balance of the cardinality of the clusters found with the hyperbins. This method was initially created as a solution to obtain cluster centroids and to use them to initialize other methods such as K-Means++. In addition, it can be used for semi-supervised classification if the number of classes is known, as in the case of this work.

Unlabeled samples were used for the execution of FAUM, using different values of k, and then the effectiveness of this method was evaluated as if it were a classifier, using the label of each sample to calculate the values of the confusion matrix and the statistical indicators associated with this matrix. The Silhouette and Dunn clustering indicators were also calculated. Since FAUM works with fixed-point or integer data (to use bit shift operations to gain efficiency), the *Mip* set was used, which is normalized and expressed integer values. In this way, the initial data was converted to PAM format [52], its consistency was verified, and FAUM was used, limiting the heuristic method to find clustering solutions with k classes. All FAUM parametrization values, with the exception of the minimum and maximum number of clusters and distance, were kept by default according to [32].

FAUM-initialized K-Means. FAUM clustering with a fixed number of clusters was used to obtain the centroids and initialize the K-Means++ method with this knowledge. In this way, K-Means++ is run deterministically to obtain the clustering results.

The unlabeled samples were used for the adjusted FAUM run, and the centroids corresponding to the different values of k explored were obtained. The effectiveness of this method was evaluated as in the previous section.

3 RESULTS

As stated above, in this research work, we applied the Matlab K-Means++ algorithm, FAUM K-Means++, FAUM in adjusted mode, and FAUM-initialized K-Means, in all cases using Euclidean distance. Additionally, FAUM-initialized K-Means was used with Chebyshev distance. The code is available at[1].

The composition of the samples used in the tests was shown in Table 1 of section 2.1. The number of biphasic P waves of interest corresponds to that indicated in [1], considering that they can be found in 3 EKG leads and that there was generally only one P wave per lead.

Following, clustering results with k = 2, 3, and 4 using the coefficients of the 3rd-degree polynomial are described. For each case, the confusion matrix was calculated. Although tests were performed considering and eliminating the independent term, the best result was obtained with k = 3. Therefore, the results take into account the 3 values of the coefficients of the 3rd-degree polynomial. Accuracy, precision, recall, and F1 score metrics were also calculated. Finally, the total metrics were calculated on the confusion matrix, weighing each of them by the number of samples of each class.

Below are the most relevant results obtained with the polynomial of 3-degree using 3 clusters. The P wave morphology types used are indicated in Table 2.

Table 2. Relevant sample for k = 3

P Wave Morphology	Number of Samples
Biphasic P wave	23
Normal P wave (includes bimodal)	18
Negative P wave	8
Total	49

The results of the confusion matrix specified for each type of P wave morphology are presented in Table 3. This matrix arises from applying the K-Means++ algorithm (both implementations) and FAUM to the coefficients of the 3-degree polynomial considering 3 clusters.

Table 3. Confusion matrix for each type of P wave morphology

	TP	TN	FP	FN	N
Biphasic	23				Total
K-Means++ Matlab	23	26	0	0	49
K-Means++ FAUM	23	26	0	0	49
FAUM adjusted	23	21	5	0	49
FAUM-initialized K-Means E	23	26	0	0	49
FAUM-initialized K-Means Ch	23	26	0	0	49

(Continued)

[1] https://github.com/francol-unicen/Clustering.git

Table 3. Confusion matrix for each type of P wave morphology *(Continued)*

	TP	TN	FP	FN	N
Positive (includes bimodal)	18				Total
K-Means++ Matlab	18	31	0	0	49
K-Means++ FAUM	18	31	0	0	49
FAUM adjusted	16	31	0	2	49
FAUM-initialized K-Means E	18	31	0	0	49
FAUM-initialized K-Means Ch	18	31	0	0	49
Negative	8				Total
K-Means++ Matlab	8	41	0	0	49
K-Means++ FAUM	8	41	0	0	49
FAUM adjusted	5	41	0	3	49
FAUM-initialized K-Means E	8	41	0	0	49
FAUM-initialized K-Means Ch	8	41	0	0	49

The results obtained for the Accuracy, precision, recall, and F1 score values are shown in Table 4.

