Computer Methods and Programs in Biomedicine 165 (2018) 53–67

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Prediction of paroxysmal Atrial Fibrillation: A machine learning based approach using combined feature vector and mixture of expert classification on HRV signal

Elias Ebrahimzadeh[a,e,f,*], Maede Kalantari[b], Mohammadamin Joulani[c], Reza Shahrokhi Shahraki[d], Farahnaz Fayaz[e], Fereshteh Ahmadi[e]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] Department of Biomedical Engineering, Faculty of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran
[c] Student Research Committee, Faculty of Medicine, Iran University of Medical Sciences, Tehran, Iran
[d] Faculty of Electrical and Biomedical Engineering, University of Sheikhbahaee, Isfahan, Iran
[e] Biomedical Engineering Department, School of Electrical Engineering, Payame Noor University of North Tehran, Tehran, Iran
[f] Seaman Family MR Research Center, Hotchkiss Brain Institute, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

ARTICLE INFO

Article history:
Received 7 April 2018
Revised 17 June 2018
Accepted 25 July 2018

Keyword:
Paroxysmal atrial fibrillation
Heart rate variability
Feature reduction
Local subset feature selection
Mixture of Expert

ABSTRACT

Background and Objective: Paroxysmal Atrial Fibrillation (PAF) is one of the most common major cardiac arrhythmia. Unless treated timely, PAF might transform into permanent Atrial Fibrillation leading to a high rate of morbidity and mortality. Therefore, increasing attention has been directed towards prediction of PAF, to enable early detection and prevent further progression of the disease. Notwithstanding the pharmacological and electrical treatments, a validated method to predict the onset of PAF is yet to be developed. We aim to address this issue through integrating classical and modern methods.

Methods: To increase the predictivity, we have made use of a combination of features extracted through linear, time-frequency, and nonlinear analyses performed on heart rate variability. We then apply a novel approach to local feature selection using meticulous methodologies, developed in our previous works, to reduce the dimensionality of the feature space. Subsequently, the Mixture of Experts classification is employed to ensure a precise decision-making on the output of different processes. In the current study, we analyzed 106 signals from 53 pairs of ECG recordings obtained from the standard database called Atrial Fibrillation Prediction Database (AFPDB). Each pair of data contains one 30-min ECG segment that ends just before the onset of PAF event and another 30-min ECG segment at least 45 min distant from the onset.

Results: Combining the features that are extracted using both classical and modern analyses was found to be significantly more effective in predicting the onset of PAF, compared to using either analyses independently. Also, the Mixture of Experts classification yielded more precise class discrimination than other well-known classifiers. The performance of the proposed method was evaluated using the Atrial Fibrillation Prediction Database (AFPDB) which led to sensitivity, specificity, and accuracy of 100%, 95.55%, and 98.21% respectively.

Conclusion: Prediction of PAF has been a matter of clinical and theoretical importance. We demonstrated that utilising an optimized combination of — as opposed to being restricted to — linear, time-frequency, and nonlinear features, along with applying the Mixture of Experts, contribute greatly to an early detection of PAF, thus, the proposed method is shown to be superior to those mentioned in similar studies in the literature.

1. Introduction

Electrocardiography (ECG) detects, measures, and records the electrical activity of the heart. Most of cardiac diseases are known to have associations with short term, i.e. beat-to-beat, or long term changes of the rhythm of cardiac excitation. Therefore, analyzing

E-mail address: e_ebrahimzadeh@ut.ac.ir (E. Ebrahimzadeh).
* Corresponding author.

https://doi.org/10.1016/j.cmpb.2018.07.014
0169-2607/

characteristic changes in ECG signals allows for diagnosing different heart disorders.

Among the most common cardiac arrhythmia in the general population is Atrial fibrillation (AF). Although not immediately life-threatening itself, secondary complications, especially thromboembolism, can imply dramatic consequences and pose a major risk of stroke to such an extent where about 15% of strokes occur in people with AF [1]. In the United States alone, AF affects an estimated 2.2 million people, with an increased incidence in the elderly population [2] The risk of stroke, resulted from thrombus formation in the heart's poorly contracting chambers, is increased fivefold in patients with AF, and the risk of death is almost doubled [3,4]. As catastrophic as the end results could be, there is still evidence that suggests atrial fibrillation remains undetected in up to 40% of patients [5–7]. Increasing with age, the prevalence of AF is 0.5% for the group aged 50–59 years, and rises to approximately 10% in the group aged 80–89 years [8].

In general, AF does not present a life threatening condition, can occur without symptoms or, in many cases, may even stay unnoticed for a while. Nevertheless, it is often associated with a high risk of cardiovascular morbidity and mortality [9]. The aim of therapy is to prevent stroke and regain sinus rhythm [10], however, prevention and treatment of AF is still far from satisfactory.

Clinically, AF presents itself in different forms, commonly starting as paroxysmal, and becoming more persistent with time. Paroxysmal AF (PAF) refers to attacks of AF which last from 2 min to less than 7 days and spontaneously revert to normal sinus rhythm, i.e. is self-terminating. Permanent AF, on the other hand, lasts more than 7 days while sinus rhythm cannot be restored or maintained, i.e. is non-terminating. Chronic AF may be the end result of PAF in about 30% of the group of PAF patients [11,12].

About 18% of PAF evolve to permanent AF over a course of 4 years [10]. Since there is rarely an impulse in a PAF patient heart which is able to depolarize the atria, some distorted small waves might be observed instead of a normal P-wave. As paroxysmal atrial fibrillation may finally become a critical disease which results in heart strokes and thromboembolisms. A major benefit of automatic detection of patients suffering from PAF is the ability to develop a preliminary time and cost effective screening procedure during a short-time visit to clinics [13]. The maintenance of sinus rhythm can lead to decreased symptoms and possibly a decrease in the atrial remodeling that causes increased susceptibility to future episodes of PAF [14]. In addition, there may be a reduction in the risk of strokes and thromboembolic events. Although Pharmacological or electrical treatments are available, there is still a need for a reliably validated method for predicting the onset of PAF. Older age is traditionally believed to be the strongest predictor for the development of AF [15,16]. Over the last decades, several studies have focused on finding algorithms able to predict PAF through the analysis of surface electrocardiographic (ECG) records. Such researches can mainly be classified into premature atrial complexes (PAC) detection, and heart rate variability (HRV) analysis. Since approximately 93% of PAF episodes are triggered by PACs [17], several methods proposed in the literature use the first method, i.e. detection of PACs, as a means to predict PAF. Zong et al. [18] studied the number and timing of PACs in the ECG episodes. Having detected PACs in 30-min ECG segments from Atrial Fibrillation Prediction Database (AFPDB), they revealed that not only the number of PACs increases in episodes preceding PAF, but also these complexes occur mostly towards the end of the episodes. They achieved a sensitivity of 79% for predicting the onset of PAF. Thong et al. [19] developed an algorithm based on a predictor that used three criteria: the number of isolated PACs that are not followed by a regular RR interval, runs of atrial bigeminy and trigeminy, and the length of any short run of paroxysmal atrial tachycardia. They showed that an increase in activity detected by any of these three criteria is an indication of an imminent episode of PAF.

Alternatively, HRV analysis has been applied as another basis for prediction-oriented studies. Lynn and Chiang [20] proposed an algorithm predicated on non-linear features calculated from return map and difference map of HRV signal, which was reported to have acquired a sensitivity of 64%. Yang and Yin [21] developed a symbolic dynamic approach, known as Footprint analysis, to investigate heart rate dynamics before PAF attacks. Vikman et al. [22] calculated the approximate entropy (ApEn) and short term scaling exponent α_1 of HRV over 20-min periods and concluded that a reduced complexity of RR interval dynamics and altered fractal properties usually precede the onset of PAF, as indicated by decreasing value of ApEn and α_1. Chesnokov et al. [13] combined complexity and spectral analysis of the 30-min HRV segment from AFPDB and noticed statistically significant increase in the very low frequency (VLF) band, low frequency (LF) band and high frequency (HF) band for the records immediately before PAF compared to distant ones, but LF/HF ratio did not discriminate these two groups with statistical significance. What also came to their attention was that complexity features like ApEn, sample entropy (SmEn) and their multiscale versions exhibit smaller values in episodes preceding PAF compared to distant ones.

Over the years, several other researches have been conducted in attempts to increase the prediction time as well as improving the quality and effectiveness of the prediction procedure with the aid of classifiers and feature extraction methods regarding different processing domains, although the reported results do not necessarily include the interpretation of clinical signs.

What is more is that the studies in the field of PAF prediction through classification have not yet presented markers that can classify the classes in an accurate and a precise manner. In fact, a major challenge in such studies is that they are mainly focused on certain features from one or two processing domains in particular. We, however, have come to realize that the combination of informative features from all processing domains encapsulates the advantages of each domain, and can therefore be regarded as a Golden Package, yielding the best possible result. With that in mind, we aim to present the best combination of features, that is the optimal combination of suitable markers which are elicited from different processing domains and can be compatible with the proposed classifier. In other words, we seek to achieve the best results through using the optimized combination of suitable markers as well as using effectual classifiers.

It should be noted that the current study does not seek to once again define applicable and effective features for detection and prediction of PAF, as there has already been adequate discussion addressing feature extraction in our previous works and other recent studies. What seems to be lacking attention is the need for an appropriate strategy to manage the extracted features to such an extent that the best separability is presented. To this end, deploying a suitable tactic to select extracted features could bring about outstanding results for prediction of the spontaneous onset of PAF. We have accordingly, applied a novel and automated approach to Local Feature Subset Selection with the assistance of the most rigorous methodologies, which have formerly been developed in previous works of our team, for extracting features from nonlinear, time-frequency and classical processes. The proposed methods enable us to select features that differ from one another in each 5-min before the PAF through the agency of selecting optimal features in each 5-min interval of the signal as an episode.

Furthermore, as done in previous studies, having extracted the HRV signal from the ECG and divided the signal into 5-min intervals, linear features were elicited, and Wigner Ville transform was applied to extract Time-Frequency, and thereupon, non-linear features [23–27].

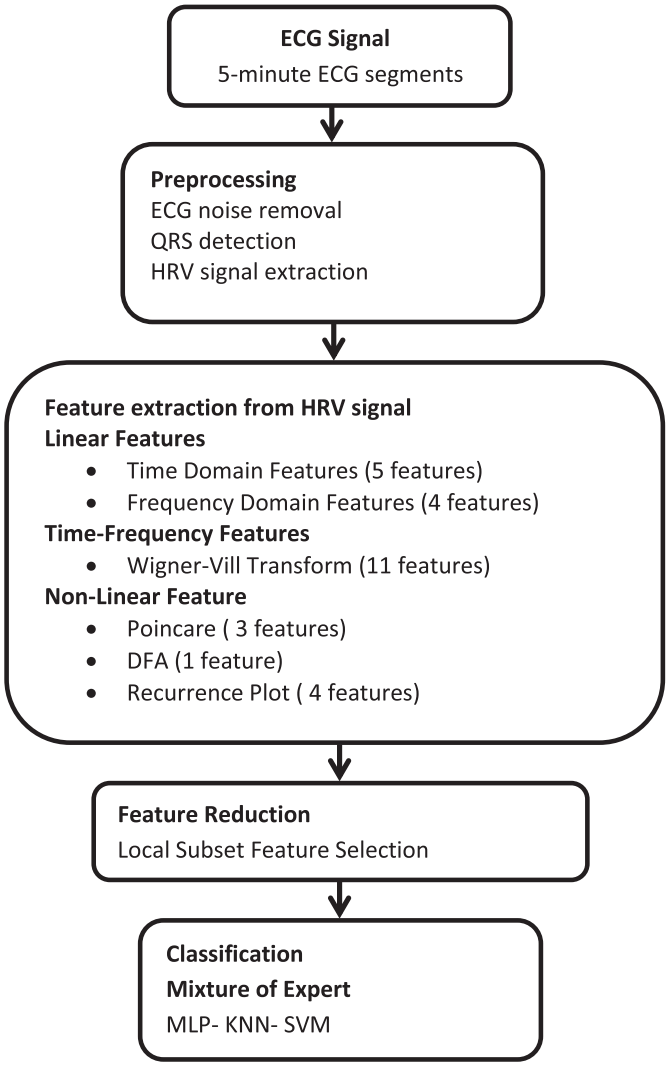


Fig. 1. Block diagram of the proposed approach for prediction of PAF.

