Sex Differences in the ECG Interpretation: a Functional Data Analysis Perspective

Vera Maioli [1], Letizia Clementi[1,2], Marco D. Santambrogio[1]
[1]Politecnico di Milano, Milan, Italy,
[2] CADS, Center for Analysis Decisions and Society, Human Technopole, Milan, Italy
vera.maioli@mail.polimi.it
{letizia.clementi, marco.santambrogio}@polimi.it

***Abstract*— Prompt diagnosis and correct therapy are essential for the treatment of cardiovascular diseases. A proper evaluation of electrocardiography (ECG) tracings could help to analyze the different effects of drug therapies in men and women and therefore would allow the prescription of more specific drugs and dosages. For this purpose, it is essential to know the morphology of the ECG and to know how to read it correctly. In recent years there has been an increasing awareness of the differences between a male and female ECG trace and, therefore, of the need to take them into account to obtain a correct diagnosis. This work proposes a method for the recognition of the subject's sex starting from the morphology of the ECG trace alone to highlight sex differences.**

This technique employs Functional Data Analysis (FDA), a statistical approach specifically developed for the analysis of curves and surfaces. We proved the adequacy of our method by evaluating its ability to classify the traces correctly and is proposed as a smart analysis tool to be employed in the development of *ad hoc* cardiac drugs. The procedure foresees a preprocessing of the signal through a Butterworth filtering, wavelet-based smoothing, and alignment of the traces. We then classify the signals through a cluster analysis in the form of multivariate functional k-mean procedures. The result is a semi-automatic assignment of the sex of the subject to which the ECG belongs. The method reaches better performances when considering younger subjects because morphological differences between sexes are more evident in this subpopulation, as previously highlighted in literature. More specifically, the accuracy is 77.8% in the younger population and 71.4% in elderly subjects. The technique hereby proposed is a valuable tool for the exploration of ECG tracings to be employed in clinical and pharmacological research.

***Index Terms*— ECG, FDA, gender medicine, cluster analysis**

I. Introduction

Cardiovascular diseases are one of the leading cause of death worldwide [1]. For this reason, correct evaluation and treatment of risk factors are essential. In this context, the evaluation of the patient's health or pathological state is done through the analysis of the ECG traces. Thus it is necessary to be aware of the characteristics of the ECG, in order to be able to recognize a potential pathology.

It has been known for several years that the risk of heart disease is not the same in men and women and that the response to therapies also varies with sex [2]. For this reason the study of sex differences is fundamental to allow adequate prevention and treatment. In the clinical field, major differences between the two sexes were reported during the early adulthood, while they become minor with advancing years: in men under the age of 50 the percentage of patients is higher, but in the case of disease the prognosis for women is worse. In particular the percentage of death related to cardiovascular disease is significantly higher in women compared to men (51% vs 42%), nonetheless the risk of experiencing the first cardiovascular complication is lower and it emerges in the female population ten years later than in the male one [2] [3].

These differences also influence the pharmacodynamics of drugs when administerd to men and women. In fact, the latter often present adverse drug reactions, in the form of torsades de pointes (QT interval prolongation in ECG trace), skin diseases and bleeding complications. This may be due to the fact that most clinical trials enrol mainly middle-aged men and, therefore, drug therapy is often inadequate, or even adverse, for women [4].

The analysis of the ECG is a valuable diagnostic tool that could be even more effective if sex differences were correctly considered. The diagnostic process is based on the identification of reference points, and the comparison of associated parameters, such as wave amplitude and intervals between successive waves, with physiological values. Previous studies in literature have analyzed the differences between such parameters in male and female ECG tracings [3]. The most relevant findings are summarized in Table I.

TABLE I:

Sex differences in ECG parameters:mean values and the standard deviation in healthy female subjects and healthy male subjects [5] [6] [7].

	Male	**Female**
Heart rate [bpm]	70	74
PR interval [msec]	167.32 ± 16.6	155.77 ± 20.5
QRS duration [msec]	90.15 ± 13.7	67.87 ± 9.5
QT interval [msec]	405.47 ± 12.10	425.87 ± 14.40

The complete assessment of these differences is far to be reached and the reasons accounting for them remain largely unknown [8]. Further studies assessing sex differences in the ECG would allow better diagnoses, and also more efficacious therapies. Bridging the care gap between men and women would improve both the analysis and the treatment of many women affected by cardiovascular diseases. The purpose of this study thus is to propose a smart tool for the analysis and classification of ECG traces.

