Heliyon 7 (2021) e08244

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Methodology for the prediction of paroxysmal atrial fibrillation based on heart rate variability feature analysis

Henry Castro [a,b,*], Juan D. Garcia-Racines [b], Alvaro Bernal-Norena [b]

[a] Universidad Santiago de Cali, Calle 5 No.62-00 Cali, Colombia
[b] Universidad del Valle, Calle 13 No. 100-00 Cali, Colombia

ARTICLE INFO

Keywords:
HRV
PAF prediction
Paroxysmal atrial fibrillation
Recursive feature elimination
Machine learning

ABSTRACT

Atrial fibrillation (AF) is the most clinically diagnosed arrhythmia, as its prevalence increases with age, and its initial stage is paroxysmal atrial fibrillation (PAF). This pathology usually triggers hemodynamic disorders that can generate cerebrovascular accidents (CVA), causing morbidity and even death. The aim of this study is to predict the occurrence of PAF episodes in order to take precautions to prevent PAF episodes. The PhysioNet AFPDB prediction database was used to extract 77 heart rate variability (HRV) features using time domain, geometrical analysis, Poincaré plot, nonlinear analysis, detrended fluctuation analysis, autoregressive modeling, fast Fourier transform (FFT), Lomb-Scargle periodogram, wavelet packet transform (WPT) and bispectrum measurements. The number of features was reduced using the near-zero value, correlation, and recursive feature elimination (RFE) methods for time windows of 1, 2, 5, 10, and 30 min. Feature selection was performed using backwards selection, genetic algorithm, analysis of variance (ANOVA), and non-dominated sorting genetic algorithm (NSGA-III) methods, and then random forest, conditional random forest, k-nearest neighbor (KNN), and support vector machine (SVM) classification algorithms were applied and evaluated using 10-fold cross-validation. The proposed method achieved a precision of 93.24% with a 5-minute window and 89.21% with a 2-minute window, improving performance in predicting PAF when compared with similar studies in the literature.

1. Introduction

The analysis of heart rate variability (HRV) time series is of utmost importance from a clinical point of view due to its high correlation with the autonomic nervous system (ANS) [1]. HRV measurement is an early predictive tool to detect cardiovascular diseases. The indicators of the progression of paroxysmal atrial fibrillation (PAF) to persistent or permanent PAF have not been fully identified; therefore, detecting atrial fibrillation (AF) in its early form is important to avoid the risks of stroke, heart failure, and/or mortality [2]. In its initial stage, PAF complications can be avoided if they are predicted early [3].

In 30-minute electrocardiogram (ECG) recordings, premature ventricular contraction (PVC) is an important feature that indicates the future appearance of PAF [4]. Furthermore, PAF appearance is linked to a considerable increase in the number of atrial and ventricular ectopic beats [5]. Based on this information, previous works have detected an early estimate of PAF using the following features:

A time-domain analysis distinguishes two types of HRV indices: fast beat-to-beat indices and slower fluctuation indices. Both indices are calculated from RR or NN intervals in a chosen time window [6].

A Poincaré plot of the "width" of the graph is a measure of the activity of the parasympathetic nervous system, and this method allows for the immediate recognition of ectopic beats [7, 8].

A Lomb-Scargle periodogram is used to estimate the power spectral density (PSD) of an HRV signal. Spectrum characteristics can discriminate between the sympathetic and parasympathetic content of the HRV signal, which is affected before PAF attacks [6]. It is generally accepted that the spectral power in the high frequency (HF) band (0.15–0.4 Hz) of the HRV signal reflects respiratory sinus arrhythmia (RSA) and thus cardiac vagal activity. On the other hand, the low-frequency band (LF) (0.04–0.15 Hz) is related to the control of baroreceptors and is mediated by both the vagal and sympathetic systems [9].

In a geometrical method, the triangular interpolation of NN interval (TINN) metrics and HRV index generally reflect HRV and are more influenced by lower frequencies than by high frequencies [1].