In the Material and Methods section, the best combination of features is selected based on their ability to generate the highest degree of differentiation between the two classes. This is accomplished through applying Local Feature Subset Selection for 5-min intervals. Employing the Mixture of Experts classification proceeds to ensure precise decision-making on the output of different processes. By means of Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Mixture of Expert classifiers, the two classes are ultimately separated. In the Result section, the outcome of the obtained prediction would be analyzed and compared to the results of similar studies. The most informative and effective features would afterwards be presented according to the highest accuracy of classification. Fig. 1 illustrates the block diagram of our approach for predicting the onset of PAF.

2. Material and methods

2.1. Dataset used

In line with previous works in the literature [28–33], 106 data from 53 pairs of ECG recordings (each pair is recorded from different PAF patients) are obtained from the standard database called Atrial Fibrillation Prediction Database (AFPDB) [34]. Each pair of data contains one 30-min ECG segment that ends just before the onset of PAF event and another 30-min ECG segment at least 45 min distant from the onset. Each ECG segment contains two-channel traces from Holter recording with sampling rate of 128 Hz and 12-bit resolution. In this paper, the 5-min HRV segment that is at least 45 min distant from PAF event is assigned a class label of "non-PAF", while the HRV segment that immediately precedes PAF event is given a class label of "PAF". Fig. 2 shows a sample of ECG signal of a person with PAF several seconds before the occurrence of PAF.

2.2. Preprocessing

At first, noise reduction is performed to have two major types of noise removed; i.e. Baseline Wandering caused by the patient's respiration, which contains low frequency components and, Power Line Interference, which contains high frequency components. To remove the baseline drift, ECG signals of both healthy and patient subjects were filtered using a moving average filter. With the moving average filter, an array of raw (noisy) data [y1, y2, ..., yN] can be converted to a new array of smoothed data. The smoothing process is similar to a low-pass filter response:

$$y_s = \frac{1}{2N+1}\sum_{k=-N}^{N} y(i+k) \tag{1}$$

In this equation, ys (i) is the smoothed value of the ith data point, and N denotes the number of data neighbors on each side of ys (i). The number $2N+1$ is usually used as the window size. We apply a two-stage moving average filter on each ECG signal. First, signals pass through a moving average filter with a window size of 1/3 signal length. Then, this pre-filtered signal moves through a moving average filter with a window size of 2/3 signal length. The drift of the signal will be removed by subtracting the output of this filter from original data [35]. A median filter with a notch filter is then used to remove the power line frequency [36]. The median filter is a nonlinear digital filter. The mathematical concept of median is used to describe the middle of the data. The median filter accordingly uses both past and future values for predicting the current point. We can describe the operation of the median filter as follows:

$$y_p(n) = median(x(n+M1)\ldots x(n)\ldots x(n-M2)) \tag{2}$$

Where:

$$median(x_1 \ldots\ x_N) = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & N\,odd \\ \frac{1}{2}\left[x\left(\frac{N}{2}\right) + x\left(\frac{N}{2}+1\right)\right] & N\,even \end{cases} \tag{3}$$

And where $x_{(n)}$ is the n th smallest of values x_1 through x_N . For the notch filter, we use the following formula [39]:

$$\begin{aligned} y_k\ (n) = \frac{1}{2}\ &[(1+a_2)x(n) - 2a_1x(n-1) \\ &+(1+a_2)x(n-2) + a_1y(n-1) - a_2y(n-2)\,, \end{aligned} \tag{4}$$

Where:

$$\begin{aligned} a_1 &= \frac{2\cos(w_0)}{1+\tan\left(\frac{\Omega}{2}\right)} \\ a_2 &= \frac{1-\tan\left(\frac{\Omega}{2}\right)}{1+\tan\left(\frac{\Omega}{2}\right)} \end{aligned} \tag{5}$$

In Eqs. (3) and (4), x(n) is the input signal and $y_k(n)$ is the output of the notch filter. $\omega 0$ and Ω are the notch frequency and 3-dB rejection bandwidth, respectively. We used filtered signals in all procedures. The Pan–Tompkins [40] algorithm is used for detection of the QRS-complex, and especially the R wave, helping us determine the RR-intervals and HRV signal (the RR-interval between two consecutive beats). After these processes, the HRV signal is ready to have features extracted from it. Fig. 3 demonstrates the extracted HRV signals of 5-min ECG segments from patients, immediately before PAF onset and distant ones.

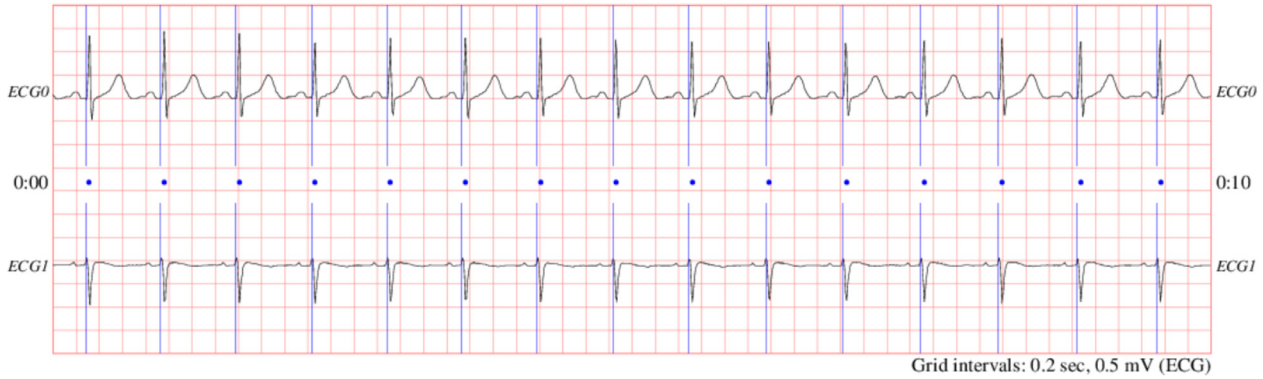


Fig. 2. ECG signal of a person with paroxysmal atrial fibrillation.

When extracting features, our main focus, according to our recent studies [23–27], lies on eliciting 5 time, 4 frequency, 11 time-frequency and 4 nonlinear features.

2.3. Classical features analysis

This stage focuses on classic linear features. A total number of 9 features are extracted including 5 features in the time domain and 4 in the frequency domain.

2.3.1. Time-Domain features

Statistical time-domain measures are divided into two classes:

- Direct measurements of RR intervals
- Measurements from the differences between RR intervals

2.3.1.1. Direct measurements of RR intervals. These features include two basic time domain variables that are calculated as:

$$\text{MNN}: RR\text{m} = \frac{1}{N}\sum RR(i) \tag{6}$$

$$\text{SDNN} = \sqrt{\frac{1}{N}\sum (RR(i) - RR_m)^2} \tag{7}$$

2.3.1.2. Measurements from the differences between RR intervals.

- The square root of the mean of the squares of differences between adjacent RR Intervals (RMSSD).

$$\text{RMSSD} = \sqrt{\frac{1}{\text{N}}\sum (\text{RR(i}+1) - \text{RR(i)})^2} \tag{8}$$

- The standard deviation of differences between adjacent RR intervals (SDSD).

$$\text{SDSD} = \sqrt{\frac{1}{\text{N}}\sum_{\text{i=1}}^{\text{n}} (\text{RR}_{(\text{dif})}(\text{i}+1) - \overline{\text{RR}_{(\text{dif})}})^2} \tag{9}$$

$$\overline{\text{RR}} = \frac{1}{\text{N}}\sum (\text{RR(i}+1) - \text{RR(i)}) \tag{10}$$

$$\text{RR(dif)} = (\text{RR(i}+1) - \text{RR(i)}) \tag{11}$$

- The proportion derived from dividing the number of interval differences of RR intervals greater than 50 ms by the total number of RR intervals (PNN50)

$$\text{PNN50} = \frac{[\text{RR(i}+1) - \text{RR(i)} > 50ms]}{\text{total}\left(\text{RR}_{(\text{dif})}\right)} \tag{12}$$

2.3.2. Frequency domain features

For spectral analysis, we used power spectral density (PSD) estimate of the HRV signal. Spectrum features are able to discriminate between the sympathetic and parasympathetic contents of the HRV signal, which are affected before PAF attacks. It is generally accepted that the spectral power in the high frequency (HF) band (0.15–0.4 Hz) of the RR intervals reflects the respiratory sinus arrhythmia (RSA) and thus, cardiac vagal activity. On the other hand, the low frequency (LF) band (0.04–0.15 Hz), is associated with the baroreceptor control and is mediated by both vagal and sympathetic systems [37]. The LF, HF, and VLF (Very Low Frequency) bands PSD as well as the ratio of the LF and HF bands power spectral density (LF/HF) are used as the frequency domain features of the RR interval signal. The power spectral density (PSD), shown in Fig. 4, was computed by Burg parametric method.

2.4. Time-Frequency domain analysis

Among the various time-frequency methods, the Smoothed Pseudo Winger Ville distribution (SPWVD) is preferred in this paper. It provides better time frequency resolution than nonparametric linear methods, an independent control of time and frequency filtering, and power estimates at lower variance with parametric methods when rapid changes occur [23–27,38].

The SPWVD of the discrete signal x (n) is defined by [53].

$$\text{X(n, m)} = 2\sum_{\text{k}=-\text{N}+1}^{\text{N}-1} |\text{h(k)}|^2 \sum_{\text{p}=-\text{M}+1}^{\text{M}-1} \text{g(p)}\text{r}_\text{x}(\text{n}+\text{p}, \text{k})\text{e}^{-\frac{\text{j}2\pi\,\text{km}}{\text{N}}} \tag{13}$$

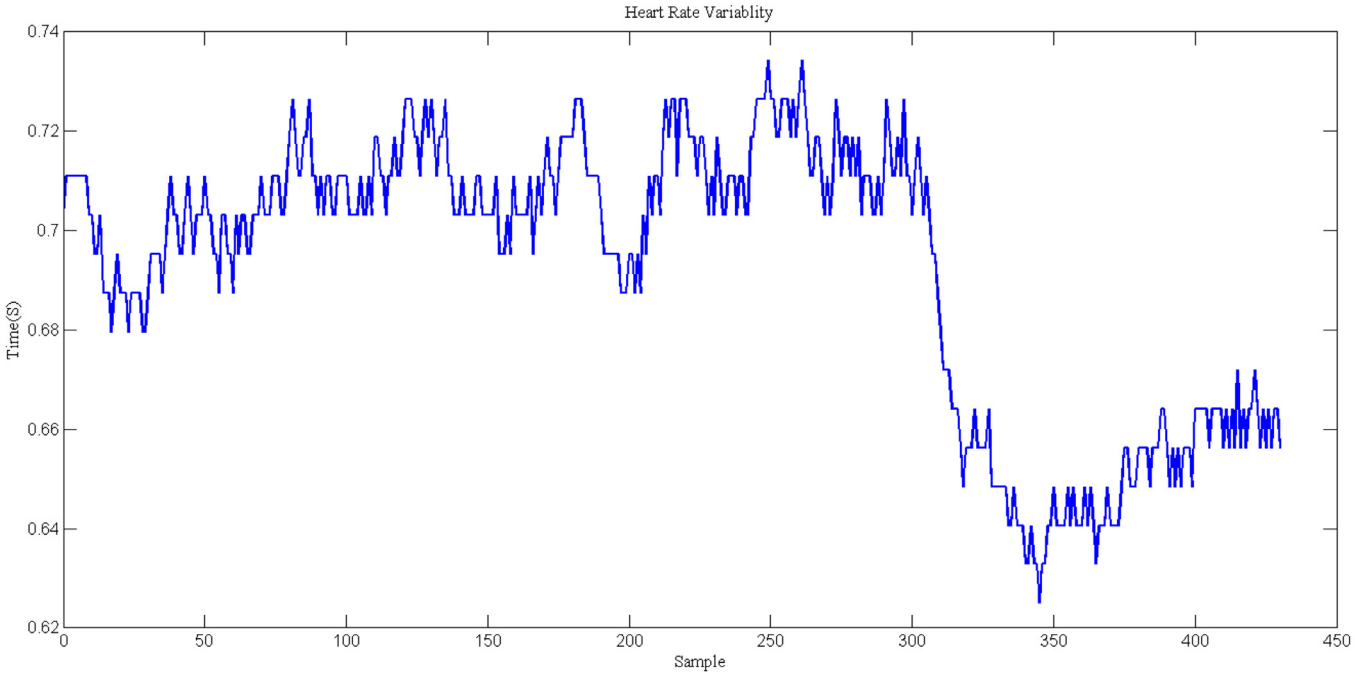
where n and m are the discrete time and frequency indexes, respectively, h(k) is the frequency smoothing symmetric normed window of length 2 N *1, g(p) is the time smoothing symmetric normed window of length 2 M *1 and r_x(n, k) is the instantaneous autocorrelation function, defined as:

$$\text{r}_\text{x}(\text{n}, \text{k}) = \text{x(n}+\text{k).x}*\text{(n}-\text{k}) \tag{14}$$