II. Methods

This Section will describe the dataset used in the present work, and the analysis performed, which include the pre-processing of the different traces, the synchronization and the perform of cluster analysis. All the steps were implemented employing the SciPy [9] and the Scikit-fda [10] Python libraries.

A. Data

In this work we analysed 34 ECG traces, taken from "PhysioNet, Fantasia Database" [11] [12]. This database was chosen because it contains the results of several ECG traces in very similar conditions: the subjects underwent 120 minutes of supine resting, while watching the movie Fantasia (Disney, 1940) to help maintain wakefulness, and continuous (ECG) registration is acquired. Furthermore, subjects have been previously examined to ensure their health situation. Table II shows the characterization of the analysed population.

TABLE II: Population characterization

	Number of subjects	**Young subjects** (20 - 35 years old)	**Old subjects** (65 - 85 years old)
Male	18	9	9
Female	16	9	7

B. Pre-processing

Before the evaluation of the differences between male and female traces, it is necessary to pre-process them, in order to eliminate the noise and make the different ECGs comparable. The following steps were made on each trace before the evaluation through FDA:

1) **Extraction of single beat**: all the evaluations on the ECG traces were made considering a single heartbeat in order to reduce computational costs and time. For all the traces the second recorded beat was considered. For the identification it was necessary to previously detect all R peaks.
2) **Zero averaging**: to ensure that the different beats are comparable to each other every trace has its own mean subtracted.
3) **Smoothing**: the ECG traces provided by the database are discreet and affected by noise. Therefore filtration and smoothing were necessary.
   A low pass Butterworth filter was used for noise removal.As regards the smoothing operation, it consists in highlighting significant patterns, attenuating the residual noise generated by environmental, electrical, electronic or physiological artifacts. Even a small failure of monotonicity in a curve can have serious consequences for the corresponding growth velocity, and even more so for acceleration curves.
   There are several methods for the trace smoothing, in particular in this study we employed the wavelets method: unlike conventional filters [13], the wavelet-based method is able to better reconstruct a function, especially if not periodic, such as the ECG. In particular, the Symlet wavelet family is applied (Fig. 1) as this basis is sufficiently smooth to allow the computation of the first derivative of the estimated functional data.

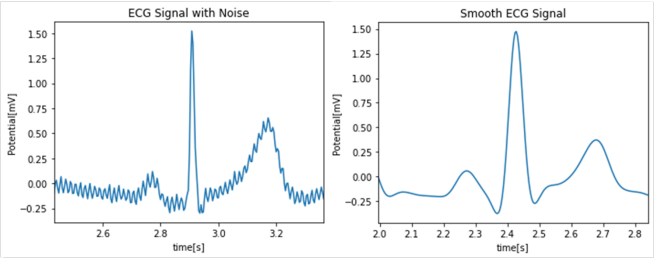

Fig. 1: On the left the raw signal, on the right the same signal after being filtered with Butterworth low-pass and smoothed with wavelet method

C. Landmarks registration

Within the heartbeat we can define events that occur in sequence and at certain times. Each individual has his own "biological time", that means that every event happens with different timing. This means that if we overlap the filtered beats as they are, they would all be staggered. Therefore, a landmark registration is required, that means a synchronization based on reference points.

The landmarks considered are the peaks of the characteristic waves of the ECG, therefore the peaks P, Q, R, S and T. In reference to these points, we define a warping function $h(t)$, that *"maps the time"* into its physiological counterpart. For this purpose, we define a function $h_i(t)$ for subject i such that at physiological time t this subject has a chronological timing of $h_i(t)$. Therefore, the warping functions modify the geometry of the curves so that the time of their peaks matches [14]. The average values of the different landmarks are shown in the Table III, while the warping functions are represented in Fig. 2. In order to report all the curves in a standardized time, it is necessary to "accelerate" or "slow down" the timing. This is done by the function:

$$V^*(t) = V[h(t)] \tag{1}$$

which scales the curves so that the various peaks occurs at the "right time". Fig. 3 shows the registered curves.