Table 4. Performance metrics for 3-degree polynomial and k = 3

	Acc	Prec	Rec	F1 Score
Biphasic				
K-Means++ Matlab	1.00	1.00	1.00	1.00
K-Means++ FAUM	1.00	1.00	1.00	1.00
FAUM adjusted	0.90	0.82	1.00	0.90
FAUM-initialized K-Means E	1.00	1.00	1.00	1.00
FAUM-initialized K-Means Ch	1.00	1.00	1.00	1.00
Positive (includes bimodal)				
K-Means++ Matlab	1.00	1.00	1.00	1.00
K-Means++ FAUM	1.00	1.00	1.00	1.00
FAUM adjusted	0.96	1.00	0.89	0.94
FAUM-initialized K-Means E	1.00	1.00	1.00	1.00
FAUM-initialized K-Means Ch	1.00	1.00	1.00	1.00
Negative				
K-Means++ Matlab	1.00	1.00	1.00	1.00
K-Means++ FAUM	1.00	1.00	1.00	1.00
FAUM adjusted	0.94	1.00	0.63	0.77
FAUM-initialized K-Means E	1.00	1.00	1.00	1.00
FAUM-initialized K-Means Ch	1.00	1.00	1.00	1.00

The total indicators obtained on the confusion matrix, weighting each of them by the number of samples in each class, are shown in Table 5.

Table 5. Total performance metrics for 3-degree polynomial and k = 3

	Total Acc	Total Prec	Total Rec	Total F1 Score
K-Means++ Matlab	1.00	1.00	1.00	1.00
K-Means++ FAUM	1.00	1.00	1.00	1.00
FAUM adjusted	0.93	0.92	0.90	0.89
FAUM-initialized K-Means E	1.00	1.00	1.00	1.00
FAUM-initialized K-Means Ch	1.00	1.00	1.00	1.00

3.1 Results of the silhouette clustering indicator

The silhouette indicator was calculated for the polynomial of degree, and considering 2, 3, and 4 clusters. The most notable results for the three clusters are shown below.

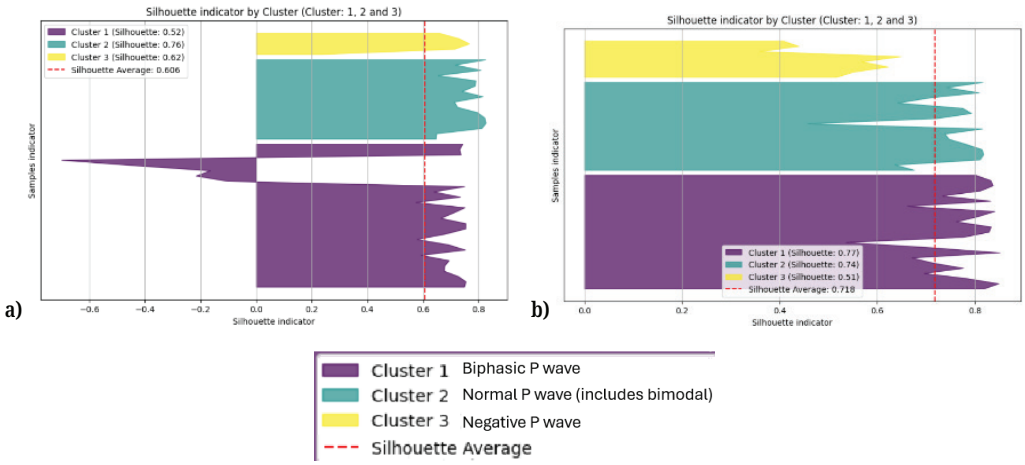

Fig. 6. Silhouette indicator a) FAUM in adjusted mode, b) FAUM initialized K-means and Chebyshev distance

In Figure 6, a) using FAUM in adjusted mode, Cluster 1 corresponding to the biphasic P waves shows that some samples are close to another cluster. However, in b), where FAUM initialized K-Means and Chebyshev distance is applied, all the biphasic P wave samples are cohesive to their cluster. This last result is also maintained when applying the indicator with the results obtained from K-Means++ Matlab, K-Means++ of FAUM, and FAUM-initialized K-Means and Euclidean distance.