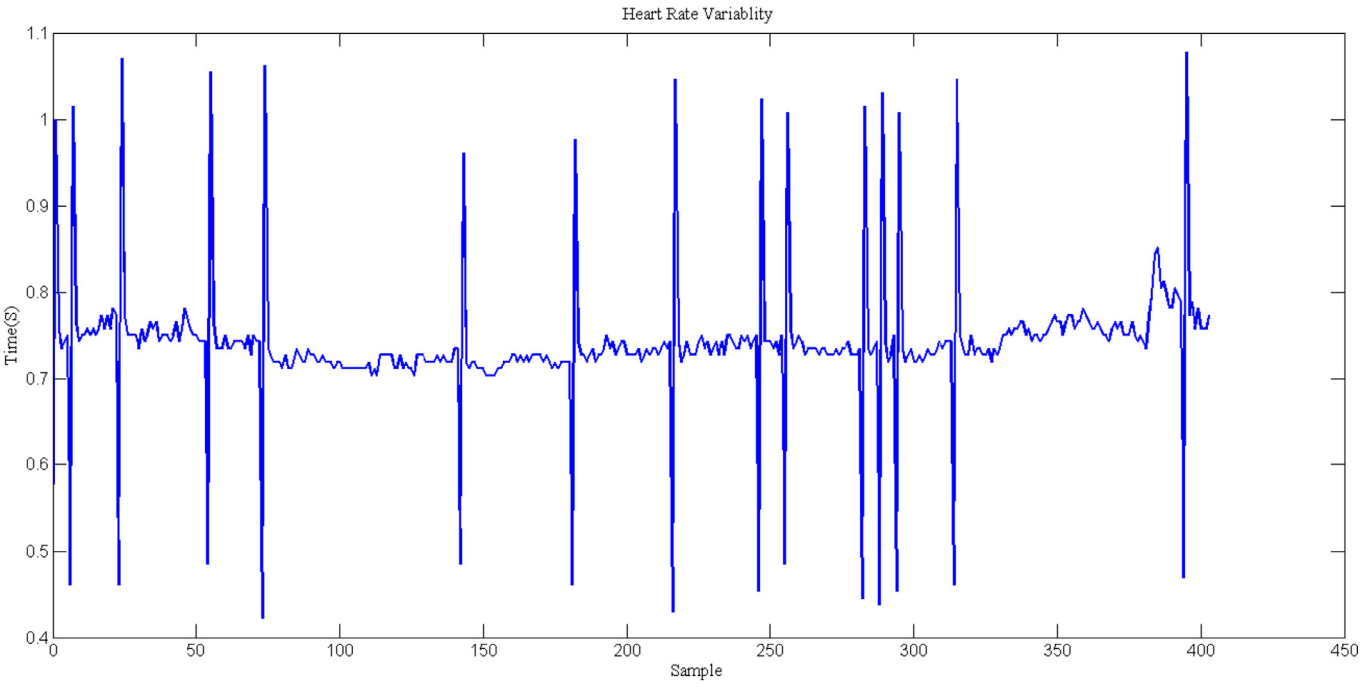
Fig. 5 shows the result of applying Wigner Ville transform to the HRV signal. Then, according to studies [23–27,38], 11 features are extracted as follows:

2.5. Nonlinear analysis

The complex non-stationary behavior of cardiovascular system calls for a non-linear analysis to explore dynamic, non-linear features as well as time-frequency features in HRV signal. With that in mind, we have conducted two nonlinear analyses, which resulted in extracting 8 different parameters of RR intervals, described as below [23–27].



(a)



(b)

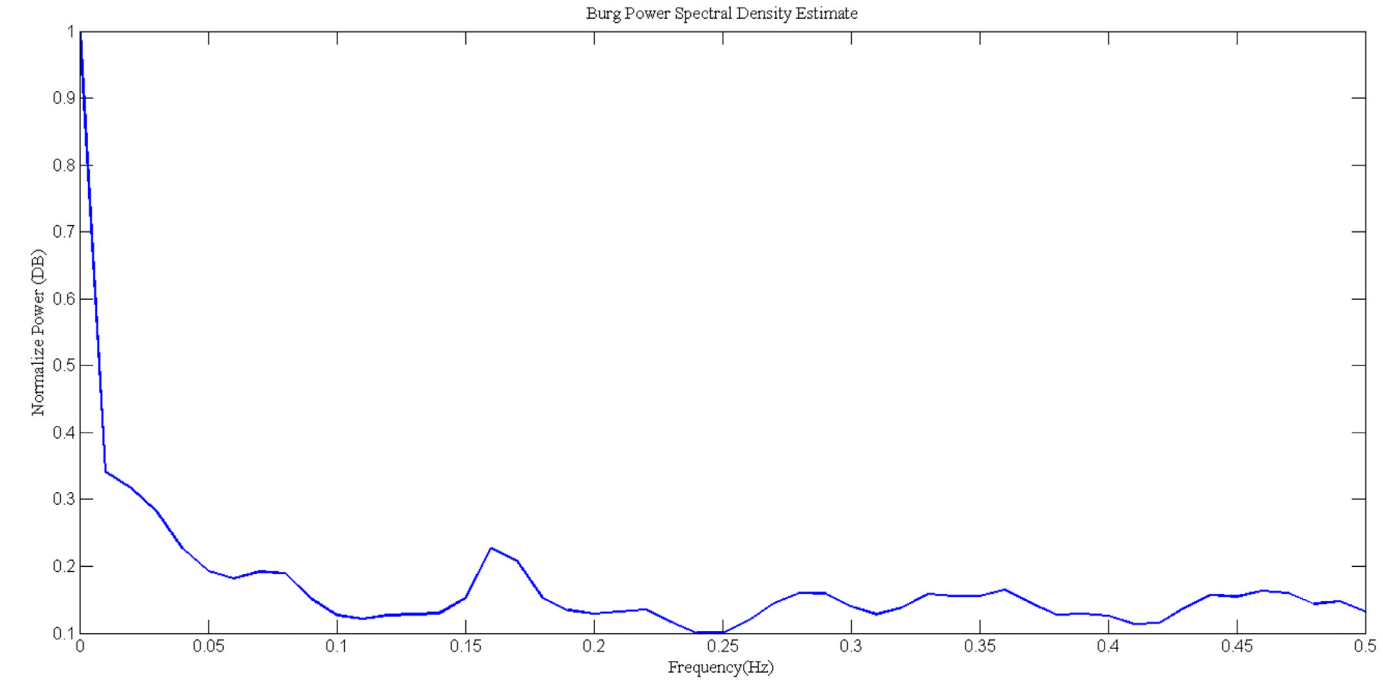
Fig. 3. The extracted HRV signal of 5-min ECG segment (a) distant from the onset of PAF (record n01 of AFPDB), and (b) immediately before the onset of PAF (record n04 of AFPDB).

2.5.1. Poincare' plot

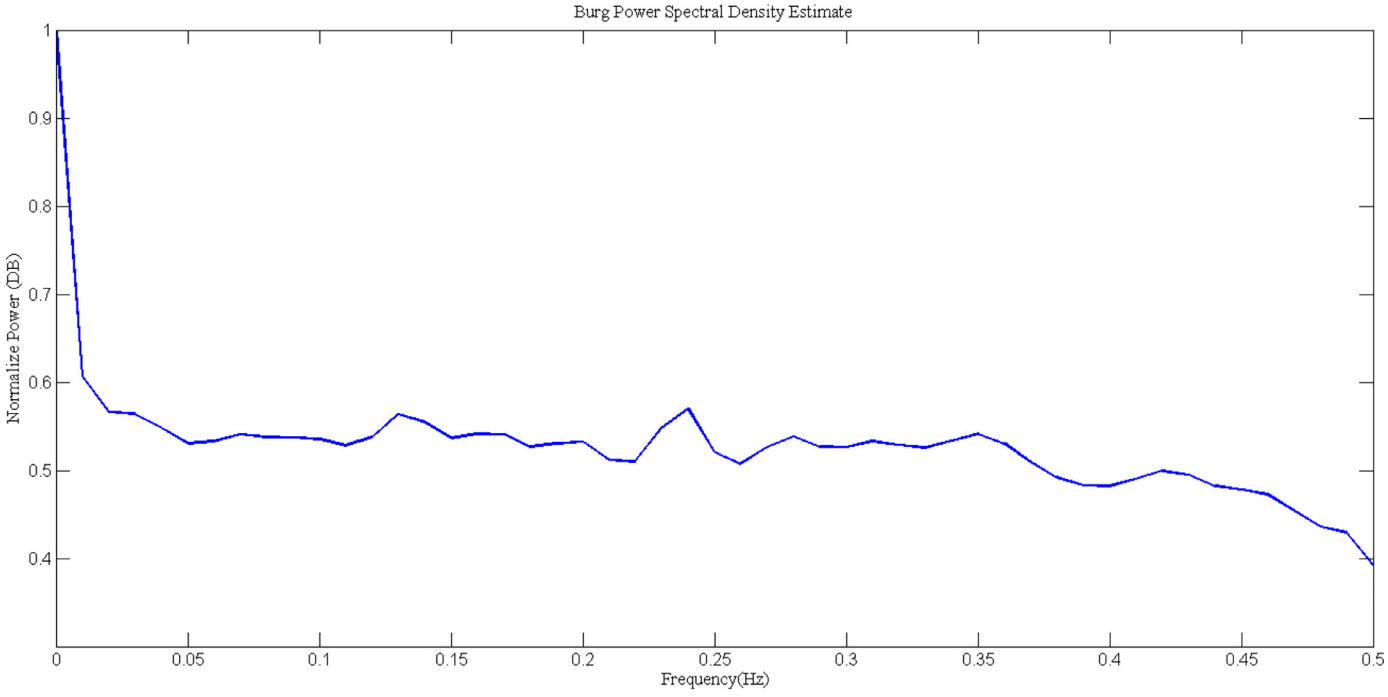
The Poincaré plot analysis is a geometrical and nonlinear quantitative method to assess the dynamics of HRV [54]. It is a two dimensional scatter plot in which RRn is plotted against RRn + 1 (n = 1,. . .,N-1 and N : length of RR interval time series). From the cloud of points of this diagram, the following linear indices are calculated as proposed by Brennan et al. [39]:

- SD1 [ms]: Standard deviation of the short-term RR interval variability (10)
- SD2 [ms]: Standard deviation of the long-term RR interval variability (11)
- SD1/SD2: Ratio of SD1 to SD2 [40].

$$SD1 = \sqrt{\text{variance}\left(\frac{NN_n - NN_{n+1}}{\sqrt{2}}\right)} \tag{15}$$



(a)



(b)

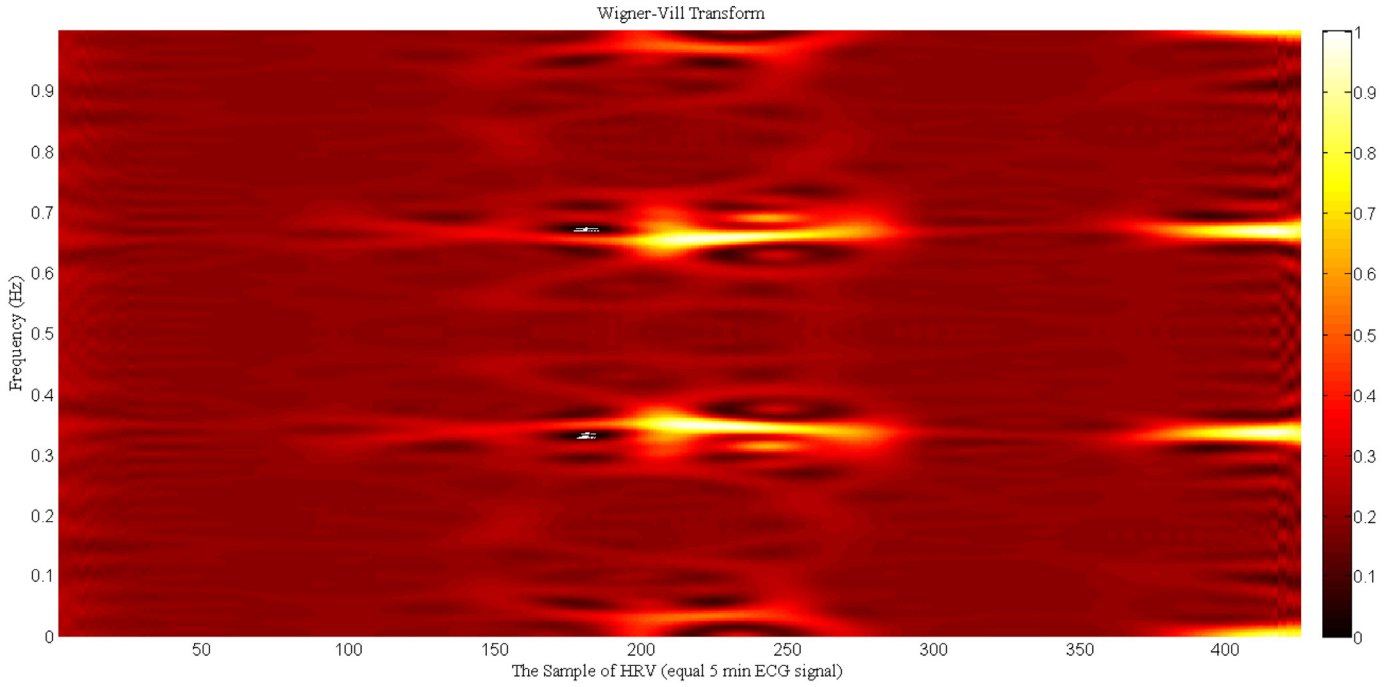
Fig. 4. The power spectral density of 5-min HRV segment (a) distant from the onset of PAF (record n01 of AFPDB), and (b) immediately before the onset of PAF (record t16 of AFPDB).

$$SD2 = \sqrt{\text{variance}\left(\frac{NN_n + NN_{n+1}}{\sqrt{2}}\right)} \quad (16)$$

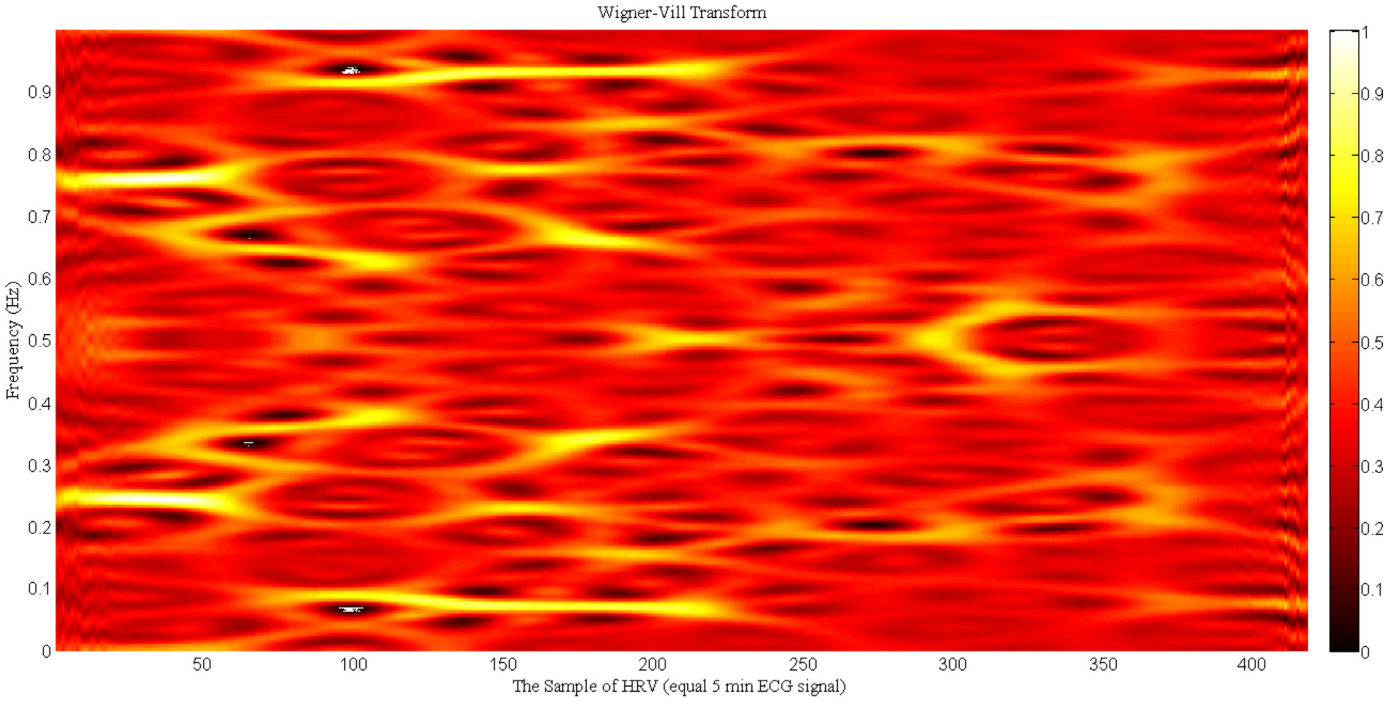
Fig. 6 shows the Poincaré plot of HRV of a non-PAF subject.

2.5.2. DFA analysis method

Detrended Fluctuation Analysis (DFA), introduced by Peng et al. [41], is a commonly employed method to quantify the fractal scaling properties of time series. To conduct such an analysis, the RR interval time series is integrated as y(k) (k = 1,…,N with N—length



(a)



(b)

Fig. 5. The Wigner Ville transform of the HRV signal (a) distant from the onset of PAF (b) immediately before the onset of PAF.
MAX_w, MIN_w, DIF_w, STD_w, E_{vlf}, E_{vf}, E_{HF}, F_{vlf}, F_{vf}, F_{HF}, W_{dif}.

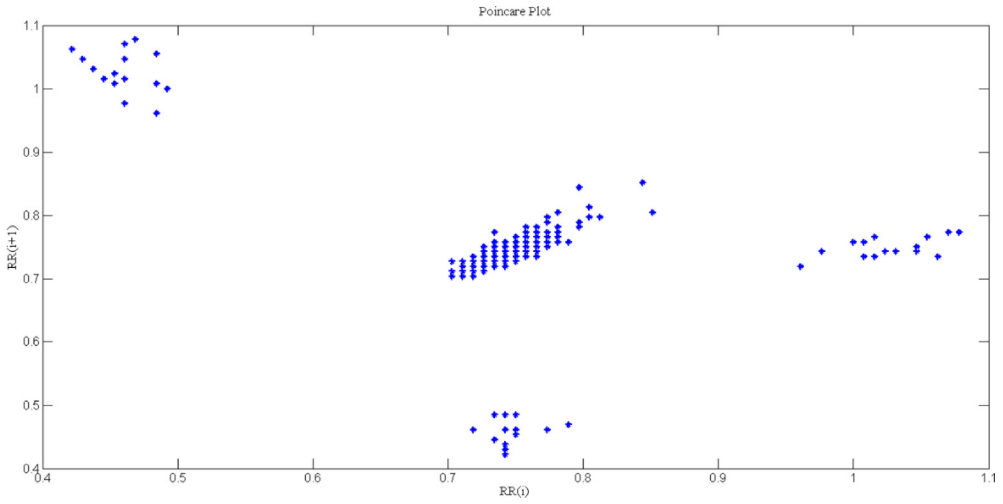


Fig. 6. Poincare plot of the non-PAF subject HRV.

of time series) and divided into equal and non-overlapping segments of length n. The local trend $y_n(k)$ in each segment is obtained by least-squares fitting and subtracted from y(k). At last, root-mean-square fluctuation values F(n) are calculated (12), and scaling exponents are estimated as the slope of the double-log plot of F(n) against n [40].

$$F(n) = \sqrt{\frac{1}{N}\sum_{K=1}^{N}[y(k) - y_n(k)]^2} \tag{17}$$

2.5.3. Recurrence plot

Recently, features like entropy and Lyapunov exponents have gained increasing attention as they are considered to be appropriate indices of time series. [42]. The RP is a tool for analysis that was introduced in the late 1980s [43]. RP is a visualization (or a graph) of a square matrix, in which the matrix elements correspond to those times at which a state of a dynamical system recurs (columns and rows correspond to a certain pair of times). Technically, the RP reveals all the times when the phase space trajectory of the dynamical system roughly visits the same area in the phase space. The RP is suitable for the analysis of physiological signals, which are often nonstationary [44].

2.6. Local subset feature selection

To reduce the feature space dimensionality, we have employed one of the recent and successful feature selection methods called Bandit [23,24] in which features are selected based on defined regions in the feature space.

Several methods have been proposed in the literature for feature selection most of which tend to select global features; i.e., they select a single subset of features for classification in the whole sample space. Nevertheless, there might be cases in which different subsets of features are the most informative ones for classification in different parts of the sample space [45]. The proposed method is premised on the idea of formulating the problem of feature subset selection as a sequential decision-making problem through the agency of feature trees. As mentioned, we are interested in partitioning the sample space into a number of localities and selecting features for each of them. In this regard, in order to profit from axis-aligned localities, as depicted in Fig. 7, it is assumed that the partitioning can be represented using a univariate binary decision tree whose benefit is to ensure that the representation of the localities will depend only on a limited number of features.

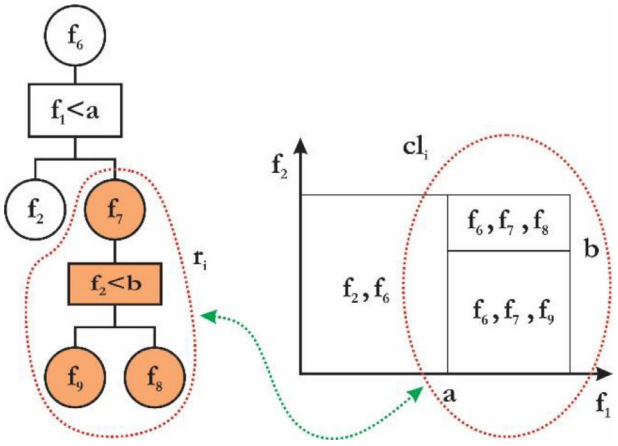


Fig. 7. An instance from different localities and features in tree representation.

