ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.1314

Hierarchical Clustering Method for Bayès Syndrome Detection

Lorena G. Franco, Ing.[1] [0000-0002-7089-3313], Luis A. Escobar, Med.[2], Rubén Wainschenker[3] [0000-0002-7089-3313], Dr., Antoni Bayès de Luna, Dr. [4] [0000-0003-1676-207X], and José M. Massa, Dr.[5] [0000-0002-7456-9676]

[1] Universidad Tecnológica Nacional FRD, Argentina, francol.edu.ar@gmail.com.

[2] Fac. Medicina, Universidad CES, Colombia, lescobar9448@gmail.com.

[3, 5,1] INTIA. Fac. Cs. Exactas Universidad Nacional del Centro de Buenos Aires, Argentina, ruben.wain@gmail.com, jmassa@exa.unicen.edu.ar.

[4] Fundación Investigación Cardiovascular Programa Cardiovascular-ICCC, Institut de Recerca del Hospital de la Santa Creu I Sant Pau, IIB-Sant Pau, Barcelona, España, abayes@santpau.cat

Abstract– ***Bayès Syndrome is manifested in the cardiac cycle of an electrocardiogram. It presents associations with multiple medical conditions, being of interest in its identification at an early stage. In this article, we applied the Hierarchical Clustering method, through Matlab implementation, to identify each signal in 4 groups or categories of interest for diagnosing Bayes Syndrome. Different values were configured for the method parameter. The best value was obtained with the 'ward' option with a normalized sample concerning signal amplitude and time, achieving a total f1 Score of 0.88. The performance of K-Means++ and FAUM from previous work was compared with the results obtained from Hierarchical Clustering for a sample with normalized amplitude. The total f1 Score indicator for Hierarchical Clustering was lower than the value obtained from the two K-Means++ implementations and higher than the adjusted FAUM value***

Keywords- Bayès Syndrome, ECG, Hierarchical Tree, K Means++, FAUM.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

Hierarchical Clustering Method for Bayès Syndrome Detection

Lorena G. Franco, Ing.[1] [0000-0002-7089-3313], Luis A. Escobar, Med.[2], Rubén Wainschenker[3] [0000-0002-7089-3313], Dr., Antoni Bayès de Luna, Dr. [4] [0000-0003-1676-207X], and José M. Massa, Dr.[5] [0000-0002-7456-9676]

[1] Universidad Tecnológica Nacional FRD, Argentina, francol.edu.ar@gmail.com.

[2] Fac. Medicina, Universidad CES, Colombia, lescobar9448@gmail.com.

[3, 5,1] INTIA. Fac. Cs. Exactas Universidad Nacional del Centro de Buenos Aires, Argentina, ruben.wain@gmail.com, jmassa@exa.unicen.edu.ar.

[4] Fundación Investigación Cardiovascular Programa Cardiovascular-ICCC, Institut de Recerca del Hospital de la Santa Creu I Sant Pau, IIB-Sant Pau, Barcelona, España, abayes@santpau.cat

Abstract*– El Síndrome de Bayès se manifiesta en el ciclo cardíaco de un electrocardiograma. Presenta asociaciones con múltiples afecciones médicas, resultando de interés su identificación en una etapa temprana. En este artículo se aplicó el método de agrupamiento o clustering, Clustering Jerárquico, con la implementación de Matlab para identificar cada señal en 4 grupos o categorías de interés para el diagnóstico del Síndrome de Bayès. Se configuraron diferentes valores para el parámetro del método. El mejor valor se obtuvo con la opción 'ward' con una muestra normalizada con respecto a la amplitud y el tiempo de la señal, lográndose un f1 Score total de 0.88. Se comparó la performance de K-Means++ y FAUM de un trabajo anterior con los resultados obtenidos del Clustering Jerárquico para una muestra con la amplitud normalizada. El indicador f1 Score total del Clustering Jerárquico resultó inferior al valor obtenido de las dos implementaciones de K Means++ y superior al valor de FAUM ajustado.

Keywords*-- Síndrome de Bayès, ECG, Árbol Jerárquico, K-Means++, FAUM.

I. INTRODUCCIÓN

En el contexto de un proyecto de investigación sobre la aplicación de técnicas computacionales de análisis de señales cardiológicas, en este trabajo se presentan los resultados de la aplicación del método de *clustering Clustering Jerárquico* utilizado con una muestra normalizada respecto de la amplitud de la señal y con la misma muestra normalizada en amplitud y tiempo. También se realizó una comparación con los métodos de agrupamiento *K-Means++* (implementación en *Matlab* y *FAUM*) y *FAUM en modo ajustado*[1]. Estos métodos se utilizaron para clasificar la morfología de la onda P del electrocardiograma (ECG). Se aplicaron de forma semi-supervisada y predictiva para posteriormente evaluar su aplicación en la detección del Síndrome de Bayès. Este Síndrome ha sido estudiado en las últimas décadas por quien le da el nombre, el Dr Antonio Bayès de Luna [2-4].