D. Derivatives

The study of the first and second derivatives is fundamental for the evaluation of the morphology of the curve as it allows the analysis of the local properties and so the evaluation of any alteration of the curve. In particular, the first derivative corresponds to the velocity: the comparison between derivatives of different traces helps to understand the timing of the different events. The second derivative, on the other hand, corresponds to the acceleration and helps us to understand the curvature of the tracings.

TABLE III: Mean and standard deviation of the landmarks timing of all the traces

	Mean value	Standard deviation
P [sec]	0.25	0.07
Q [sec]	0.30	0.06
R [sec]	0.34	0.06
S [sec]	0.38	0.07
P [sec]	0.54	0.09

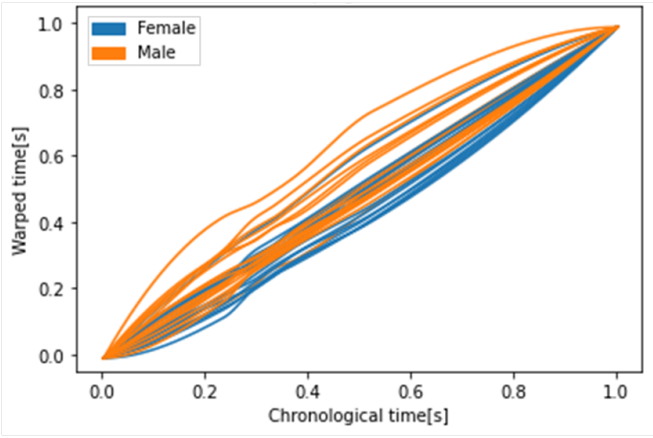

Fig. 2: Time warping function *h(t)*. The curves above the diagonal represent the slowest ECG traces for which the warping function had to "speed up" the physiological time. The opposite is true for the curves below the diagonal. It is possible to observe how males mainly occupy the upper portion of the Figure, while female's timing results slower.

E. Cluster Analysis

Once the signal has been filtered and registered it is possible to proceed with its evaluation through a multivariate functional k-means clustering procedure.

Cluster analysis consists in grouping data based on similarities. For each grouping a centroid is identified and, based on the distance from it, a certain function is assigned to its grouping through an optimization algorithm. The distance can be calculated according to different equations. In this study the Euclidean distance is considered:

$$d\mathbf{F_i}(t),\mathbf{F_j}(t) = \sqrt{\sum_{r=1}^{n}\int_T (\mathbf{F_i}(t)-\mathbf{F_j}(t))^2 dt} \qquad (2)$$

This equation depends only on the function of the ECG trace, while it would be interesting to consider also other equations for the computation of the distance. In particular it would be interesting to consider the role of the first derivative of the curve to distinguish between different morphologies. The evaluation of the distance minimization, also using the definition of the distance as a function of:

- both ECG trace and its derivative;
- only of the ECG trace derivative;

and the respective costs, would allow a better decision on the optimal cluster assignment method.

The computation of the centroids of each cluster takes place through an iterative algorithm in which we want to minimize the distance between the centroids and the functions belonging to that group. The number of groupings is defined a priori.

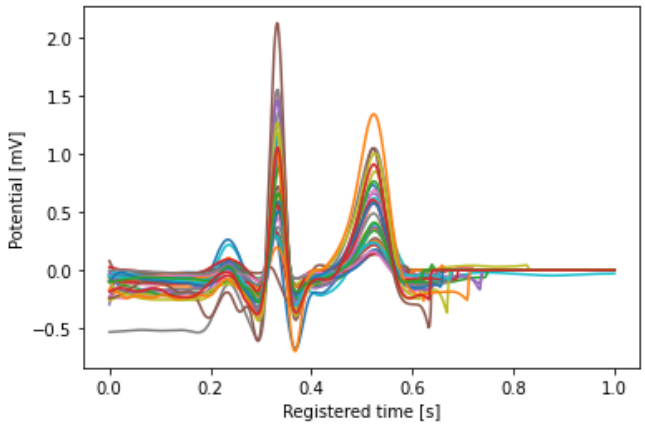

Fig. 3: Smoothed and registered ECG traces of 34 healthy subjects. The registration was done by synchronizing the landmarks through the warping function

In this study, fuzzy c-means clustering was chosen to be applied in the cluster analysis because previous literature has shown as it can be considered a more accurate algorithm compared to the standard k-means algorithm. Unlike the k-means algorithm, where the data points exclusively belong to one cluster, in the case of the fuzzy c-means algorithm, the data points can belong to more than one cluster with a certain percentage. Fuzzy c-means clustering gives comparatively better results for overlapped data sets [15].