A decision tree is typically comprised of two types of nodes; splitting nodes, and leaf nodes. The former signifies a split in the sample space and thus, have two children; the latter, on the other hand, is attributed to a single locality. It is worthwhile noting that the decision boundary in splitting nodes is determined based on the mean value of the corresponding feature over the training samples corresponding to that sub-tree.

In pursuance of local feature selection, the notion of decision tree has been developed to obtain a unified model, known as the feature tree, to represent both localities and corresponding selected features. In this regard, another type of node is put forward by the name of Feature Node which signifies a feature that is assigned to all of its descendant localities, and may have at most one child. Additionally, the concept of a Compound Locality is used to refer to each sub-tree that corresponds to a set of neighbor localities. This representation facilitates the selection of similar features for neighbor localities as they are known to be more likely to share a number of features. (see Fig. 7).

A feature node is represented by a circle with a single feature inside, and a splitting node is depicted by a rectangle containing a feature and a threshold. Localities are shown by leaves. The sub-tree r_i corresponds to the compound locality cl_i which consists of two single localities. It should be noted that the mutual features of neighbor localities are factored together in the parent feature node.

Applying a feature tree enables us to assign a training/ test sample to a unique leaf. This will be achieved through attributing

the sample to one of the descendants in the root repeatedly until it is assigned to a leaf. Therefore, in each locality we accumulate a subset of training samples and a subset of features, i.e. the set of feature nodes from the leaf to the root. In order to classify a test sample, we first assign it to a locality based on the feature tree and subsequently classify it in the locality using the corresponding features and training samples.

Selecting suitable local features is, in this context, a matter of paramount importance which requires us to adopt a criterion in order to compare different feature trees. We expect that the selected features make the classification more convenient. In this respect, the samples of different classes should be separable in the new space which is constructed by the selected features. One realization of this requirement is a space where each sample and its neighbors are likely to belong to the same class. Taking this into consideration, we assume S and ft to be the training set and the feature tree respectively. Given ft and an arbitrary sample x, we can find the subset of S that belongs to the same locality as x. Let $s \subset L(x, s, ft, k)$ be the k-nearest neighbors of x among the members of this subset. The score, i.e. fitness, of ft with respect to the training set S is calculated as follows

$$SCORE(ft) = \frac{1}{K.|S|} \sum_{x \in s} \sum_{y \varepsilon L(x,s,ft,k)} \begin{cases} 1 & label(y) = label(x) \\ 0 & otherwise \end{cases} \quad (18)$$

where label (.) gives the label, i.e. the class, of a sample.

It is noteworthy that, in this setting, each node of the tree is equivalent to a state for the Reinforcement Learning (RL) machine. This state consists of the sequence of nodes from the root to the current node in the tree. The RL agent selects an action at every state it arrives at. The decision expands the feature tree at its current node. Here, selecting an action for each node means choosing the node type (a feature node, a splitting node or a leaf) and the corresponding feature index. Therefore, the set of all possible actions in each state is

$$\text{Actions} = \{f_1, f_2, .., F_f, S_1, S_2, \ldots, S_F, T\}$$

in which F is the number of features, fi and si correspond to a feature node and a splitting node respectively, and T is the terminating action. The termination action finishes feature selection in the current node, leaving it as a leaf.

Each learning episode starts from the root node of the tree which, depending on the type of the node, leads to one of the below mentioned cases:

- A feature node f_i is selected: the node will be transformed to a feature node (with *i*-th feature). As a feature node has only one child, the next state would be the child node. The procedure then continues by selecting an action for the child.
- A splitting node s_i is selected: the node will be transformed to a splitting node (with i-th feature) that has two children. The procedure then runs independently for each of the two siblings to construct the corresponding sub-trees.
- Terminating action *T* is selected: the node will be transformed to a leaf with no children. Therefore, feature selection and locality formation will be terminated in the current branch.

The learning continues by generating another episode. The learning process ends when the number of episodes exceeds a predefined threshold. Fig. 8 presents the pseudo-code of one learning episode where *st* is the current state and *R* is the share of return corresponding to the action which is selected in *st*.

```
// This Function recursively construct a feature tree and evaluate it.
// In the first call of this function, the state is the root node of feature tree.

1:  [Returns R] = select_actions_recursively ( state St)

2:  a = select_action (St)                          // action selection
3:  if (a is the terminating action)                // if so, feature tree construction is terminated
4:     R = calculate_return (St)                    // this locality are classified, and return
5:  end                                             // function is calculated
6:  if (a is a feature node)
7:     R = select_action_recursively ( next_state (St,a) ) // feature tree continues to be
8:  end                                                    // constructed in the child

9:  if (a is a split node)                          // if so, feature tree continues to be
                                                    // constructed in the two children states
10:    R1 = select_action_recursively ( left_child (St,a) )
11:    R2 = select_action_recursively ( right_state (St,a) )

       // Return of each node is the CCR of evaluation samples in its sub-tree.
       // Therefore, the CCR of the two children should be weighted by the number
       // of the samples that they have,  i.e. nLeftsSaples and nRightSaple

12:    R = (R1*nLeftSamples + R2*nRightSamples) / (nLeftSamples + nRightSamples)
13: end
14: updateQValue (s , a , R)
15: Return R
16: END
```

Fig. 8. Pseudo code of the function as a description of the foundation of feature selection. It demonstrates feature selection during Reinforcement Learning.

2.7. Classification

In this work, we used four classifiers, the K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Multilayer Perceptron (MLP) and Mixture of Expert (ME) to differentiate between signals before PAF and distant from PAF. To evaluate the performance of the classifiers, a ten-fold cross-validation method is employed. The 56 ECG signals are divided into ten parts with the number of signals in each part being equal except for two or three groups. One part was used to test the classifier and the other nine parts were used to train the classifier. This process was repeated ten times for each different test set, and the average performance for accuracy, sensitivity, and specificity was calculated. In order to increase the accuracy, we repeated each of the ten processes fifteen times.

2.7.1. Multilayer perceptron (MLP)

This classifier makes use of a three-layer MLP with the error back propagation algorithm and variable learning rate. The input layer has the same number of nodes as the input's vector length for each interval time [46–48]. The output layer, on the other hand, contains one node accounting for a possibility of only 2 classes to be classified. All the possible combinations of the selected numbers of neurons in the hidden layer were selected and trained, leading to the optimized number being 4. It is worth mentioning that the training process has solely been conducted on the training data. Once the training error has dropped to a minimum, we move on to testing the network using the testing data. In fact, never has the data network observed the testing data when selecting the optimal architecture, which brings an improvement on the proper generalization of the network.

The output nodes and the hidden layer use a linear transfer function and a sigmoid function, respectively. Network training proceeds until the mean value is less than 0.01, or the number of training iterations reaches 1000.

2.7.2. K-Nearest neighbor

This classifier stores labeled feature vectors and calculates the minimum distance between stored and new feature vectors

[49,50]. The basic steps of the KNN algorithm are: (a) Computing the distances between all samples that have already been classified into clusters; (b) finding the k samples with the smallest distance values; and (c) approving new data. A new sample will be added (classified) to the largest cluster out of k selected samples. We tested the values of k from 1 to 15 to make a comparison with [51] and found that k = 5 gets the best results with this classifier. We used three k values (5, 10, 15) to reduce the complexity of tables.

2.7.3. Support vector machine

In this work, the SVM classifier is used for classification of episodes before PAF and distant from PAF. SVM is a machine-learning technique which has established itself as a powerful tool in many classification problems. Simply stated, the SVM identifies the best separating hyperplane (the plane with maximum margins) between the two classes of the training samples within the feature space by focusing on the training cases placed at the edge of the class descriptors. In this way, not only an optimal hyperplane is fitted, but also fewer training samples are effectively used; thus high classification accuracy is achieved with small training sets.

This classifier is a well-known supervised learning model that analyzes data used for classification and regression. The main idea of SVM is to map the input data from the N-dimensional input space, through some nonlinear mapping, to the M-dimensional feature space M > N, where the data classes can be linearly separated [32]. In other words, the SVM is an extension of nonlinear models of the generalized portrait algorithm based on statistical learning theory [52,53]. The goal of regression is to determine the best model from a set of models (named Estimating Functions) to approximate future values accurately. The generic support vector regression estimating function is:

$$f(x)=(w \cdot \Phi(x))+b \tag{19}$$

where $w \subset R_n$, $b \subset R$ and Φ is a nonlinear function that maps x into a higher dimensional space. W and b are the weight vector and bias, respectively. The weight vector (w) can be written as:

$$w = \sum_{i=1}^{L} \left(\propto_i - \propto_i^*\right) \tag{20}$$

By substituting Eq. (19) into Eq. (20), the generic equation can be rewritten as:

$$f(x) = \sum_{i=1}^{L} (\propto_i - \propto_i^*)(\Phi(x_i).\Phi(x)) + b \tag{21}$$

$$f(x) = \sum_{i=1}^{l} \left(\propto_i - \propto_i^*\right) k(xi.\ x) + b \tag{22}$$

In Eq. (21), the function k (x_i . x) = ($\Phi(x_i)$. $\Phi(x)$) is replaced with the dot product and is known as the kernel function and $\propto = (\propto 1, \propto 2, \ldots, \propto l)$ is the vector of nonnegative Lagrange multipliers.

The choice of kernel functions and kernel parameters depends mainly on the application. Among the useful kernel functions are radial basis functions (RBFs) and polynomial kernel functions. The formulas of these kernel functions are shown below respectively:

$$\left\{ \frac{-|x - xi|^2}{2\sigma^2} \right\} \tag{23}$$

$$[(x * x_i)+1]^d \tag{24}$$

Where "σ" and "d" are kernel width and order respectively which were experimentally defined to achieve the best classification result. In this work, RBFs and polynomial kernel functions were used with different sigma values (σ = 0.8, 1, 1.2) and orders (d = 1, 2, 3), respectively.

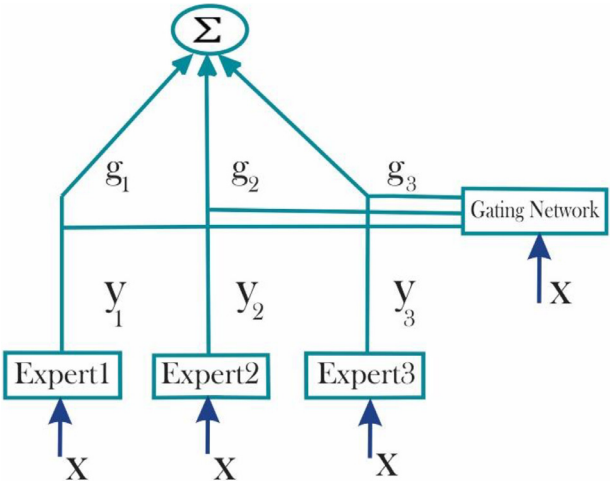


Fig. 9. Architecture of a Mixture of Expert Network.

2.8. Mixture of expert

Empirical studies verify that a given learning algorithm can outperform all others for a particular problem or for a specific subset of the input data. Be that as it may, finding a single expert to achieve the best results on the overall problem domain seems unlikely. To benefit from the different behaviors of the base classifiers, and to improve the performance in classification, there have recently been widespread interests in the use of model combination approaches for classification. They raise hopes that if a learner fails, the overall system is capable of capturing the information contained in all the learners and recovering the error. Among the combining methods, Mixture of Experts (ME) is a modular architecture of neural networks that has gained considerable attention over the past few years. Based on a Divide and Conquer (D&C) principle, ME typically comprises a gating network that partitions the input space into smaller problems and several expert networks that solve each subproblem [54,55].