* Corresponding author.
E-mail address: hecastro1@gmail.com (H. Castro).

https://doi.org/10.1016/j.heliyon.2021.e08244
Received 3 June 2021; Received in revised form 11 August 2021; Accepted 20 October 2021

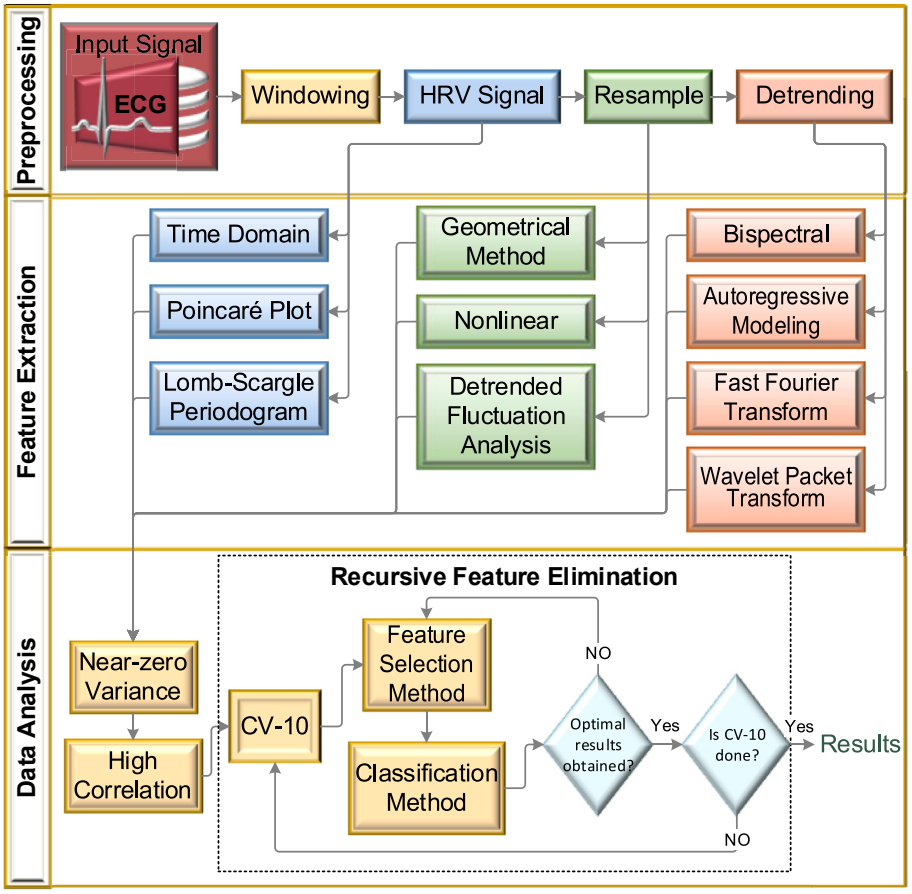

Figure 1. Research method overview.

Nonlinear analysis allows for evaluating the functioning of the cardiovascular system and discriminating PAF events by measuring the regularity of the HRV signal through the entropy per sample [9, 10, 11].

Detrended fluctuation analysis is used to quantify the fractal scale properties of short-interval RR signals, and the fluctuations are related to a scaling exponent (or self-similarity factor), $\propto$. $\propto$ can be seen as an indicator of the "roughness" of the original time series: the higher the value of $\propto$, the smoother the time series will be [12]. For normal subjects (healthy young people) $\propto$ is closer to 1, and this value falls in different ranges for various types of cardiac abnormalities [13].

Bispectral analysis is a technique used to reveal the time-phased relationships between noisy interacting oscillators, and it has been used to study the nature of the coupling between cardiac and respiratory activity [9, 14].

Autoregressive modeling is used to classify normal sinus rhythm (NSR) and various cardiac arrhythmias, including premature atrial contraction (PAC). Autoregressive (AR) coefficients were calculated using the Burg algorithm, and the AR modeling results showed that an order of sixteen was sufficient to model the HRV signals [15].

In fast Fourier transform (FFT), the HRV signal can be analyzed using different higher-order spectra (known as polyspectrals), which are spectral representations of higher-order moments or accumulations of a signal. A time-dependent spectral analysis of HRV was found to be valuable in explaining patterns of heart rate control during reperfusion [10].

Wavelet packet transform (WPT) is a useful method for R-R interval analysis given that it highlights time-dependent changes in the frequency spectrum [16]. WPT applies low-pass and high pass filters determined by a mother wavelet, this process yields a set of packages each of which describes a specific sub-band of the spectrum. Consequently, it is important to choose an appropriate mother wavelet function. According to [17], Daubechies wavelet functions are the most suitable to be used on ECG and HRV signals.