En los últimos años se ha asociado al Síndrome de Bayès a diferentes patologías, no sólo del Sistema circulatorio. El concepto de Bloqueo Interauricular (BIA) es el más frecuente y relevante a nivel auricular. Se dividió el BIA de la misma manera que a nivel ventricular, sinoauricular y auriculoventricular en primer grado o parcial, tercer grado o avanzado y segundo grado o intermitente [3, 5-9].

Bayès de Luna et al. [2] analizaron ECGs, demostrando una prevalencia de BIA avanzado del 1%, mientras que cuando se seleccionó solo a los pacientes con cardiopatía estructural la prevalencia fue del 2%.

En función de los trabajos de investigación consultados es importante mencionar que se ha asociado el BIA con alteraciones médicas como fibrilación auricular (FA), isquemia miocárdica, agrandamiento de la aurícula izquierda y émbolos sistémicos [10]. El BIA es considerado como un factor de riesgo para accidente cerebrovascular cardioembólico [11]. En la mediana edad, el BIA avanzado triplica el riesgo de FA y casi duplica el riesgo de ictus. La duración de la onda P también se asocia con mortalidad cardiovascular y muerte súbita cardíaca. En edades muy avanzadas la presencia de BIA también se asocia con la mortalidad total. Hay estudios que demuestran que la prevalencia de demencia se incrementaba progresivamente al pasar de onda P normal a BIA parcial, BIA avanzado y FA [12]. Teniendo en cuenta lo expuesto resulta de interés su reconocimiento en una etapa temprana. El diagnóstico del bloqueo parcial o avanzado puede realizarse analizando el ECG.

Se han desarrollado muchos métodos y herramientas para el análisis de ECG [13-14] como la detección de la onda P [15-18] en la bibliografía actual, pero no existen métodos que identifiquen el BIA. En cuanto al análisis de las ondas del ciclo cardíaco, un número importante de métodos aplican técnicas basadas en el análisis de frecuencia como Wavelets y Fourier, entre otros. Una alternativa a las múltiples opciones existentes consiste en explorar el problema desde el punto de vista de la clasificación. Las técnicas para estos problemas han registrado una importante mejora en su eficacia y eficiencia en los últimos años, impulsadas por los problemas clasificados como Big Data [19]. En función del contexto descripto, el interés se centró en técnicas de agrupamiento de forma semi-supervisada en las que una muestra de cada clase se etiqueta manualmente.

El avance de la tecnología resulta relevante para la industria biomédica. Los nuevos electrocardiógrafos generan los resultados del ECG en más de un formato. Sin embargo, esta ventaja no suele ser aprovechada. Muy pocos centros y especialistas en Cardiología almacenan los electrocardiogramas en formato digital. Frente a este contexto

y debido a que los ECG disponibles son resultado del seguimiento a lo largo de los años de pacientes que presentaron BIA, se procesaron electrocardiogramas que se encuentran en soporte papel y por lo tanto, fue necesaria su digitalización. La misma se realizó teniendo en cuenta que se debe preservar principalmente la onda P. En trabajos anteriores [20] los autores exploraron técnicas de digitalización y segmentación orientadas a preservar esta onda. También se aplicaron métodos de *clustering* [21].

En el agrupamiento jerárquico se busca construir la relación jerárquica entre los datos que se deben agrupar. Si cada punto de datos representa un grupo individual al principio y, luego, los dos grupos más vecinos se fusionan en un nuevo grupo hasta que solo queda un grupo [22]. Este método de *clustering* proporciona dendrogramas que visualizan las relaciones jerárquicas entre los clústeres, y los mapas autoorganizados brindan una cuadrícula bidimensional que visualiza algunas topologías específicas dentro del conjunto de datos. En el mejor de los casos, estos métodos descubren los grupos en los que un conjunto determinado de datos se agrupan cuando existen relaciones jerárquicas naturales.

Un tema de interés, dentro del *Clustering Jerárquico*, actualmente es el análisis de las diferentes configuraciones en términos de eficiencia espacial, temporal y desempeño junto con aspectos de robustez, tal como se puede ver en [23-26].

En este trabajo se propone aplicar técnicas para agrupar ondas P en diferentes grupos correspondientes a morfologías utilizadas en el diagnóstico del BIA: Onda P normal, Onda P bimodal (bloqueo de primer grado), Onda P con morfología negativa y por último onda P (bifásica) o ± (bloqueo de tercer grado). Se decidió aplicar el método *Clustering Jerárquico* para el conjunto de señales disponibles. En el método se probarán diferentes valores para el parámetro *linkage*. Los resultados más relevantes se utilizaran para compararlos con los valores obtenidos de otros métodos de *clustering* aplicados al mismo conjunto de datos. Al igual que en trabajos anteriores [21, 27-28], se evaluó la técnica de *clustering* no con los indicadores de desempeño propios de los métodos de agrupamiento tal como el índice de Dunn o el índice Silohuette [29]. Para analizar los resultados obtenidos se utilizarán los indicadores de análisis de la matriz de confusión: *Precision*, *Accuracy*, *Recall* y *f1 Score*.