Fig. 9 shows the architecture of an ME network. The gating network receives the vector x as input and produces scalar outputs that are partition of unity at each point in the input space. Each expert network produces an output vector for an input vector. The gating network provides linear combination coefficients as veridical probabilities for expert networks and, therefore, the final output of the ME architecture is a convex weighted sum of all the output vectors produced by expert networks. Let N be the number of expert networks in the ME architecture. All the expert networks are linear with a single output nonlinearity that is also referred to as "generalized linear". The i th expert network produces its output $o_i(x)$ as a generalized linear function of input x.

$$o_i(x) = f(W_i x) \tag{25}$$

Where W_i is a weight matrix and f(.) is a fixed continuous nonlinearity. The gating network is also generalized linear function, and its i th output, $g(x,v_i)$, is the multinomial logit or softmax function of intermediate variables ξi:

$$g(x, v_i) = \frac{e^{\xi i}}{\sum_{K=1}^{N} e^{\xi i}} \tag{26}$$

Where $\xi i = V_i^T X$ and v_i is a weight vector. The overall output of o(x) of the ME architecture is

$$o(x) = \sum_{K=1}^{N} g(x, v_k) o_k(x) \tag{27}$$

From a probabilistic standpoint, for an input-output pair (x,y), the values of g(v,x) are interpreted as the multinomial probabilities

associated with the decision that terminated in a regressive process that maps x to y. Once the decision has been made, resulting in a choice of regressive process i, the output y is chosen from a probability density p(y_x, Wi) in which Wi denotes the set of parameters or weight matrix of the i th expert network in the model. Thus, the total probability of generating y from x is the mixture of the probabilities of generating y from each component densities, where the mixing proportions are multinomial probabilities: [56]

$$P(Y|x, \Phi) = \sum_{K=1}^{N} g(x, v_k) P(Y|x, W_k) \quad (28)$$

Where the Φ is the set of all the parameters including both expert and gating network parameters. The probabilistic model allows for the ME architecture to be treated as maximum likelihood problem, therefore, we have applied an expectation maximization (EM) algorithm for adjusting the parameters of the architecture, which enables the ME model to be significantly effective in coping with complexity issues regarding non-trivial time series and classification complications [57–63]. In this framework, a number of relatively small expert networks can be used together with a gating network designed to divide the global classification task into simpler subtasks [56].

In the current study, having considered a number of commonly used artificial neural networks, we achieved the best network as the one with 12-20-1 input network and the expert network of 12-15-1. The applied MLP in the input network was formed in three layers using error back propagation algorithm with variable learning rate. The input layer has a number of nodes equal to the input vector length (12 nodes). The output layer consists of one node with its value being 0 or 1. We tried to optimize the neural network architecture through changing the number of hidden layer neurons which led us to choosing a three-layered neural network with 20 neurons in the hidden layer using the standard Sigmoid activation function. Network training continued until the mean square error became less than 0.01 or the number of training iterations reached 1000. Accordingly, the 12-15-1 structure was selected for the expert networks, accounting for the higher power of the input network, which in fact is the selector.

2.9. Evaluation

The ability of the proposed method for prediction of PAF is evaluated using Accuracy (AC), Sensitivity (SN), Specificity (SP) and Precision (P). In the following Eqs. 29–(32), TP refers to true positives (correctly predicted PAF), TN refers to true negatives (correctly predicted non-PAF), FN refers to false negatives (incorrectly predicted non-PAF) and FP refers to false positives (incorrectly predicted PAF).

Accuracy (AC): the ratio of correct predictions to the total predictions

$$AC = \frac{TP + TN}{TP + TN + FN + FP} \quad (29)$$

Sensitivity (SN): the ratio of true positives to the total positives.

$$SN = \frac{TP}{FN + TP} \quad (30)$$

Specificity (SP): the ratio of true negatives to the total negatives

$$SP = \frac{TN}{TN + FP} \quad (31)$$

Precision (P): the ratio of predicted positive cases that were correct

$$P = \frac{TP}{FP + TP} \quad (32)$$

Table 1
Results of the ANOVA test.

Feature	*P*-Value
Linear features	0.15
Time-frequency features	$0 \leq 0.0001$
Non-linear features	$0 \leq 0.0001$

Table 2
Performance of the KNN and SVM classifiers with different parameters.

Classifier	Sensitivity	Specificity	Accuracy
SVM-Poly (d = 1)	81.38%	97.11%	94.53%
SVM-Poly (d = 2)	83.25%	97.41%	92.23%
SVM-Poly (d = 3)	82.64%	97.40%	93.71%
SVM-RBF (σ = 0.8)	82.60%	98.78%	95.23%
SVM-RBF (σ = 1.0)	81.18%	98.86%	95.18%
SVM-RBF (σ = 1.2)	81.45%	97.62%	94.77%
KNN (K = 5)	77.27%	96.38%	92.38%
KNN (K = 10)	78.13%	96.22%	89.18%
KNN (K = 15)	76.45%	94.23%	87.65%

3. Result

We extracted 28 features from each episode: nine linear features, including 5 features from time domain and 4 from frequency domain, 11 time-frequency features, which contain time and frequency information simultaneously, and 8 nonlinear features, i.e. SD1, SD2, SD1/SD2, DFA features and sample entropy. Then, by applying local subset feature selection, 12 features have been selected. Our studies show that these selected features can be used as suitable markers for prediction of PAF. Fig. 4 illustrates the power spectral density, Wigner-Ville transform plot, and Poincaré plot of episodes before PAF and distant from PAF.

Table 1 depicts the power of the 12 selected features in discriminating episodes immediately before PAF and distant from the onset of PAF, which was evaluated by analysis of variance (ANOVA) on the learning set. According to this table, all of the time-frequency features and non-linear features had acceptable discriminant power, as was implied by the p-value. The results of ANOVA test in spectrum features also showed that LF and HF power differ significantly between episodes prior to PAF and distant from PAF onset, however, due to the high p-values of frequency peaks in LF and HF bands, these features could not distinguish between the two groups with statistical significance. Therefore, we used 10 features with statistically significant discrimination, excluding the frequency peaks in LF and HF bands from the total 12 aforementioned features to form the feature vector. Eventually, we used 50 ECG segments to train the SVM classifier and 56 segments to test the algorithm. The training and test sets belong to different subjects and there is no overlapping between the two sets.

To optimize the learning cost and the prediction performance, the SVM classifier parameters and kernel width σ must be chosen with more care and caution. To this end, we compared the performance of the networks by evaluating the error function using an independent validation set, and selected the network which has the smallest error with respect to the validation set. Since this procedure can itself lead to some overfitting to the validation set, the performance of the selected network was confirmed by test data set. In fact, we have divided the training data into train and validation sets with the proportions being 70% & 30% for training and validation. The optimum values of the parameters, chosen when the error on the validation dataset reaches a minimum, are achieved as 0.8 and 3 for σ and order (d), respectively.

Table 2 summarizes the performance of the two classifiers, i.e. SVM and KNN, having variable parameters. The features are used

Table 3
The statistical measures for validation by SVM, KNN and MLP classifiers.

	TP	FP	FN	TN
SVM	26	2	1	27
KNN	24	4	2	26
MLP	25	3	2	26
ME	27	1	0	28

Table 4
Results of the classifier performance in percentage, for 10-fold cross validation.

Classifier	Accuracy	Specificity	Sensitivity	Precision
SVM	94.64	93.10	96.29	92.85
KNN	89.28	86.66	92.30	85.71
MLP	91.07	89.65	92.59	89.28
ME	98.21	96.55	100	96.42

as input of Local Subset Feature Selection algorithm to reduce the number of original features and to enhance the performance of the proposed algorithm, then we use SVM and KNN classifiers for classification. We have evaluated the performance of the classifiers with different kernels and Ks, such as polynomials with order 1, 2, and 3 and RBFs with different sigma values ($\sigma = 0.8$, 1, 1.2) for the SVM classifier and we tested values of k from 1 to 15 and found that $k = 5$ gets the best results in the KNN classifier. We show that three k values (5, 10, 15) reduce the complexity of the tables. It is highlighted that the input of the classifiers is the HRV features that are extracted during the Local Feature Subset Selection phase.

To evaluate the performance of the proposed method, three measures are used based on Eqs. (29)–(32). If for example a PAF episode is correctly classified as the PAF episode, it is said that the episode is classified TP. On the other hand, if a non-PAF episode is classified as non-PAF episode, it is said that the episode is classified TN. Any non-PAF episode which is classified as a PAF episode by mistake will produce an FP; while any PAF episode which is mistakenly classified as a non-PAF episode will produce an FN result. The results of classification of test data for each class are summarized in Table 3. The obtained accuracy, sensitivity, specificity, and precision of the proposed method, as shown in Table 4, were achieved as 98.21%, 95.55%, 100% and 96.42%, respectively. The proposed methodology presents better results than the other existing approaches.

As is well known, the Sensitivity and Specificity of test (the false positive and false negative rates) alongside the incidence of the disease with P (PAF) are necessary values to help calculate other important quantities such as P (PAF|+).

At first, for our data set picking a threshold of 0.75 gives us the following results:

Then, the sensitivity and specificity values are computed for this particular decision rule. Fig. 10 manifests the value of FN, FP, TN, TP with the value of the threshold being 0.75. These calculations for MLP classifier are as follows:

$$\text{Sensitivity} = P(\text{Test}+|\text{Condition}+) = TP/(TP+FN) = 25(25+2) = 0.925$$

$$\text{Specificity} = P(\text{Test}-|\text{Condition}-) = TN/(FP+TN) = 26(3+26) = 0.896$$

After selecting the classifier, in order to evaluate the separability of features, the extracted features are compared to each other in both individual (linear, nonlinear and time-frequency) and optimal combinational mode. Table 5 demonstrates the separability

Table 5
The accuracy of MLP, SVM, KNN and ME classifiers with the selected features subsets (individually and in combinational) for PAF and non-PAF.

	SVM	KNN	MLP	EM
Linear	84.92%	76.83%	78.25%	82.43%
Time- Frequency	91.82%	78.13%	80.66%	95.38%
Non-Linear	89.23%	81.67%	82.15%	96.12%
Combinational	93.76%	92.38%	91.90%	98.21%

of linear, nonlinear and time-frequency features and also combinational mode for two classes, i.e. PAF and non-PAF. As can be seen in this table, combinational features are more capable of classification of classes which is why they have been used in this study as input features vector to predict PAF.

4. Discussion

In this work, a new approach to predict the onset of PAF is presented. First, we extracted a number of Linear, Time-Frequency (TF) and Nonlinear features. Successively, we reduced the feature space dimensionality by applying an improved machine learning method known as Time Local Subset Feature Selection. The Mixture of Expert is finally used for classification of episodes distant from and prior to PAF.

4.1. Common approaches in the literature

Conventionally, spectral analysis of the HRV signal is believed to be of great use in predicting PAF. This is mainly attributed to the increased power in LF and HF bands of HRV immediately before the PAF event [13]. Moreover, Poincaré plot evidently provides valuable insight into the dynamic of HRV signal, in the sense that SD1, SD2, and SD1/SD2, which quantify this plot, exhibit lower values in non- PAF episodes compared to PAF episodes.

The Complexity analysis of the HRV signal has also demonstrated its usefulness in estimating regularity of the signal. Studies have revealed a significant decrease in the complexity of R-R intervals and altered fractal properties in short-term R-R interval dynamics preceding PAF. It is considered to be a marker of both altered regulation of sinus node behavior and an increase of atrial firing from a single ectopic focus, which together can trigger the spontaneous onset of AF.

Complexity features such as approximate entropy (ApEn) and sample entropy (SmEn) have also exhibited smaller values in episodes preceding PAF in comparison with distant ones. [22,32,31].

Accordingly, the current study has elicited a set of these features (14 features) and used them for training an ME-based classifier that is able to identify which episode precedes the onset of PAF.