As a first approach, a single analysis was carried out on all the traces, defining two clusters, considering one for men and one for women. The assignment of the clusters was made to optimize the results, that means the distance minimization. Further analysis were considered necessary and other two methods were performed: only the young subjects and only the old subjects. The aim is to confirm the experimental evidence previously discussed by comparing the ECGs through two different cluster analyzes, one for young individuals and one for elderly individuals, each with two groupings: men and women. For the different groups of subjects, the clustering optimization algorithm led to the definition of different centroids in order to minimize the distance of the ECG curves from the centroid of the cluster they belong to.

The results of the cluster analysis were then compared with real data through a confusion matrix and, in order to evaluate the validity of the method, the accuracy was assessed using the formula:

$$acc = \frac{tp+tn}{tp+fp+tn+fn} \qquad (3)$$

Where *tp*, *tn*, *fp* and *fn* indicate respectively the results that are true positive, true negative, false positive and false negative.

For a further analysis of the shape of the ECG curve, the cluster analysis was carried out also on its derivatives in order

to study the slope and its variation.

III. Results

The first classification approach was applied to the curves of the traces. The result of the different methods are the centroids of the clusters shown in Fig. 4. The two graphs relate only to the young and old population, as it is important to distinguish the subjects also according to age. The same

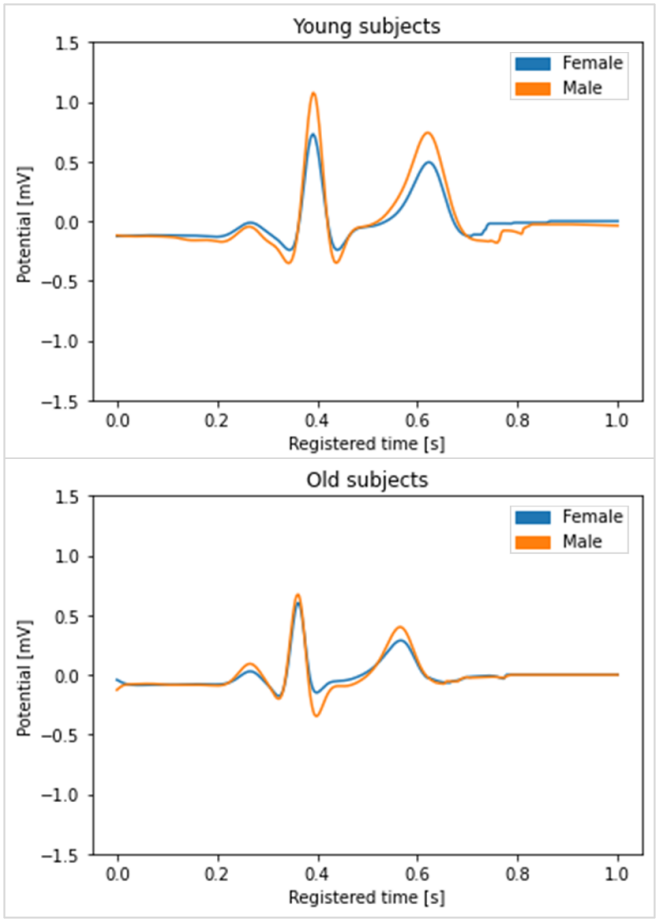

Fig. 4: Comparison of the centroids found according to the assignment to the clusters with different methods: in the first plot are represented the young subjects, in the second plot elderly subjects.

cluster analysis was also carried out on the derivatives of the ECG curves to evaluate the differences, as well as a function of the vertical displacement of the curves, that means the amplitudes of the signals, also of variations in slopes and curvatures. A comparison between the centroids obtained through the classification with the methods previously described is shown in Fig. 5 for the first derivative and in Fig. 6 for the second derivative.