In addition, different techniques have been described for the prediction of PAF from technical to clinical points of view. The computers in cardiology (CinC) Challenge 2001 by PhysioNet obtained an accuracy of 82% [18]. Thong et al. [11] obtained a sensitivity and specificity of 84% and 88%, respectively, by analyzing premature atrial complexes (which trigger 93% of PAF episodes). Boon et al [19] achieved an accuracy of 87.7% with a window length of 5 min. Chazal et al [20], using a window length of 10 min, achieved an accuracy of 90.4%. Mohebbi et al [21] used a 30-minute window of length for the accuracy of PAF with an accuracy of 92.86%.

This paper defines a methodology to find an optimal set of HRV features, a classifier, and a validator to create a robust system that allows predicting the appearance of a PAF event with a high degree of precision.

2. Research method

In this research the methodology was developed using RStudio software, PBC V1.3.1093 and MatLab R2020a.

The proposed methodology is divided into 3 main stages: preprocessing, feature extraction, and data analysis. In the first stage, the extraction and preprocessing of the HRV signal are carried out from the ECG signal. In the second stage, 10 different methods are used to extract 77 HRV features. In the last stage, these features are analyzed and selected until the optimal combination is found to predict PAF.

In Figure 1, the general scheme of the proposed methodology is shown, Which consists of three main stages:

1. Preprocessing: Where an HRV signal, extracted from an ECG signal, it is resampled and its trend is removed.

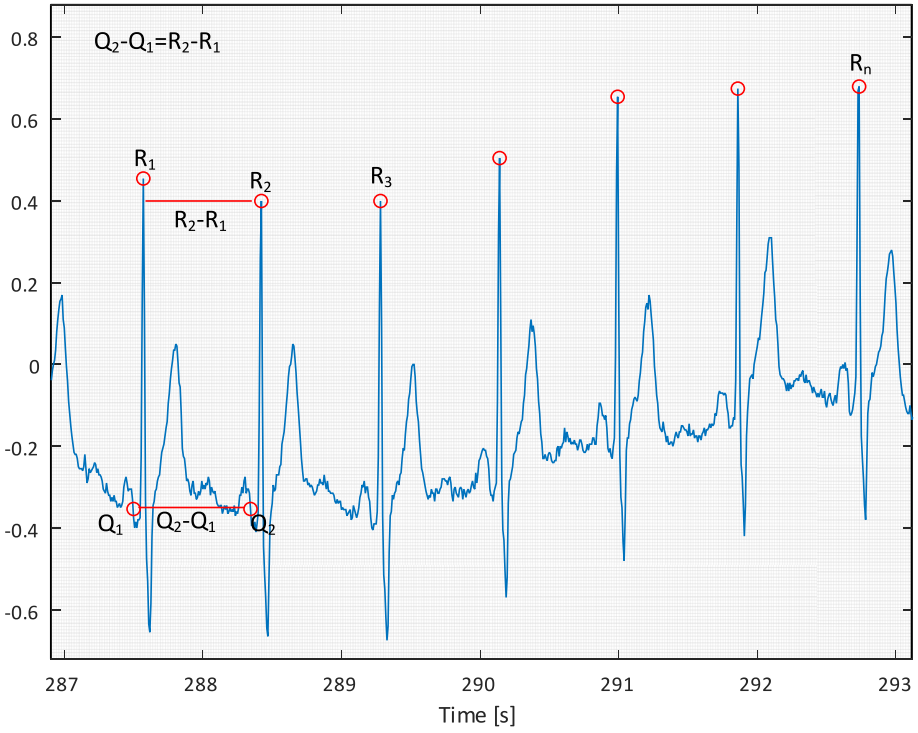

Figure 2. HRV extraction from RR intervals of an ECG signal.

2. Feature extraction: Where 10 different methods are used to extract 77 HRV features.
3. Data Analysis: Where three methods, including a recursive feature elimination method, are used to find the optimal number of features to predict a PAF.