3.2 Results of the Dunn Clustering Index

The value obtained for the 3-degree polynomial with 3 clusters is 0.47 for Matlab K-Means++, FAUM K-Means++, and FAUM-initialized K-Means, applying in all cases a Euclidean distance. On the other hand, for FAUM-initialized K-Means and Chebyshev distance, the value obtained was 0.41. The worst result was obtained

with FAUM in adjusted mode. The values obtained for the 3-degree polynomial with 2 and 4 clusters were lower.

4 DISCUSSIONS

In this section we present an analysis of the results and possible explanations in relation to relevant aspects of the problem. The main difficulty is to distinguish the samples belonging to each one of the 4 morphologies. When there are 4 clusters, the one corresponding to the bimodal morphology overlaps with the one of the normal morphology. This can be observed in Figure 7, where the violet circle that represents the values of the non-independent coefficients of the polynomial of 3 degrees of the bimodal morphology is located inside the cloud that represents the normal morphology identified by the yellow circles. The bimodal and normal P wave morphologies are positive P waves.

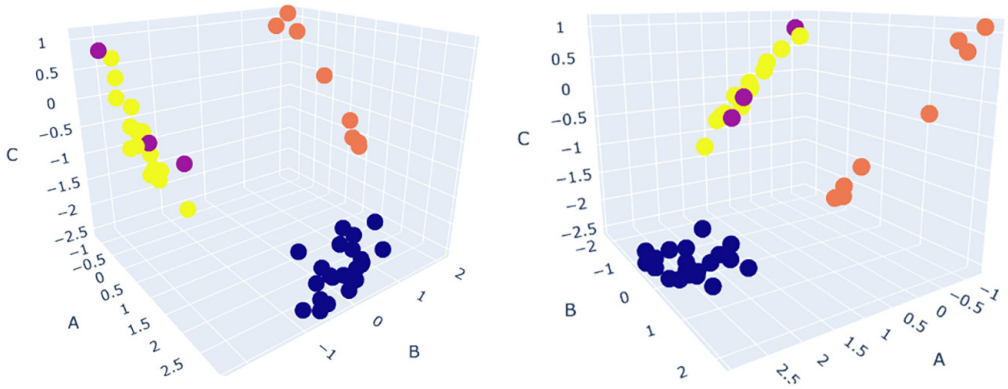

Fig. 7. Chart of the 3-degree coefficients with the P wave morphology clusters from Table 1

The highest Dunn index is obtained for 3 clusters, thus indicating that using 3 clusters is more compact and the clusters are further apart than considering more or fewer clusters.

When the results for the same degree but with 2 clusters are analyzed, some of the negative waves are clustering with the biphasic waves, which makes sense since the P waves that present the 3rd-degree block have a positive and then negative morphology.

Although the indicators have not been placed in the results section when trying to use 5 clusters in a polynomial of 3 degrees, the biphasic waves are still correctly identified, but a subcluster within the negative waves and another subcluster within the positive waves are obtained. This result is the same for K-Means++ in either implementation, while with adjusted FAUM it does not detect all the biphasic P waves.

During the experiment, clustering methods were used with polynomials of degrees 2, 3, 4, 5, and 10. The results of using clusters of 2, 3, 4, and 5 were explored. However, A-IAB detection performance decreased in cases not detailed in the results section. As the degree of polynomial increased, performance decreased. Also, when higher than 3-degree polynomials were used, instability and overfitting were observed and hence generalization capabilities decreased.

When working with 3 degrees and considering 3 clusters, the results of FAUM initialized K-Means with Euclidean or Chebyshev distance coincide with those obtained

when applying K-Means++ in the classical form of FAUM and Matlab. However, when running adjusted FAUM, false positives of the biphasic wave are presented. In its defense, it can be said that FAUM in adjusted mode was designed to work with a larger number of samples.

It is important to note that K-Means++ is non-deterministic, as its initialization depends on random or pseudo-random numbers; in contrast, FAUM is fully deterministic. When K-Means is initialized with centroids obtained from FAUM, it also becomes fully deterministic. The inherent randomness of K-Means++ makes the comparison of its results challenging, as different runs may yield varying outcomes. This is particularly relevant when comparing the MATLAB implementation of K-Means++ with the FAUM implementation of K-Means++, since there is no straightforward way to synchronize the pseudo-random number generators (PRNGs) of to guarantee identical results.