A continuación, en la sección II se presentará una síntesis de los materiales y el método de agrupamiento propuesto. En la sección III se muestran los resultados obtenidos. Por último en la sección IV se presentan las conclusiones elaboradas y trabajos futuros.

II. Materiales y Métodos

En esta sección se presentan las características de los ECGs utilizados y los métodos de agrupamiento aplicados.

A. Materiales

Se utilizaron 49 muestras de un total de 600 ECG procedentes de las investigaciones del Dr. Bayès y su grupo de trabajo, en el contexto de un proyecto conjunto de colaboración. Estas muestras consisten en imágenes de ECG en papel, las cuales fueron escaneadas a una resolución suficiente que permitió digitalizar los 138 niveles de amplitud y la duración de la onda. Luego a dichas muestras se les aplicó un proceso de digitalización basado en el trabajo previo [20] para preservar la onda P. En la Fig. 1 se observan ejemplos de ondas P.

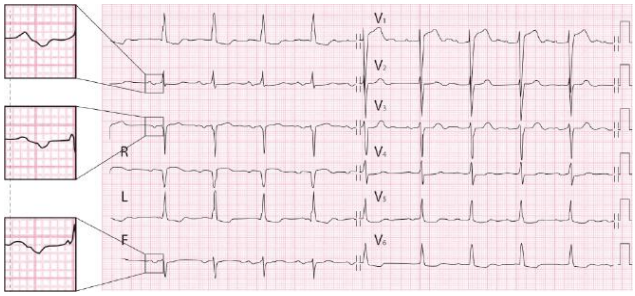

Fig. 1 ECG con BIA Avanzado.

Las imágenes obtenidas mediante el proceso de digitalización constituyen los materiales de este trabajo. Se aplicó una binarización y un umbralado. Con la intención de obtener una curva del menor espesor posible, y por lo tanto mejorar la precisión. A las imágenes obtenidas con el umbralado, se le aplicó una esqueletización por medio de una erosión múltiple iterativa. Este proceso se ilustra en las Fig. 2a., 2b., 2c. y 2d. Con el fin de establecer una referencia para los valores positivos y negativos de la onda, se aplicó un método basado en la técnica manual utilizada por los médicos que trabajan en este tema [30]. Por último se obtuvo una lista de valores de intensidad para cada columna de la imagen, correspondiente a cada elemento de muestreo temporal del ECG. Esta lista de valores se calculó inicialmente con una alta precisión en punto flotante y luego se normalizó entre -1 y1.

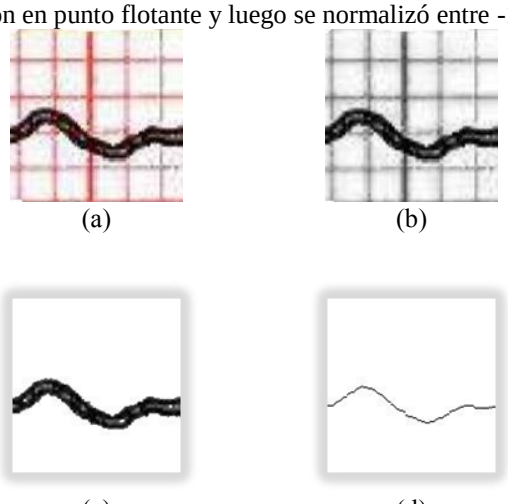

Fig. 2 Digitalización de imagen. (a) Imagen original. (b) Binarizado. (c) Umbralización. (d) Esqueletización.

B. Métodos

En esta sección se presentan los detalles relacionados a la aplicación del método de agrupamiento.

Para poder aplicar el método de *Clustering Jerárquico*, fue necesario disponer de un conjunto de datos provenientes de las 49 muestras. Estos datos consisten en un conjunto de características de cada muestra que representan la amplitud de la señal en cada uno de los tiempos de medición. Resulta relevante mencionar que este método de agrupamiento trabaja bajo la suposición que las características definidas son independientes y ortogonales entre sí. Debido a que las características provienen del proceso de medir la actividad eléctrica del corazón de un paciente, las mismas poseen una fuerte dependencia temporal entre ellas. Esto significa que la suposición de independencia y ortogonalidad de las características que es deseable en la aplicación de métodos de Machine Learning y en particular en los métodos de *clustering* [31-33], no se cumple en este caso. Sin embargo, lo que sucede en el problema planteado en este trabajo, es que si bien el comportamiento en este caso de la onda P tiene una cierta predecibilidad en cuanto a lo que se espera que sea una onda P, existen fenómenos fisiológicos, de captura de señal y del mismo Síndrome de Bayès que alteran la forma de la onda. Es por esto que se decidió trabajar con los valores de amplitud temporales como si fuesen independientes para los algoritmos, aunque no lo sean realmente.