The accuracy of the method was evaluated and the values related to the different methods are reported in Table IV. The application of the method reaches the highest accuracy values when applied to the young population for all the curves considered, in particular for the second derivative (81.11%),

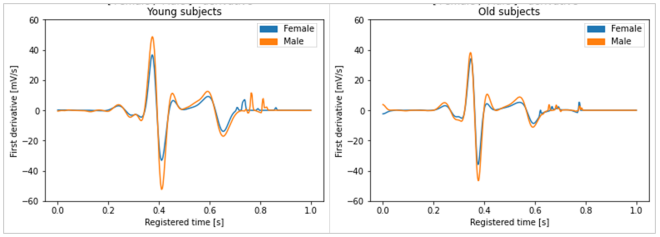

Fig. 5: Comparison of the centroids of the cluster analysis on ECG first derivatives in young (left) and old subjects (right).

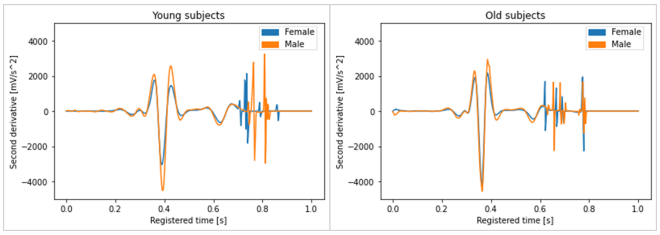

Fig. 6: Comparison of the centroids of the cluster analysis on ECG second derivatives in young (left) and old subjects (right).

while it is not sufficiently satisfying when the method is applied to all the ECG traces together.

TABLE IV: Comparison of the accuracy of different methods used for cluster analysis applied on the ECG traces, their derivatives and their second derivatives.

	Accuracy (%)		
	Traces	First derivative	Second derivative
All subjects	56.25	58.65	61.94
Young subjects	77.78	77.78	81.11
Old subjects	71.42	74.71	74.71

IV. Discussion

As it is shown in Table IV, whether the method is applied to traces, first derivatives or second derivatives, when we refer to the all subjects together we obtain a lower accuracy. The reason is that the traces were evaluated distinguishing only between sexes and not by age, which is instead an extremely relevant factor.

Moreover, the accuracy has greater value when the analysis is applied only on young patients, while it decreases in elderly patients. These results are consistent with what stated by Carbone et al. [3]: the risk of cardiovascular disease tends to be equal between elderly men and women, while the risk is different in the age group between 20 and 50 years. The different propensity to the pathology presupposes differences in the ECG tracings, confirmed in this study.

The accuracy increases when the method is applied to the first derivative of the curves, rather than to the ECG traces themselves. This depends on the fact that the main difference between male and female ECG lies in the duration of the waves, rather than in the amplitude of the tracings. As shown in Fig. 5, the greatest difference between male and female

traces is observable at a young age, while it is more similar between elderly subjects. In general, the peaks reach higher values in the ECG of younger population and the female category presents lower peaks at any age.

Finally, in the figure we can observe some oscillations especially at the end of the registered time. The reason is an amplification of the error due to the noise that occurs in the calculation of the derivatives. Such fluctuations are also present in the curves of the second derivatives, which in turn are relied to the rate of change in the slope of the curves. From the comparison in Fig. 6 it is possible to note that also in this case the differences are greater at a young age and that women tend to have smaller peaks than men at all ages. The major differences emerge in the QRS interval, which underlies ventricular depolarization [16].

From the comparison of the results obtained from the traces and their derivatives, we can assert that for a correct analysis of the ECG signal, the evaluation of the peaks values alone is insufficient because the greatest differences between the sexes are related to the duration of subintervals, rather than to the amplitudes. The analysis of the derivatives should be considered for the development of a classification tool.

However, we aim at further increasing the accuracy of the method proposed by implementing the following steps in future works:

- As mentioned in Section II, the belonging of a function to a certain cluster can be evaluated through multiple distance definitions. It would be interesting to verify if a different definition of the distance would affect the results;
- In this work the value of k was initialized *a priori*. It would be interesting to employ a Silhouette analysis comparing multivariate functional k-mean procedures with different values of k, to verify that the number of cluster employed is actually the optimal one;
- The analysis was carried out on a limited number of individuals, all healthy. A possible extension of the present work could verify if different pathologies affect the ECG traces of men and women in distinct ways.

V. Conclusions

In this work we propose a tool for the semi-automatic analysis and classification of ECGs based on FDA and fuzzy c-means clustering. The method achieves an accuracy always higher than 70% in every simulation performed considering the young and old population separately. Thus, we conclude that this method is effective in distinguishing the male and female ECG signals, especially in case of young subjects. In addition, the accuracy is greater if the clustering is performed on the derivatives of the ECG traces, rather than on the curves themselves: the greatest differences are not due to the amplitude of the signal and its peaks, but to the timing and, therefore, male and female traces are characterized by different speeds and rates of change.