Regarding future work, we plan to increase the data set size by contacting more specialized medical institutions. Supervised methods are considered, but they depend strongly on expert data labeling. Also, data augmentation will be applied after the augmenting conditions for the P-wave are defined with the experts, which are currently under discussion.

5 CONCLUSIONS

In this work, the coefficients of a 3rd-degree polynomial were chosen as the better features of the signals provided by Dr. Antonio Bayès de Luna's team to classify the EKGs and interpolate the EKG curves. By using polynomials, the aim is to characterize the morphology of each P wave, emphasizing its most significant features. From the tests using different numbers of clusters, the most effective result was achieved with 3 clusters.

By applying the K-Means++ methods (both implementations) and FAUM-initialized K-Means for both Euclidean and Chebyshev distance, 100% of the P waves presenting the A-IAB were identified. This is confirmed by the total indexes of accuracy, precision, recall, and F1 Score. By performing the analysis of clustering indexes, it is also confirmed that considering 3 clusters is the best option; clustering is more robust.

FAUM has the advantage of using fixed-point representation for the values of the characteristic vectors of the samples. As a consequence of this improvement, time efficiency is increased by a significant factor concerning other K-Means++ applications. This depends on the number of samples, characteristics, and the -bit precision of the representation of the values of the latter. Therefore, it is designed to work with large amounts of data in the context of online processing within the flow of medical practice. Although it has not been tested, based on the literature [40–42], the polynomial approximation method can be used to describe the morphology of other ECG waves. Furthermore, FAUM and K-Means++ do not depend on training with the specific morphology of the P wave, so in principle they can be generalized to other waves and pathologies.

The robustness of FAUM and FAUM-initialized K-Means++ relies not only on the temporal and spatial efficiency but also on the non-probabilistic nature of these methods compared with K-Means, hierarchical and most clustering methods.

As for limitations, in principle, the current version of FAUM considers hypercubes for dimensions, and this presents some difficulties for detecting nonlinear clusters.

In the future, we plan to incorporate nonlinear geometries, which will improve the detection of more complex clusters. Another limitation is that the number of samples did not allow us to use supervised machine learning methods that require training with a larger number of samples.

6 REFERENCES

[1] A. B. De Luna *et al.*, "Electrocardiographic and vectorcardiographic study of interatrial conduction disturbances with left atrial retrograde activation," *Journal of Electrocardiology*, vol. 18, no. 1, pp. 1–13, 1985. https://doi.org/10.1016/S0022-0736(85)80029-7

[2] A. Bayés de Luna *et al.*, Iturralde, "Interatrial conduction block and retrograde activation of the left atrium and paroxysmal supraventricular tachyarrhythmia," *European Heart Journal*, vol. 9, no. 10, pp. 1112–1118, 1988. https://doi.org/10.1093/oxfordjournals.eurheartj.a062407

[3] L. Bacharova and G. S. Wagner, "The time for naming the interatrial block syndrome: Bayes syndrome," *Journal of Electrocardiology*, vol. 48, no. 2, pp. 133–134, 2015. https://doi.org/10.1016/j.jelectrocard.2014.12.022

[4] A. Bayés de Luna, "Bloqueo a nivel auricular," *Rev. Esp. Cardiol.*, vol. 32, no. 1, pp. 5–10, 1979.

[5] D. Conde and A. Baranchuk, "What a cardiologist must know about the Bayes' syndrome," *Revista Argentina de Cardiología (RAC)*, vol. 82, no. 3, pp. 237–239, 2014. https://doi.org/10.7775/rac.v82.i3.3862

[6] D. Conde and A. Baranchuk, "Bloqueo interauricular como sustrato anatómico-eléctrico de arritmias supraventriculares: Síndrome de Bayés," *Archivos de cardiología de México*, vol. 84, no. 1, pp. 32–40, 2014. https://doi.org/10.1016/j.acmx.2013.10.004

[7] A. B. De Luna, A. Baranchuk, L. A. E. Robledo, A. M. van Roessel, and M. Martínez-Sellés, "Diagnosis of interatrial block," *Journal of Geriatric Cardiology: JGC*, vol. 14, no. 3, pp. 161–165, 2017. https://doi.org/10.11909/j.issn.1671-5411.2017.03.007