El resultado del conjunto, obtenido a partir de las características de la señal, es entonces una matriz de valores, cuyas filas contienen las 49 muestras y las columnas los 138 valores de amplitud correspondientes a cada lapso muestreado. Estas columnas están formadas con las características correspondientes a cada muestra. Se trabajó con dos tipos de normalizaciones.

Inicialmente se normalizó la amplitud de la señal del conjunto de muestras. El conjunto quedó conformado por valores de punto flotante entre -1 y 1 que representan la amplitud correspondiente a la medición. Posteriormente se obtuvo un conjunto de muestras que se normalizó con valores de punto flotante entre -1 y 1 tanto para amplitud como para el tiempo.

En este método se utilizó primero como entrada el conjunto normalizado en amplitud. Se trabajó con un valor de K igual a 4 clases, teniendo en cuenta las morfologías de interés de la onda P. Esta aplicación se realizó utilizando la implementación del *Clustering Jerárquico* de la herramienta *Matlab*. Este método implica una técnica estadística donde los grupos se crean secuencialmente mediante la fusión sistemática de *clusters* similares, en función de las medidas de distancia y *linkage* elegidas.

La función distancia utilizada fue la distancia Euclidiana, debido a que las características generales de la señal hacen que sea más beneficiosa que otras funciones distancia que podrían penalizar pequeñas diferencias en el valor de una característica de dos muestras. Es necesario aclarar que en general, la función de distancia Euclidiana no es adecuada para espacios multidimensionales, pero aun así se justifica su uso en este contexto.

Dado que el objetivo final es encontrar los cluster que son más similares, es importante encontrar el *linkage* entre dos *clusters* que tienen más de un caso o puntos de datos. Actualmente hay varias técnicas para encontrar ese vínculo entre grupos. Si una configuración del parámetro *linkage* puede funcionar bien para un tipo particular de un conjunto de datos, es posible que no funcione bien en otro tipo de conjunto de datos [24]. Teniendo en cuenta lo anteriormente expresado se decidió aplicar las siguientes configuraciones para el parámetro *linkage*: *average*, *centroid*, *complete*, *median*, *single*, *weighted* y *ward*; para analizar los resultados obtenidos.

El método también se utilizó con un conjunto de muestras normalizadas en amplitud y tiempo. Se trabajó con un valor de K igual a 4 clases y al igual que en el caso anterior se utilizó una distancia euclidiana. Se aplicaron las mismas configuraciones para el parámetro *linkage*: *average*, *centroid*, *complete*, *median*, *single*, *weighted* y *ward*.

III. RESULTADOS

En esta sección se presentan los resultados más relevantes de las pruebas que se ejecutaron con *Clustering Jerárquico*. El algoritmo de agrupamiento se aplicó para diferentes valores para el parámetro *linkage*. En el ámbito de esta investigación el valor de k que se utilizó fue de 4.

El método de *Clustering Jerárquico* se aplicó para las 49 muestras en un primer momento normalizadas en amplitud, y a continuación normalizadas en amplitud y tiempo.

Además se presentarán los valores de la matriz de confusión e indicadores obtenidos de aplicar los algoritmos de agrupamiento *K-Means++* (dos implementaciones) y *FAUM* ajustado con la misma muestra. En este caso se trabajó con la muestra normalizada en amplitud. Se utilizarán estos resultados obtenidos en un trabajo de investigación previo [20] para analizarlos en conjunto con los resultados obtenidos de aplicar *Clustering Jerárquico*.

En la Tabla I puede observarse como se encuentra conformada una de las muestras utilizadas en las pruebas. La cantidad de ondas P bifásica de interés se corresponden con lo indicado en [2], considerando que se pueden encontrar en 3 derivaciones de un ECG y teniendo en cuenta que no siempre fue posible extraer más de una onda P por derivación.

TABLA I
MUESTRA RELEVANTE

Morfología de la onda P	Cantidad de Muestras
Onda P bifásica	23
Onda P bimodal	3
Onda P negativa	8
Onda P normal	15
Total	49

Se presenta un total de 49 muestras distribuidas según su morfología.

En la Tabla II se presentan los resultados de la matriz de confusión especificada para cada tipo de morfología de la onda P y los datos correspondientes de aplicar *Clustering Jerárquico* para los 4 tipos de configuraciones del parámetro *linkage* que obtuvieron el mejor resultado: *average*, *centroid*, *median* y *ward*. Además también se obtuvieron los datos para los siguientes *linkage*: *complete*, *single* y *weighted*. El valor que se utilizó de k fue de 4 en todos los casos. Los datos de las muestras se encontraban normalizadas en amplitud.

TABLA II
MATRIZ DE CONFUSIÓN.