2.1. Data description

This research uses the AFPDB database of PhysioNet [22], which contains 50 record sets called "n" obtained from normal subjects or people who have never experienced PAF and 50 record sets called "p" obtained from people who have experienced PAF. Each record contains approximately 30 min of continuous ECG signals without any PAF content. Record sets "p" are divided into two classes: records that precede the immediate appearance of PAF (close PAF) and records that do not have PAF 45 min after its termination or 45 min before its start (distant PAF).

Each record contains 2 leads of the ECG signal and the location in time of the onset of the QRS complex. In this paper, we find the HRV signal by determining the duration of each beat through the QRS complex. The start of the QRS complex to the next QRS complex is equivalent to the RR interval, as shown in Figure 2.

Record n27 was not taken into account in this paper because previous works claimed that it contains considerable noise and greatly affects the calculation of the HRV signal [19]. The remaining 99 record sets were used to predict PAF through its classification. To compare this work with previous works [4, 9, 19, 20, 23, 24, 25, 26] having as criteria: window length, the number of features, and validation. The classification is carried out in two different ways:

- **Group 1:** Record sets are divided into 2 classes. The first class contains normal subjects and distant PAF signals, and the second class contains close PAF signals.

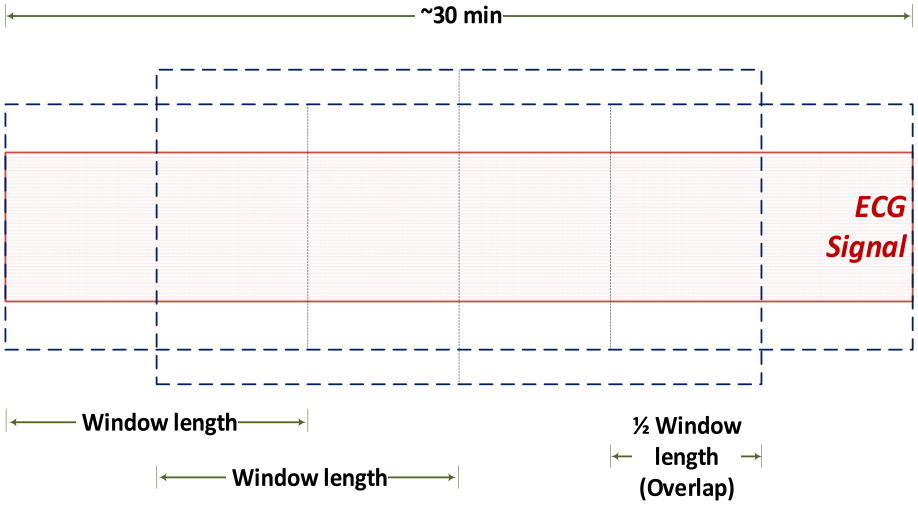

Figure 3. Windowing of ECG signal.