[8] A. Baranchuk, P. Torner, and A. B. de Luna, "Bayés syndrome: What is it?" *Circulation*, vol. 137, no. 2, pp. 200–202, 2018. https://doi.org/10.1161/CIRCULATIONAHA.117.032333

[9] M. Martínez-Sellés *et al.*, "Interatrial block and cognitive impairment in the BAYES prospective registry," *International Journal of Cardiology*, vol. 321, pp. 95–98, 2020. https://doi.org/10.1016/j.ijcard.2020.08.006

[10] M. Martínez-Sellés *et al.*, "Advanced interatrial block and P-wave duration are associated with atrial fibrillation and stroke in older adults with heart disease: The BAYES registry," *EP Europace*, vol. 22, no. 7, pp. 1001–1008, 2020. https://doi.org/10.1093/europace/euaa114

[11] P. A. Iomini, M. Martínez-Sellés, R. Elosua, A. Bayés-de-Luna, and A. Baranchuk, "Síndrome de Bayés, accidente cerebrovascular y demencia," *Archivos Peruanos de Cardiología y Cirugía Cardiovascular*, vol. 2, no. 1, pp. 27–39, 2021. https://doi.org/10.47487/apcyccv.v2i1.126

[12] R. Bejarano-Arosemena and M. Martínez-Sellés, "Interatrial block, bayés syndrome, left atrial enlargement, and atrial failure," *Journal of Clinical Medicine*, vol. 12, no. 23, p. 7331, 2023. https://doi.org/10.3390/jcm12237331

[13] D. Kitkungvan and D. H. Spodick, "Interatrial block: Is it time for more attention?" *Journal of Electrocardiology*, vol. 42, no. 6, pp. 687–692, 2009. https://doi.org/10.1016/j.jelectrocard.2009.07.016

[14] V. Ariyarajah, P. Puri, S. Apiyasawat, and D. H. Spodick, "Interatrial block: A novel risk factor for embolic stroke?" *Annals of Noninvasive Electrocardiology*, vol. 12, no. 1, pp. 15–20, 2007. https://doi.org/10.1111/j.1542-474X.2007.00133.x

[15] A. B. De Luna, M. Martínez-Sellés, A. Bayés-Genís, R. Elosua, and A. Baranchuk, "Síndrome de Bayés. Lo que todo clínico debe conocer," *Revista Española de Cardiología*, vol. 73, no. 9, pp. 758–762, 2020. https://doi.org/10.1016/j.recesp.2020.04.003

[16] L. G. Franco, L. A. Escobar Robledo, A. Bayés de Luna, and J. M. Massa, "Digitalización de Imágenes de ECG para la Detección del Síndrome de Bayés," in *XXIV Congreso Argentino de Ciencias de la Computación (La Plata)*, 2018.

[17] L. G. Franco, L. A. Escobar Robledo, A. Bayés de Luna, and J. M. Massa, "P-wave clustering methods for bayès syndrome detection," *CONAIISI*, 2020.

[18] L. G. Franco, L. A. Escobar, R. Wainschenker, A. Bayès de Luna, and J. M. Massa, "Hierarchical clustering method for bayès syndrome detection," *LACCEI*, vol. 1, no. 8, 2023. https://doi.org/10.18687/LACCEI2023.1.1.1314

[19] M. Suboh, R. Jaafar, N. Nayan, and N. Harun, "ECG-based detection and prediction models of sudden cardiac death: Current performances and new perspectives on signal processing techniques," *International Journal of Online & Biomedical Engineering*, vol. 15, no. 15, pp. 110–126, 2019. https://doi.org/10.3991/ijoe.v15i15.11688

[20] S. Hadiyoso, F. Farell, and M. D. Sulistiyo, "Image based ECG signal classification using convolutional neural network," *International Journal of Online & Biomedical Engineering*, vol. 18, no. 4, pp. 64–78, 2022. https://doi.org/10.3991/ijoe.v18i04.27923

[21] F. Gritzali, G. Frangakis, and G. Papakonstantinou, "Detection of the P and T waves in an ECG," *Computers and Biomedical Research*, vol. 22, no. 1, pp. 83–91, 1989. https://doi.org/10.1016/0010-4809(89)90017-7

[22] G. Lenis, N. Pilia, T. Oesterlein, A. Luik, C. Schmitt, and O. Dössel, "P wave detection and delineation in the ECG based on the phase free stationary wavelet transform and using intracardiac atrial electrograms as reference," *Biomedical Engineering/Biomedizinische Technik*, vol. 61, no. 1, pp. 37–56, 2016. https://doi.org/10.1515/bmt-2014-0161

[23] R. Gonzalez-Fernandez, M. Rivero-Varona, and G. M. De Oca-Colina, "Detection of P wave in electrocardiogram," in *Computing in Cardiology 2013*, 2013, pp. 515–518.