	TP	TN	FP	FN	N
Bifásica	23				
CJ Average o centroid	8	25	1	15	49
CJ Median	9	25	1	14	49
CJ Ward	9	25	1	14	49
Bimodal	3				Total
CJ Average o centroid	0	45	1	3	49
CJ Median	0	44	2	3	49
CJ Ward	0	43	3	3	49
Negativa	8				Total
CJ Average o centroid	7	31	10	1	49
CJ Median	6	30	11	2	49
CJ Ward	8	30	11	0	49
Positiva	15				Total
CJ Average o centroid	15	29	5	0	49
CJ Median	15	29	5	0	49
CJ Ward	12	29	5	3	49

En la Tabla III se aprecian los valores de *Accuracy*, *Precision*, *Recall* y *f1 Score* [34] para la muestra indicada en la Tabla I.

TABLA III
INDICADORES

	Acc	Prec	Rec	f1 Score
Bifásica				
CJ Average o centroid	0.67	0.89	0.35	0.50
CJ Median	0.69	0.90	0.39	0.55
CJ Ward	0.69	0.90	0.39	0.55
Bimodal				
CJ Average o centroid	0.92	0.00	0.00	0.00
CJ Median	0.90	0.00	0.00	0.00
CJ Ward	0.88	0.00	0.00	0.00
Negativa				
CJ Average o centroid	0.78	0.41	0.88	0.56
CJ Median	0.73	0.35	0.75	0.48
CJ Ward	0.78	0.42	1.00	0.59
Positiva				
CJ Average o centroid	0.90	0.75	1.00	0.86
CJ Median	0.90	0.75	1.00	0.86
CJ Ward	0.84	0.71	0.80	0.75

En la Tabla IV puede observarse un resumen de los indicadores totales obtenidos sobre la matriz de confusión, ponderando cada uno de ellos por la cantidad de muestras de cada clase.

TABLA IV
INDICADORES TOTALES

	Acc Total	Prec Total	Rec Total	f1 Score Total
CJ Average o centroid	0.77	0.71	0.61	0.59
CJ Median	0.78	0.71	0.61	0.60
CJ Ward	0.76	0.71	0.59	0.58

En la Tabla V se presentan los resultados de la matriz de confusión especificada para cada tipo de morfología de la onda P y los datos correspondientes de aplicar *Clustering Jerárquico* para los 4 tipos de configuraciones del parámetro *linkage* que obtuvieron el mejor resultado: *average*, *centroid*, *complete* y *ward*. Se obtuvieron también los datos para los siguientes *linkage*: *median*, *single* y *weighted*. El valor que se utilizó de k fue de 4 en todos los casos. Los datos de las muestras normalizadas a diferencia de los datos presentados en la Tabla II se normalizaron en amplitud y tiempo.

TABLA V
MATRIZ DE CONFUSIÓN.

	TP	TN	FP	FN	N
Bifásica	23				
CJ Average o centroid	23	23	3	0	49
CJ Complete	17	24	2	6	49
CJ Ward	21	23	3	2	49
Bimodal	3				Total
CJ Average o centroid	0	45	1	3	49
CJ Complete	0	35	11	3	49
CJ Ward	3	45	1	0	49
Negativa	8				Total
CJ Average o centroid	6	41	0	2	49
CJ Complete	7	41	0	1	49
CJ Ward	7	41	0	1	49
Positiva	15				Total
CJ Average o centroid	13	31	1	2	49
CJ Complete	9	31	3	6	49
CJ Ward	12	32	2	3	49

En la Tabla VI se aprecian los valores de *Accuracy*, *Precision*, *Recall* y *f1 Score* para la muestra indicada en la Tabla V.

En la Tabla VII puede observarse un resumen de los indicadores totales obtenidos sobre la matriz de confusión de la Tabla V, ponderando cada uno de ellos por la cantidad de muestras de cada clase.

En la Tabla VIII se presentarán los valores de la matriz de confusión de aplicar los algoritmos de agrupamiento *K-Means++* (implementaciones diferentes: *Matlab* y *FAUM*) y *FAUM ajustado* con la muestra de 49 señales normalizada en amplitud.

TABLA VI
INDICADORES

	Acc	Prec	Rec	f1 Score
Bifásica				
CJ Average o centroid	0.94	0.88	1.00	0.94
CJ Complete	0.84	0.89	0.74	0.81
CJ Ward	0.90	0.88	0.91	0.89
Bimodal				
CJ Average o centroid	0.92	0.00	0.00	0.00
CJ Complete	0.71	0.00	0.00	0.00
CJ Ward	0.98	0.75	1.00	0.86
Negativa				
CJ Average o centroid	0.96	1.00	0.75	0.86
CJ Complete	0.98	1.00	0.88	0.93
CJ Ward	0.98	1.00	0.88	0.93
Positiva				
CJ Average o centroid	0.90	0.93	0.87	0.90
CJ Complete	0.82	0.75	0.60	0.67
CJ Ward	0.90	0.86	0.80	0.83

TABLA VII
INDICADORES TOTALES

	Acc Total	Prec Total	Rec Total	f1 Score Total
CJ Average o centroid	0.93	0.86	0.86	0.86
CJ Complete	0.85	0.81	0.67	0.74
CJ Ward	0.92	0.88	0.88	0.88

TABLA VIII
MATRIZ DE CONFUSIÓN.