Our results evidenced the existence of significant sex-based differences. Moreover, the role of the age in shaping the ECG morphology emerged as fundamental, as well as the evaluation of the derivatives.

This work lays the foundations for the development of a sex-differentiated ECG reading method, based solely on morphology. In fact, thanks to the FDA it is possible to carry out a categorization of the ECG traces and to develop a smart analysis tool for their correct reading: a better understanding of the differences between men and women would allow, in addition to a better diagnosis, a research for the development of drugs for cardiovascular diseases that is more targeted on the specific patient.

References

[1] Jackson JK. Mc Namara K, Alzubaidi H. Cardiovascular disease as a leading cause of death: how are pharmacists getting involved? *Integr Pharm Res Pract.*, 8:1–11, 2019.

[2] Roberto Cangemi, Giulio Francesco Romiti, Giuseppe Campolongo, Eleonora Ruscio, Susanna Sciomer, Daniele Gianfrilli, and Valeria Raparelli. Gender related differences in treatment and response to statins in primary and secondary cardiovascular prevention: the never-ending debate. *Pharmacological research*, 117:148–155, 2017.

[3] Vincenzo Carbone, Franco Guarnaccia, Giovanni Carbone, Giovanni Battista Zito, Ugo Oliviero, Silvia Soreca, and Francesca Carbone. Gender differences in the 12-lead electrocardiogram: clinical implications and prospects. *Italian Journal of Gender-Specific Medicine*, 6(3):126–141, 2020.

[4] Juan Tamargo, G Rosano, T Walther, J Duarte, Alexander Niessner, Juan Carlos Kaski, C Ceconi, Heinz Drexel, Keld Kjeldsen, G Savarese, et al. Gender differences in the effects of cardiovascular drugs. *European Heart Journal–Cardiovascular Pharmacotherapy*, 3(3):163–182, 2017.

[5] Hanna Mieszczanska, Grzegorz Pietrasik, Katarzyna Piotrowicz, Scott McNitt, Arthur J Moss, and Wojciech Zareba. Gender-related differences in electrocardiographic parameters and their association with cardiac events in patients after myocardial infarction. *The American journal of cardiology*, 101(1):20–24, 2008.

[6] Peter Smetana, Velislav N Batchvarov, Katerina Hnatkova, A John Camm, and Marek Malik. Sex differences in repolarization homogeneity and its circadian pattern. *American Journal of Physiology-Heart and Circulatory Physiology*, 282(5):H1889–H1897, 2002.

[7] Ernst Simonson, HENRY BLACKBURN JR, THOMAS C PUCHNER, Pauline Eisenberg, Fernanda Ribeiro, and Manuel Meja. Sex differences in the electrocardiogram. *Circulation*, 22(4):598–601, 1960.

[8] Arthur J Moss. Gender differences in ecg parameters and their clinical implications. *Annals of noninvasive electrocardiology: the official journal of the International Society for Holter and Noninvasive Electrocardiology, Inc*, 15(1):1, 2010.

[9] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[10] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[11] Morin R Goldberger AL Lipsitz LA Iyengar N, Peng CK. Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. *Am J Physiol*, 271(4 Pt 2):1, 1996.

[12] Amaral L. Glass L. Hausdorff J. Goldberger, A. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):e215–e220, 2000.

[13] F. Gustaffson. Determining the initial states in forward-backward filtering. *Transactions on Signal Processing*, 46(1):988–992, 1996.

[14] Francesca Ieva, Anna M Paganoni, Davide Pigoli, and Valeria Vitelli. Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):401–418, 2013.

[15] Zeynel Cebeci and Figen Yildiz. Comparison of k-means and fuzzy c-means algorithms on different cluster structures. *Agrárinformatika/journal of agricultural informatics*, 6(3):13–23, 2015.

[16] Vijay S Chauhan, Andrew D Krahn, Bruce D Walker, George J Klein, Allan C Skanes, and Raymond Yee. Sex differences in qtc interval and qt dispersion: dynamics during exercise and recovery in healthy subjects. *American heart journal*, 144(5):858–864, 2002.