[24] H. K. Chatterjee, R. Gupta, and M. Mitra, "Real time P and T wave detection from ECG using FPGA," *Procedia Technology*, vol. 4, pp. 840–844, 2012. https://doi.org/10.1016/j.protcy.2012.05.138

[25] W. Li, "Wavelets for electrocardiogram: Overview and taxonomy," *IEEE Access*, vol. 7, pp. 25627–25649, 2018. https://doi.org/10.1109/ACCESS.2018.2877793

[26] M. H. Milon, "Comparison on fourier and wavelet transformation for an ECG signal," *American Journal of Engineering Research*, vol. 6, no. 8, pp. 1–7, 2017.

[27] S. Pitina and T. Dima, "Novel oversampling Fourier transform for ECG-waves," in *2020 International Semiconductor Conference (CAS)*, IEEE, 2020, pp. 93–96. https://doi.org/10.1109/CAS50358.2020.9268028

[28] A. Taghizabet, J. Tanha, A. Amini, and J. Mohammadzadeh, "A semi-supervised clustering approach using labeled data," *Scientia Iranica*, vol. 30, no. 1, pp. 104–115, 2023. https://doi.org/10.24200/sci.2022.58519.5772

[29] F. Y. Wattimena *et al.*, "Data mining application for the spread of endemic butterfly cenderawasih bay using the k-means clustering algorithm," *International Journal of Online & Biomedical Engineering*, vol. 19, no. 9, pp. 108–121, 2023. https://doi.org/10.3991/ijoe.v19i09.40907

[30] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Technical Report, Stanford, 2006.

[31] P. Berkhin, "Survey of clustering data mining techniques," Accrue Software. Inc. TR, San Jose, USA, 2002.

[32] H. J. Curti and R. S. Wainschenker, "FAUM: Fast autonomous unsupervised multidimensional classification," *Information Sciences*, vol. 462, pp. 182–203, 2018. https://doi.org/10.1016/j.ins.2018.06.008

[33] A. Starczewski and A. Krzyżak, "Performance evaluation of the silhouette index," in *International Conference on Artificial Intelligence and Soft Computing*, Cham: Springer International Publishing, 2015, pp. 49–58. https://doi.org/10.1007/978-3-319-19369-4_5

[34] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974. https://doi.org/10.1080/01969727408546059

[35] J. Cai, J. Hao, H. Yang, X. Zhao, and Y. Yang, "A review on semi-supervised clustering," *Information Sciences*, vol. 632, pp. 164–200, 2023. https://doi.org/10.1016/j.ins.2023.02.088

[36] S. S. Khan, S. Ahamed, M. Jannat, S. Shatabda, and D. M. Farid, "Classification by clustering (CbC): An approach of classifying big data based on similarities," in *Proceedings of International Joint Conference on Computational Intelligence: IJCCI 2018*. Singapore: Springer Nature Singapore, 2019, pp. 593–605. https://doi.org/10.1007/978-981-13-7564-4_50

[37] Z. H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018. https://doi.org/10.1093/nsr/nwx106

[38] Z. Yu *et al.*, "Adaptive ensembling of semi-supervised clustering solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1577–1590, 2017. https://doi.org/10.1109/TKDE.2017.2695615

[39] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1476–1485. https://doi.org/10.1109/ICCV.2019.00156

[40] S. Jokic, V. Delic, Z. Peric, S. Krco, and D. Sakac, "Efficient ECG modeling using polynomial functions," *Elektronika ir Elektrotechnika*, vol. 110, no. 4, pp. 121–124, 2011. https://doi.org/10.5755/j01.eee.110.4.304

[41] R. Borsali, A. Naït-Ali, and J. Lemoine, "ECG compression using an ensemble polynomial modeling: Comparison with the DCT based technique," *Cardiovascular Engineering: An International Journal*, vol. 4, no. 3, pp. 237–244, 2004. https://doi.org/10.1023/B:CARE.0000038780.96845.27

[42] R. Chaturvedi and Y. Yadav, "Analysis of ECG signal by polynomial approximation," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 2, no. 5, pp. 1029–1033, 2014.