	TP	TN	FP	FN	N
Bifásica	23				Total
K-Means++ Matlab	18	24	2	5	49
K-Means++ FAUM	22	23	3	1	49
FAUM ajustado	23	12	14	0	49
Bimodal	3				Total
K-Means++ Matlab	3	38	8	0	49
K-Means++ FAUM	3	45	1	0	49
FAUM ajustado	3	44	2	0	49
Negativa	8				Total
K-Means++ Matlab	7	41	0	1	49
K-Means++ FAUM	7	41	0	1	49
FAUM ajustado	6	41	0	2	49
Positiva	15				Total
K-Means++ Matlab	6	29	5	9	49
K-Means++ FAUM	12	33	1	3	49
FAUM ajustado	0	33	1	15	49

En la Tabla IX se aprecian los valores de *Accuracy*, *Precision*, *Recall* y *f1 Score* para la muestra indicada en la Tabla VIII.

TABLA IX
INDICADORES

	Acc	Prec	Rec	f1 Score
Bifásica				
K-Means++ Matlab	0.86	0.90	0.78	0.84
K-Means++ FAUM	0.92	0.88	0.96	0.92
FAUM ajustado	0.71	0.62	1.00	0.77
Bimodal				
K-Means++ Matlab	0.84	0.27	1.00	0.43
K-Means++ FAUM	0.98	0.75	1.00	0.86
FAUM ajustado	0.96	0.60	1.00	0.75
Negativa				
K-Means++ Matlab	0.98	1.00	0.88	0.93
K-Means++ FAUM	0.98	1.00	0.88	0.93
FAUM ajustado	0.96	1.00	0.75	0.86
Positiva				
K-Means++ Matlab	0.71	0.55	0.40	0.46
K-Means++ FAUM	0.92	0.92	0.80	0.86
FAUM ajustado	0.67	0.00	0.00	0.00

En la Tabla X puede observarse un resumen de los indicadores totales obtenidos sobre la matriz de confusión de la Tabla VIII, ponderando cada uno de ellos por la cantidad de muestras de cada clase.

TABLA X
INDICADORES TOTALES

	Acc Total	Prec Total	Rec Total	f1 Score Total
K-means++ Matlab	0.83	0.77	0.69	0.71
K-means++ FAUM	0.93	0.90	0.90	0.90
FAUM ajustado	0.76	0.49	0.65	0.54

IV. CONCLUSIONES Y TRABAJO FUTURO

Teniendo en cuenta los valores de los indicadores obtenidos, al aplicar el método de *Clustering Jerárquico* para la muestra normalizada en amplitud, los mejores resultados se obtienen al agrupar las ondas positivas de la muestra, logrando un *f1 Score* de 0.86. En el caso de agrupar las ondas de 3er grado (bifásica) no se consigue un buen desempeño, al tener una morfología ± a veces asigna una señal al grupo de las señales positivas y a veces al grupo de las negativas. En el caso de 1er grado (bimodal) al tener muy pocas muestras disponibles no se logra agrupar.

Respecto de los valores de configuración del parámetro *linkage* utilizados, el mejor resultado, se logró con la opción *median*, con 0.6 en el *f1 Score* de los indicadores totales. Valores muy similares aunque inferiores se obtuvieron con las opciones *average* o *centroid* y *ward*. Mientras que las opciones que tuvieron el peor desempeño fueron: *complete*, *single* y *weighted*. De estas opciones el *single*, que si no se especifica al implementar en *Matlab* el método de *Clustering Jerárquico* lo utiliza por defecto, obtuvo el peor rendimiento.

Considerando los valores de los indicadores obtenidos, al aplicar el método de *Clustering Jerárquico* para la muestra normalizada en amplitud y tiempo, los resultados son muy buenos, con un *f1 Score* en general (a excepción de un caso) mayor a 0.8 para las ondas bifásica, negativa y positiva. En el caso de 1er grado (bimodal) con una de las configuraciones de *linkage* se consigue obtener a pesar de las pocas muestras un *f1 Score* de 0.86, mientras que con las demás opciones las señales no son correctamente agrupadas.

En cuanto a las configuraciones de *linkage* utilizadas, el mejor resultado se logró con la opción *ward*, con 0.88 en el *f1 Score* de los indicadores totales. Continúa con un buen rendimiento aunque inferior la opción de *linkage average* o *centroid* con un valor de 0.86 para el valor de *f1 Score* total. Mientras que la opción *complete* permite obtener un *f1 Score* total de 0.74. Las opciones que tuvieron el peor desempeño fueron: *median*, *single* y *weighted*. Al igual que para el caso de las muestras normalizadas en amplitud, la opción *single*, que el método de *Clustering Jerárquico* lo utiliza por defecto, obtuvo el peor rendimiento.