From a clinical standpoint, many attempts have been made to scrutinize the role of P wave in the prediction of AF after coronary artery surgery. Dilvaries et al. [64] listed maximum P-wave duration (P maximum) and the difference between the maximum and the minimum P-wave duration (P dispersion) as simple electrocardiographic markers derived from the 12-lead surface electrocardiogram that are significantly different between patients with idiopathic PAF and healthy subjects. With a P maximum value of 110 ms and a P dispersion value of 40 ms, they achieved a sensitivity of 88% and 83% and a specificity of 75% and 85%, respectively.

In a similar effort, having measured the duration, the amplitude, and the dispersion of P wave from 12-lead ECG of 120 patients, Chang et al [65] came to realize that the duration of P wave was far longer in the group with AF, exclusively in leads II, III, aVF

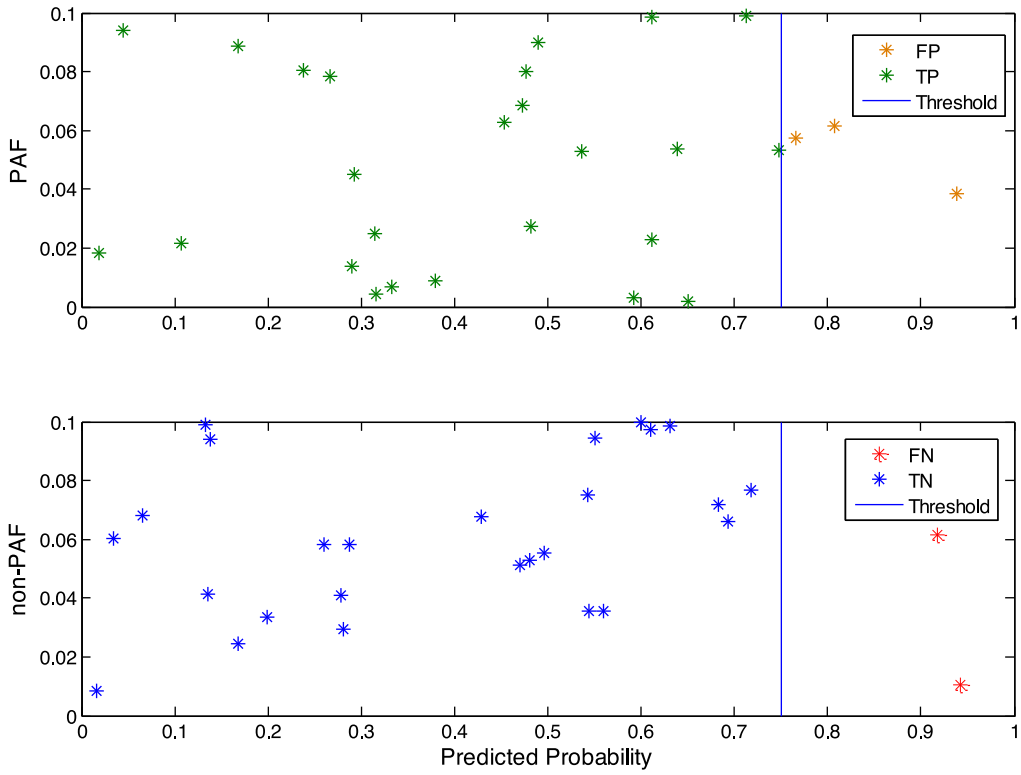


Fig. 10. The value of FN, FP, TN, TP with the prediction accuracy threshold value being 0.75. TP = 25, FP = 3, TN = 26, FN = 2.

and V2. However, they did not find a significant correlation between the P wave dispersion or amplitude and the occurrence of postoperative AF.

Budeus et al. [66] studied the P wave signal averaged ECG in 101 patients for prediction of atrial fibrillation after coronary artery bypass grafting (CABG). They found that patients with AF were older, had a longer filtered P-wave duration (FPD) and a lower mean square voltage of the last 20 ms (RMS 20) of the P wave; correspondingly, they suggested that the P-wave signal averaged ECG as well as an analysis of chemoreflexsensitivity can preoperatively predict AF after CABG, leading to a sensitivity of 78%, a specificity of 75% and a positive predictivity value of 64%.

The authors of [18,20,21] participated in Computers in Cardiology Challenge 2001 (PAF prediction challenge) among whom Zong et al. [18] used the number and timing of atrial premature complex (APC) as their main predictor and received the highest score at the time. Nevertheless, in a further attempt at achieving better results in 2004, Thong et al [19] proposed another method, a key component of which was an analysis of isolated PACs not followed by a regular R-R interval. They reported a sensitivity of 89% and a specificity of 91% for prediction the onset of PAF.

Chesnokov [32] in 2008 used complexity and spectral analysis of HRV, but they did not yield a promising sensitivity for prediction of PAF events.

4.2. Comparison with previous studies

To assess the performance of analytical methods in classification, we have compared the classification rates for nonlinear and TF features both separately and combined. The outcomes are then compared to the previously reported results which have applied similar methods of evaluation.

When compared to Thong et al. [19] method, a conspicuous observation to emerge, aside from a better performance, is that the algorithm proposed in the present study does not need the both records of a subject to identify imminent episodes of PAF.

We also simulated and evaluated Boone et al. [33] method with AFPDB database in order to enable a meaningful comparison between our algorithm and theirs. Remarkably, the predictive accuracy improved from 87.7% to 98.21%, which provides compelling evidence that an optimal combination of nonlinear and time-frequency processing methods followed by ME classification simply excels at genetic algorithm methods. The same logic accounts for the dramatic improvement of separability in Mohebbi et al. [31] study.

Furthermore, the ME classifier is shown to have offered better performance in classification of the episodes than other commonly used classifiers such as SVM, MLP and k-NN.

In addition to the mentioned experimental researches, several methods have been developed for automatic prediction of PAF using ECG signal. Table 4 summarizes a number of these methods, all of which have made use of AFPDB database, along with their reported results with reference to the commonly used measures of sensitivity, and specificity.

The algorithm proposed in the present study demonstrates outstanding performance in terms of both sensitivity and specificity, outperforming the existing approaches, as shown in Table 4.

4.3. Major findings

While there appears to be no distinct difference between ECG signals that are leading to PAF and those that are distant from PAF, HRV signal contains critical information within its nonlinear nature. Accordingly, to benefit from time domain and frequency features at the same time, we have made use of the time-frequency analyses, particularly Wigner-Ville transform, which ensure both classic and nonlinear methods are included and applied in a complementary fashion. The results denote that combining the most

Table 6
A comparison of the presented methods in other papers and the current method, for predicting the onset of paroxysmal atrial fibrillation.

Author and year	Signal length (Min)	Feature extraction	Performance evaluation method	SEN (%)	SPE(%)	ACC (%)
Boon et al. 2016 [29]	15	HRV features	10-fold CV	77.4	81.1	79.3
Boon et al. 2018 [33]	10	HRV features	10-fold CV	58.5	81.1	68.9
Yang and Yin, 2001 [21]	10	HRV based footprint analysis	Single Hold	_	_	57.0
Hickey and Heneghan, 2002 [28]	10	Spectral based HRV features	5-fold CV	53.0	80.0	70.0
Boon et al. 2018 [33]	5	Spectral based HRV features	5-fold CV	51.0	79.0	68.0
Boon et al. 2018 [33]	5	HRV features	10-fold CV	86.8	88.7	87.7
Zong et al. 2001 (18)	30	Number and timing of PACs	Single Hold	79	_	80.0
Hickey et al. 2002 [28]	30	PACs detection and spectral based HRV features	5-fold CV	79.0	72.0	72.0
Thong et al. 2004 [19]	30	PACs analysis	Single Hold	89.0	91.0	90.0
Costin et al. 2013 [30]	30	HRV features and morphological variability of QRS complexes	Single Hold	89.3	89.4	89.4
Mohebbi et al. 2012 [31]	30	HRV features	Single Hold	96.2	93.1	94.5
Cheskonov 2008 [32]	30	HRV based spectral features	Single Hold	72.7	88.2	80.0
Lynn and Chiang [20]	30	HRV based return map and Poincare Plot features	Single Hold	_	_	64.0
The proposed method	5	HRV features	10-fold CV	100	95.5	98.2

Note: CV: Cross validation

informative features extracted from different domains along with performing the local subset feature selection method and using the ME classifier lead to a more accurate predictor of PAF onset.

From a clinical perspective, the achieved results of PAF prediction have further strengthened our confidence in enabling timely treatment and increasing the survival rate.

4.4. Study limitations

It is worthwhile noting that the results presented in this paper, similar to those of other studies listed in Table 6, are restricted by the lack of a prospective head-to-head evaluation with clinically derived, real world data.

5. Conclusions

In this paper, we have proposed an effective machine learning based methodology to classify ECG with the aim of predicting the onset of Paroxysmal Atrial Fibrillation. The prediction performance of this algorithm has been shown to be superior to the previously developed methods in terms of both sensitivity and specificity, with their obtained values being 100% and 95.5% respectively.

We strongly believe that what makes this method particularly effective is that it uses an optimal combination of — as opposed to being restricted to — linear, time-frequency, and nonlinear features which are cautiously chosen to represent the frequency content, the phase relations between frequency components, and non-linear dynamics of the HRV signal before the occurrence of PAF, respectively. The results highlight that there are statistically significant differences related to these features in PAF episodes and non-PAF episodes, thus confirming that they are in fact potent and reliable predictors. In addition, Mixture of Experts, having shown to be a precise classifier, contributes greatly to optimizing the performance of this method.

Finally, it is our belief that the presented approach would lend itself well for use by doctors and clinicians, as it makes possible to develop an early-detection system for the idiopathic onset of PAF, which can alert a patient prior to the occurrence of the event.

Conflict of interest statement

There are no conflicts of interest please just include the text

References

[1] W.M. Feinberg, J.L. Blackshear, A. Laupacis, R. Kronmal, R. Hart, Prevalence, age distribution, and gender, Arch. Intern. Med. 155 (1995) 469–473.
[2] A.S. Go, E.M. Hylek, K.A. Phillips, Y. Chang, L.E. Henault, J.V. Selby, D.E. Singer, Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study, Jama 285 (2001) 2370–2375.
[3] P.A. Wolf, R.D. Abbott, W.B. Kannel, Atrial fibrillation as an independent risk factor for stroke: the Framingham Study, Stroke 22 (1991) 983–988.
[4] E.J. Benjamin, P.A. Wolf, R.B. D'Agostino, H. Silbershatz, W.B. Kannel, D. Levy, Impact of atrial fibrillation on the risk of death: the Framingham heart study, Circulation 98 (1998) 946–952.
[5] A.M. Patel, D.C. Westveer, K.C. Man, J.R. Stewart, H.I. Frumin, Treatment of underlying atrial fibrillation: paced rhythm obscures recognition, J. Am. Coll. Cardiol. 36 (2000) 784–787.
[6] A. Filippi, G. Bettoncelli, A. Zaninelli, Detected atrial fibrillation in north Italy: rates, calculated stroke risk and proportion of patients receiving thrombo-prophylaxis, Family Pract. 17 (2000) 337–339.
[7] M. Sudlow, H. Rodgers, R.A. Kenny, R. Thomson, Population based study of use of anticoagulants among patients with atrial fibrillation in the community, BMJ 314 (1997) 1529.
[8] W.B. Kannel, P.A. Wolf, E.J. Benjamin, D. Levy, Prevalence, incidence, prognosis, and predisposing conditions for atrial fibrillation: population-based estimates 1, Am. J. Cardiol. 82 (1998) 2N–9N.
[9] I. Savelieva, A. Camm, Silent atrial fibrillation—another Pandora's box, Pacing Clin. Electrophysiol. 23 (2000) 145–148.
[10] S.M. Al-Khatib, W.E. Wilkinson, L.L. Sanders, E.A. McCarthy, E.L. Pritchett, Observations on the transition from intermittent to permanent atrial fibrillation, Am. Heart J. 140 (2000) 142–145.
[11] V. Markides, R.J. Schilling, Atrial fibrillation: classification, pathophysiology, mechanisms and drug treatment, Heart 89 (2003) 939–943.
[12] N. Takahashi, A. SEKI, K. Imataka, J. Fujii, Clinical features of paroxysmal atrial fibrillation, Jpn. Heart J. 22 (1981) 143–149.
[13] Y. Chesnokov, A. Holden, H. Zhang, Screening patients with paroxysmal atrial fibrillation (PAF) from non-PAF heart rhythm using HRV data analysis, Comput. Cardiol. 2007 (2007) 459–462 IEEE.
[14] E.N. Prystowsky, Management of atrial fibrillation: therapeutic options and clinical decisions, Am. J. Cardiol. 85 (2000) 3–11.
[15] D. Amar, H. Zhang, D.H. Leung, N. Roistacher, A.H. Kadish, Older age is the strongest predictor of postoperative atrial fibrillation, Anesthesiology 96 (2002) 352–356.