[43] M. Karczewicz and M. Gabbouj, "ECG data compression by spline approximation," *Signal Processing*, vol. 59, no. 1, pp. 43–59, 1997. https://doi.org/10.1016/S0165-1684(97)00037-6

[44] M. Brito, J. Henriques, P. Carvalho, B. Ribeiro, and M. Antunes, "An ECG compression approach based on a segment dictionary and bezier approximations," in *2007 15th European Signal Processing Conference*, IEEE, 2007, pp. 2504–2508.

[45] R. Nygaard, D. Haugland, and J. H. Husy, "Signal compression by second order polynomials and piecewise non interpolating approximation," Department of Electrical and Computing Engineering, Tech. Rep, 1999.

[46] W. Philips and G. De Jonghe, "Data compression of ECG's by high-degree polynomial approximation," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 4, pp. 330–337, 1992. https://doi.org/10.1109/10.126605

[47] N. N. Anuar *et al.*, "Cardiovascular disease prediction from electrocardiogram by using machine learning," *International Journal of Online & Biomedical Engineering*, vol. 16, no. 7, pp. 34–48, 2020. https://doi.org/10.3991/ijoe.v16i07.13569

[48] L. Vicent *et al.*, "Baseline ECG and prognosis after transcatheter aortic valve implantation: The role of interatrial block," *Journal of the American Heart Association*, vol. 9, no. 22, p. e017624, 2020. https://doi.org/10.1161/JAHA.120.017624

[49] A. Bayés de Luna, *ECGs for Beginners*. Hoboken, NJ: John Wiley & Sons, 2014. https://doi.org/10.1002/9781118821350

[50] A. Jaeger and D. Banks, "Cluster analysis: A modern statistical review," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 15, no. 3, p. e1597, 2023. https://doi.org/10.1002/wics.1597
[51] https://sourceforge.net/projects/faum/
[52] NetPbm home page, "Netpbm," 2017. http://netpbm.sourceforge.net/

7 AUTHORS

Lorena G. Franco is a systems engineer and a PhD student in the Computational and Industrial Mathematics program at INTIA, UNCPBA. She is the Research and Development Coordinator in the Department of Information Systems Engineering at FRD, National Technological University. Her work focuses on signal-processing methods and computational techniques applied to ECG analysis for the detection of Bayés Syndrome (E-mail: lorenafranco@intia.exa.unicen.edu.ar).

Luis A. Escobar Robledo holds a Physician degree from CES University, Colombia, and is now a full-time PhD student at the CMU-PITT Computational Biology PhD Program at the University of Pittsburgh School of Medicine. Has several contributions to electrocardiography and cardiac diseases.

Rubén Wainschenker holds a PhD in Physics from the University of Buenos Aires (UBA). His research interests include signal processing and data science. He has supervised several PhD students and is a retired professor from the Department of Computer and Systems at UNCPBA, where he taught permanent courses in image processing and digital electronics.

Antoni Bayès de Luna is a distinguished Spanish cardiologist and Emeritus Professor at the Universitat Autònoma de Barcelona. He led the Cardiology Department at Fundación Investigación Cardiovascular Programa Cardiovascular-ICCC, Hospital de la Santa Creu i Sant Pau, and is internationally known for his contributions to electrocardiography, authoring influential textbooks and hundreds of scientific publications.

Hugo Curti holds a master's degree in systems engineering from UNCPBA University. His research focuses on data science methods for clustering and classification across a variety of problem domains. He is a full-time professor in the Department of Computer and Systems at UNCPBA, where he teaches courses in computer architecture and cybersecurity.

José M. Massa holds a PhD in Computational and Industrial Mathematics from UNCPBA University, Argentina. His research focuses on programming languages and medical informatics. He is a full-time professor in the Department of Computer and Systems at UNCPBA, an advisor to the High-Performance Computing Center, and a member of the Academic Committee of several doctoral programs.