En este trabajo se decidió comparar los valores de los indicadores obtenidos al aplicar el método de *Clustering Jerárquico* para las 49 muestras normalizadas en amplitud con los indicadores obtenidos de aplicar los métodos *K-Means++* (implementaciones diferentes: *Matlab* y *FAUM*) y *FAUM* ajustado que fueron obtenidos en un trabajo previo [21] para el mismo conjunto de datos.

Al comparar el mejor *f1 Score* obtenido con el *Clustering Jerárquico* correspondiente a las señales positivas se logró un valor de 0.86 superando al valor de *FAUM* ajustado y *K-Means++* implementación de *Matlab* e igualando al mejor valor obtenido con *FAUM ajustado*. En el resto de las morfologías fue superado el *f1 Score* por *K-Means++* y *FAUM.* Si se tiene en cuenta el valor de 0.6 del *f1 Score* total del *Clustering Jerárquico*, es inferior al valor obtenido de las dos implementaciones de *K-Means++* y superior al valor de *FAUM ajustado*.

A modo de análisis de interpretabilidad de resultados, una hipótesis acerca del motivo de la mejora en el desempeño lograda con la configuración *ward* respecto del resto de las configuraciones del parámetro *linkage* para el experimento donde las muestras están normalizadas en amplitud y tiempo, puede deberse a que la configuración *ward* considera la asignación de una muestra a un *cluster* según la minimización de la varianza del *cluster* y no según la distancia al *centroide* o al elemento más cercano del *cluster*. Considerando que las muestras se agrupan en nubes de puntos que no poseen una morfología regular, la configuración *ward* logra un agrupamiento que conserva la morfología natural de cada *cluster*.

Respecto a la aplicabilidad de los resultados obtenidos, el presente trabajo es una contribución para lograr disponer en un futuro de una herramienta que sea capaz de poder integrarse con los sistemas de adquisición y almacenamiento de ECGs en una Institución médica con el fin de identificar automáticamente grupos de pacientes que requieran una consulta adicional con un médico en relación a una potencial afección relacionada al Síndrome de Bayès.

En trabajos futuros se planea i) aplicar *K-Means++* y *FAUM* al conjunto de 49 muestras normalizadas con respecto a la amplitud y el tiempo para poder comparar los indicadores obtenidos en este trabajo, ii) *FAUM ajustado* modificando sus parámetros de entropía y cardinalidad, iii) Utilizar centroides de referencia para inicializar los *clusters* de forma predefinida con la idea de utilizar estos algoritmos u otros como clasificadores, iv) utilizar descriptores polinómicos para señales.

REFERENCES

[1] Curti, H. J., and Wainschenker, R. S. "FAUM: Fast Autonomous Unsupervised Multidimensional classification". Information Sciences, 2018, vol. 462, pp. 182-203.

[2] Bayés de Luna. A., et al. "Electrocardiographic and vectorcardiographic study of interatrial conduction disturbances with left atrial retrograde activation". Journal of electrocardiology, 1985, vol. 18, no 1, pp. 1-13.

[3] Bayés de Luna. A., et al. "Interatrial conduction block and retrograde activation of the left atrium and paroxysmal supraventricular tachyarrhythmia". European heart journal. 1988, vol. 9, no 10, pp. 1112-111

[4] Bacharova. L., and Wagner. G. S. "The time for naming the Interatrial Block Syndrome: Bayes Syndrome". Journal of Electrocardiology. 2014, vol. 48, no 2, pp. 133-134.

[5] Bayés de Luna. A. "Bloqueo a Nivel Auricular". Rev Esp Cardiol. 1979, vol. 32, no 1, pp. 5-10.

[6] Conde. D., and Baranchuk. A. "What a Cardiologist must know about Bayes' Syndrome". Revista Argentina de Cardiología. 2014, vol. 82, no 3, pp. 237-239.

[7] Conde. D., and Baranchuk. A. "Bloqueo interauricular como sustrato anatómico-eléctrico de arritmias supraventriculares: síndrome de Bayés". Archivos de cardiología de México, 2014, vol. 84, no 1, pp. 32-40.

[8] Bayés de Luna. A., Baranchuk. A., Robledo. L. A. E., van Roessel. A. M., and Martínez-Sellés. M. "Diagnosis of interatrial block". Journal of geriatric cardiology: JGC, 2017, vol. 14, no 3, pp. 161.

[9] Baranchuk. A., Torner. P., and Bayés de Luna. A. "Bayés Syndrome What Is It?". Circulation. 2018, vol. 137, no 2, pp.200-202.

[10] Kitkungvan. D., and Spodick. D. H. "Interatrial block: is it time for more attention?". Journal of electrocardiology, 2009, vol. 42, no 6, pp. 687-692.