[16] A. Ruigómez, S. Johansson, M.-A. Wallander, L.A.G.a. Rodrıǵuez, Incidence of chronic atrial fibrillation in general practice and its treatment pattern, J. Clin. Epidemiol. 55 (2002) 358–363.
[17] C. Kolb, S. Nürnberger, G. Ndrepepa, B. Zrenner, A. Schömig, C. Schmitt, Modes of initiation of paroxysmal atrial fibrillation from analysis of spontaneously occurring episodes using a 12-lead Holter monitoring system, Am. J. Cardiol. 88 (2001) 853–857.
[18] W. Zong, R. Mukkamala, R. Mark, A methodology for predicting paroxysmal atrial fibrillation based on ECG arrhythmia feature analysis, Comput. Cardiol. 2001 (2001) 125–128 IEEE.
[19] T. Thong, J. McNames, M. Aboy, B. Goldstein, Prediction of paroxysmal atrial fibrillation by analysis of atrial premature complexes, IEEE Trans. Biomed. Eng. 51 (2004) 561–569.
[20] K. Lynn, H. Chiang, A two-stage solution algorithm for paroxysmal atrial fibrillation prediction, Comput. Cardiol. 2001 (2001) 405–407 IEEE.
[21] S. Vikman, T.H. Mäkikallio, S. Yli-Mäyry, S. Pikkujämsä, A.-M. Koivisto, P. Reinikainen, K.J. Airaksinen, H.V. Huikuri, Altered complexity and correlation properties of RR interval dynamics before the spontaneous onset of paroxysmal atrial fibrillation, Circulation 100 (1999) 2079–2084.
[22] A. Yang, H. Yin, Prediction of paroxysmal atrial fibrillation by footprint analysis, Comput. Cardiol. 2001 (2001) 401–404 IEEE.
[23] E. Ebrahimzadeh, M.S. Manuchehri, S. Amoozegar, B.N. Araabi, H. Soltanian-Zadeh, A time local subset feature selection for prediction of sudden cardiac death from ECG signal, Med. Biol. Eng. Comput. (2017) 1–18.
[24] E. Ebrahimzadeh, B.N. Araabi, A novel approach to predict sudden cardiac death using local feature selection and mixture of expert, Comput. Intell. Electr. Eng. 7 (2016) 15–32.
[25] E. Ebrahimzadeh, M. Pooyan, Early detection of sudden cardiac death by using classical linear techniques and time-frequency methods on electrocardiogram signals, J. Biomed. Sci. Eng. 4 (2011) 699.
[26] E. Ebrahimzadeh, M. Pooyan, Prediction of sudden cardiac death (SCD) by using time-frequency domain methods and nonlinear analysis from ECG signals, Comput. Intell. Electr. Eng. 3 (2013) 15–26.
[27] E. Ebrahimzadeh, M. Pooyan, A. Bijar, A novel approach to predict sudden cardiac death (SCD) using nonlinear and time-frequency analyses from HRV signals, PloS One 9 (2014) e81896.
[28] B. Hickey, C. Heneghan, Screening for paroxysmal atrial fibrillation using atrial premature contractions and spectral measures, Comput. Cardiol. 2002 (2002) 217–220 IEEE.
[29] K. Boon, M. Khalil-Hani, M. Malarvili, C. Sia, Paroxysmal atrial fibrillation prediction method with shorter HRV sequences, Comput. Methods Prog. Biomed. 134 (2016) 187–196.
[30] H. Costin, C. Rotariu, A. Păsăricǎ, Atrial fibrillation onset prediction using variability of ECG signals, advanced topics in electrical engineering (ATEE), in: 2013 8th International Symposium on, IEEE, 2013, pp. 1–4.
[31] M. Mohebbi, H. Ghassemian, Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal, Comput. Methods Prog. Biomed. 105 (2012) 40–49.
[32] Y.V. Chesnokov, Complexity and spectral analysis of the heart rate variability dynamics for distant prediction of paroxysmal atrial fibrillation with artificial intelligence methods, Artif. Intell. Med. 43 (2008) 151–165.
[33] K. Boon, M. Khalil-Hani, M. Malarvili, Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III, Comput. Methods Prog. Biomed. 153 (2018) 171–184.
[34] P hysionet AFPDB database, http://www.physionet.org/physiobank/database/afpdb.
[35] G.D. Clifford, F. Azuaje, P. Mcsharry, ECG statistics, noise, artifacts, and missing data, Adv. Methods Tools for ECG Data Anal. 6 (2006) 18.
[36] R. Acharya, S.M. Krishnan, J.A. Spaan, J.S. Suri, Advances in Cardiac Signal Processing, Springer, 2007.
[37] T.F.o.t.E.S.o. Cardiology, Heart rate variability, standards of measurement, Physiol. Int. Clin. Use Circ. 93 (1996) 1043–1065.
[38] E. Ebrahimzadeh, S.M. Alavi, A. Bijar, A. Pakkhesal, A novel approach for detection of deception using smoothed pseudo Wigner–Ville distribution (SPWVD), J. Biomed. Sci. Eng. 6 (2013) 8.
[39] M. Brennan, M. Palaniswami, P. Kamen, Do existing measures of Poincare plot geometry reflect nonlinear features of heart rate variability, IEEE Trans. Biomed. Eng. 48 (2001) 1342–1347.
[40] A. Voss, R. Schroeder, A. Heitmann, A. Peters, S. Perz, Short-term heart rate variability—influence of gender and age in healthy subjects, PloS one 10 (2015) e0118308.
[41] C.K. Peng, S. Havlin, H.E. Stanley, A.L. Goldberger, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, Chaos 5 (1995) 82–87.
[42] J.-P. Eckmann, D. Ruelle, Ergodic theory of chaos and strange attractors, in: The Theory of Chaotic Attractors, Springer, 1985, pp. 273–312.
[43] J.-P. Eckmann, S.O. Kamphorst, D. Ruelle, Recurrence plots of dynamical systems, EPL (Europhysics Letters) 4 (1987) 973.
[44] R. Sun, Y. Wang, Predicting termination of atrial fibrillation based on the structure and quantification of the recurrence plot, Med. Eng. Phys. 30 (2008) 1105–1111.
[45] M.-H.Z. Ashtiani, M.N. Ahmadabadi, B.N. Araabi, Bandit-based local feature subset selection, Neurocomputing 138 (2014) 371–382.
[46] E. Ebrahimzadeh, S.M Alavi, F. Samsami khodadad, Implementation and designing of lie-detection system based on electroencephalography (EEG), Ann. Mil. Health Sci. Res. 11 (2013) 20–26.
[47] M. Nikravan, E. Ebrahimzadeh, M.R. Izadi, M. Mikaeili, Toward a computer aided diagnosis system for lumbar disc herniation disease based on mr images analysis, Biomed. Eng. 28 (2016) 1650042.
[48] E. Ebrahimzadeh, M. Pooyan, S. Jahani, A. Bijar, S.K. Setaredan, ECG signals noise removal: selection and optimization of the best adaptive filtering algorithm based on various algorithms comparison, Biomed. Eng. 27 (2015) 1550038.
[49] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. (CSUR) 34 (2002) 1–47.
[50] D.T. Larose, Discovering Knowledge in data: an Introduction to Data Mining, John Wiley & Sons, 2014.
[51] U.R. Acharya, H. Fujita, V.K. Sudarshan, V.S. Sree, L.W.J. Eugene, D.N. Ghista, R. San Tan, An integrated index for detection of sudden cardiac death using discrete wavelet transform and nonlinear features, Knowl. Based Syst. 83 (2015) 149–158.
[52] J.A. Suykens, T. Van Gestel, B. De Moor, J. Vandewalle, Basic methods of least squares support vector machines, Least Squares Support Vector Mach. (2002) 71–116.
[53] E.E. Osuna, Support vector machines: hraining and applications, Massachusetts Inst. Technol. (1998).
[54] S. Masoudnia, R. Ebrahimpour, Mixture of experts: a literature survey, Artif. Intell. Rev. 42 (2014) 275–293.
[55] P. Moerland, Some methods for training mixtures of experts, IDIAP (1997).
[56] E.D. Übeyli, İ. Güler, Features extracted by eigenvector methods for detecting variability of EEG signals, Pattern Recognit. Lett. 28 (2007) 592–603.
[57] S.E. Yuksel, J.N. Wilson, P.D. Gader, Twenty years of mixture of experts, IEEE Trans. Neural Netw. Learn. Syst. 23 (2012) 1177–1193.
[58] C.A. Lima, A.L. Coelho, F.J. Von Zuben, Hybridizing mixtures of experts with support vector machines: investigation into nonlinear dynamic systems identification, Inf. Sci. 177 (2007) 2049–2074.
[59] S. Amoozegar, M. Pooyan, E. Ebrahimzadeh, Classification of brain signals in normal subjects and patients with epilepsy using mixture of experts, Comput. Intell. Electr. Eng. 4 (2013) 1–8.
[60] E. Ebrahimzadeh, F. Fayaz, F. Ahmadi, M.R. Dolatabad, Linear and nonlinear analyses for detection of sudden cardiac death (SCD) using ECG and HRV signals, Trends Med. Res. 1 (1) (2018) 1–8.
[61] E. Ebrahimzadeh, F. Fayaz, F. Ahmadi, M. Nikravan, A machine learning-based method in order to diagnose lumbar disc herniation disease by MR image processing, MedLife Open Access 1 (2018) 1–10.
[62] E. Ebrahimzadeh, F. Fayaz, F. Ahmadi, M.R. Dolatabad, Linear and nonlinear analyses for detection of sudden cardiac death (SCD) using ECG and HRV signals, Trends Med. Res. 1 (1) (2018), doi:10.15761/TR.1000105.
[63] E. Ebrahimzadeh, F. Fayaz, F. Ahmadi, M. Nikravan, A machine learning-based method in order to diagnose lumbar disc herniation disease by MR image processing, MedLife Open Access 1 (1) (2018).
[64] P.E. Dilaveris, E.J. Gialafos, S.K. Sideris, A.M. Theopistou, G.K. Andrikopoulos, M. Kyriakidis, J.E. Gialafos, P.K. Toutouzas, Simple electrocardiographic markers for the prediction of paroxysmal idiopathic atrial fibrillation, Am. Heart J. 135 (1998) 733–738.
[65] C.-M. Chang, S.-H. Lee, M.-J. Lu, C.-H. Lin, H.-H. Chao, J.-J. Cheng, P. Kuan, C.-R. Hung, The role of P wave in prediction of atrial fibrillation after coronary artery surgery, Int. J. Cardiol. 68 (1999) 303–308.
[66] M. Budeus, M. Hennersdorf, S. Röhlen, S. Schnitzler, O. Felix, K. Reimert, P. Feindt, E. Gams, H. Wieneke, S. Sack, Prediction of atrial fibrillation after coronary artery bypass grafting: the role of chemoreflexsensitivity and P wave signal averaged ECG, Int. J. Cardiol. 106 (2006) 67–74.