[11] Ariyarajah. V., Puri. P., Apiyasawat. S., and Spodick. D. H. "Interatrial block: A novel risk factor for embolic stroke?".Annals of Noninvasive Electrocardiology, 2007, vol. 12, no 1, pp. 15-20.

[12] de Luna. A. B., Martínez-Sellés. M., Bayés-Genís. A., Elosua. R., and Baranchuk. A. "Síndrome de Bayés. Lo que todo clínico debe conocer". Revista Española de Cardiología. 2020, vol. 73, no 9, p. 758-762.

[13] Bailey. J. J., Berson. A. S., Garson Jr. A., Horan. L. G., Macfarlane. P. W., Mortara. D. W., and Zywietz. C. "Recommendations for Standardization and Specifications in Automated Electrocardiography: Bandwidth and Digital Signal Processing. A report for health professionals by an ad hoc writing group of the Committee on Electrocardiography and Cardiac Electrophysiology of the Council on Clinical Cardiology. American Heart Association.". Circulation. 1990, vol. 81, no 2, pp. 730-739.

[14] Yochum. M., Renaud. C., and Jacquir. S. "Automatic detection of P. QRS and T patterns in 12 leads ECG signal based on CWT". Biomedical Signal Processing and Control, 2016, vol. 25, pp. 46-52.

[15] Gritzali. F., Frangakis. G., and Papakonstantinou. G. “Detection of the P and T-waves in an ECG”. Computers and Biomedical Research. 1989, vol. 22, no 1, pp. 83-91.
[16] Lenis. G., Pilia. N., Oesterlein. T., Luik. A., Schmitt. C., and Dössel. O. “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, 2016, vol. 61, no 1, pp. 37-56.
[17] Gonzalez-Fernandez. R., Rivero-Varona. M., and de Oca-Colina. G. M. “Detection of P wave in electrocardiogram”. En Computing in Cardiology 2013, IEEE. 2013, pp. 515-518.
[18] Chatterjee. H. K., Gupta. R., andMitra. M. “Real time P and T wave detection from ECG using FPGA”. Procedia Technology, 2012, vol. 4, pp. 840-844.
[19] Ismail. A., Shehab. A., and El-Henawy. I. M. “Healthcare Analysis in Smart Big Data Analytics: Reviews. Challenges and Recommendations”. En Security in Smart Cities:Models, Applications. and Challenges. Springer, Cham. 2019, pp. 27-45.
[20] Franco. L. G., Escobar Robledo. L. A., Bayés de Luna. A., and Massa. J. M. “Digitalización de Imágenes de ECG para la Detección del Síndrome de Bayés”. En XXIV Congreso Argentino de Ciencias de la Computación. La Plata, 2018.
[21] Franco. L. G., Escobar Robledo. L. A., Bayés de Luna. A, and Massa. J. M: "P-Wave Clustering Methods for Bayès Syndrome Detection". CONAIISI, 2020.
[22] XU, D. and TIAN, Y. “A comprehensive survey of clustering algorithms”. Annals of Data Science, 2015, vol. 2, p. 165-193.
[23] Ackerman, M., and Ben-David, S. “A characterization of linkage-based hierarchical clustering”. The Journal of Machine Learning Research, 2016, vol. 17, no 1, p. 8182-8198.
[24] Jarman, A. M. “Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method”. Georgia Southern University, 2020.
[25] Charikar, M., Chatziafratis, V., and Niazadeh, R. “Hierarchical clustering better than average-linkage”. En Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2019. p. 2291-2304.
[26] Yim, O., and Ramdeen, K. T. “Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data”. The quantitative methods for psychology, 2015, vol. 11, no 1, p. 8-21
[27] Franco, L. G., Escobar, L. A., de Luna, A. B., and Massa, J.M. “Clustering of derivative and integrative p-wave features for Bayès Syndrome detection”. In 17th International Symposium on Medical Information Processing and Analysis, SPIE, 2021. Vol. 12088, pp. 410-421.
[28] Franco, L. G., Escobar, L. A., de Luna, A. B., and Massa, J.M. “Augmentation and Clustering of P-Wave for Bayès Syndrome Detection”. CONAIISI, 2022.
[29] Kumar, V., Chhabra, J. K., and Kumar, D. “Performance evaluation of distance metrics in the clustering algorithms”. INFOCOMP Journal of Computer Science, 2014, vol. 13, no 1, p. 38-52.
[30] Bayés de Luna. A. “ECGs for beginners”. John Wiley & Sons, 2014.
[31] Murtagh. F., and Contreras. P. “Methods of Hierarchical Clustering”. CoRR. abs/1105,0121, 2011.
[32] Aggarwal. C. C., and Reddy. C. K. “Data Clustering: Algorithms and Applications”. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra, 2014.
[33] Domingos. P., and Pazzani. M. “Beyond independence: Conditions for the optimality of the simple Bayesian classifier”. En Proc. 13th Intl. Conf. Machine Learning, 1996, pp. 105-112.
[34] Powers, D. M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061, 2020.