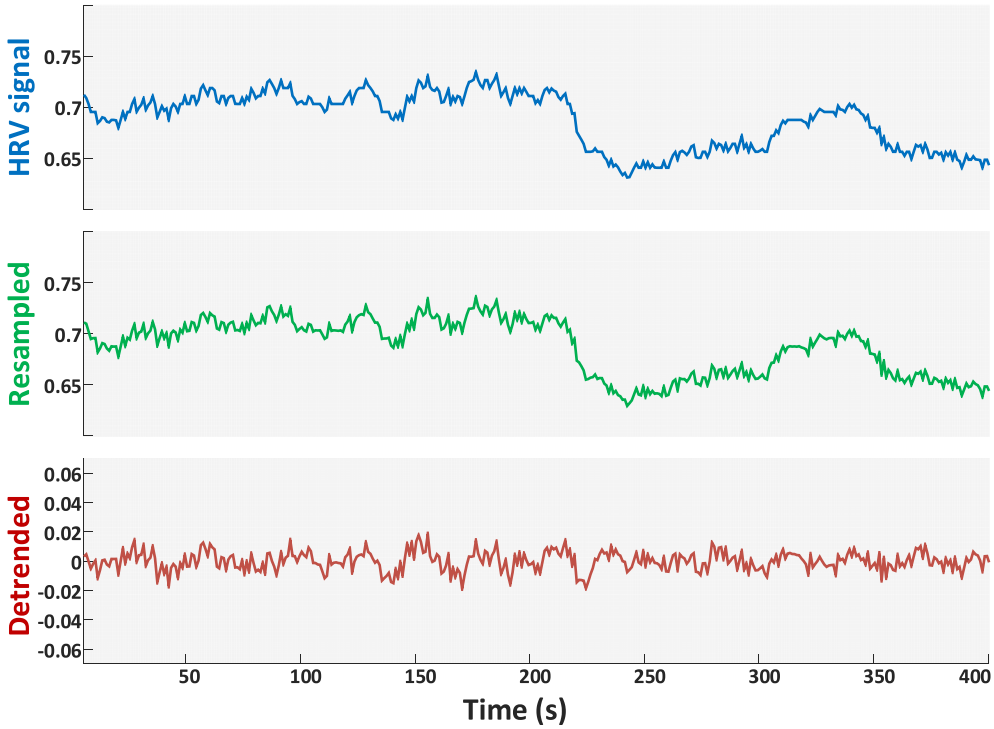

Figure 4. HRV signal preprocessing. Raw HRV signal (blue), resampled HRV signal (green), detrended HRV signal (red).

- **Group 2:** Record sets are divided into 2 classes. The first class contains distant PAF signals, and the second class contains close PAF signals.

2.2. Preprocessing

Previous work has used ECG signals of different durations for the prediction of PAF [4, 9, 19, 20, 23, 24, 25, 26]. In this paper, signals with durations of 30, 10, 5, 2, and 1 min were used, and their effectiveness was compared. These windows length were chosen to compare the results obtained by this study with previous work done by other authors.

To obtain these signals, a windowing process was performed by dividing the original ECG signal into overlapping segments by 50%, as shown in Figure 3.

Once this process had been carried out, the HRV signal was extracted from each record by measuring the time elapsed between two consecutive beats or two consecutive R peaks (RR interval), as shown in Figure 2 and Algorithm 1.

Algorithm 1. Windowing process.

```
ECG(t) // ECG signal
winSize // Window size in seconds
winOlap // Window overlap in seconds

gap ← winSize - winOlap

// Number of windows
nWin ← RoundDown(length(ECG)/gap)

for w from 0 to nWin-1
win[w] ← ECG(t >= w*gap and t < winSize +
              w*gap)
  Check Rpeaks location in win[w]

  for each peak_n in Rpeaks
       winHRV[w] ← peak_n - peak_n-1
```

The time domain, Poincaré plot, and Lomb-Scargle periodogram feature extraction methods can work with raw HRV signals; however, the other methods require uniform sampling. Due to the nature of obtaining the HRV signal, the time between samples is directly affected by the instantaneous heart rate of the ECG signal. To correct this, resampling of the HRV signal is performed at 7 Hz using the cubic spline method, which allows an ECG signal up to 210 bpm to be correctly represented [27].

Methods based on frequency domain analysis require that, in addition to uniform sampling, the HRV signal has no trend. To achieve this feature, the wavelet package decomposition method was used to eliminate frequencies lower than 0.04 Hz corresponding to the trend of the signal [28, 29], as illustrated in Figure 4.

2.3. HRV feature extraction

In this stage, a raw HRV signal was used to extract 8 features by time-domain analysis, 3 features by Poincaré plot, and 5 features by Lomb-Scargle periodogram [30]. Additionally, the resampled HRV signal was used to extract 2 features by the geometrical method, 1 feature by nonlinear methods, and 2 features by detrended fluctuation analysis. Finally, the resampled and detrended HRV signal was used to extract 45 features by bispectral analysis, 3 features by autoregressive modeling, 3 features by fast Fourier transform, and 5 features by wavelet packet transform. Table 1 summarizes these characteristics and describes the references used for their calculation.

2.3.1. Extracted features from raw HRV signal

The time-domain analysis allows us to statistically describe the HRV signal: AVNN is the mean value of the signal NN interval, SDNN is the standard deviation, SDSD is the standard deviation of the difference between consecutive HRV values, RMSSD is the root mean square of successive differences between consecutive HRV values; NN50 is the total number of consecutive HRV values whose difference is greater than 50 ms, NN20 is the total number of consecutive HRV values whose difference is greater than 20 ms, pNN50 is the percentage of the total consecutive HRV values whose difference is greater than 50 ms, and pNN20 is the percentage of total consecutive HRV values whose difference is greater than 20 ms. These features are calculated using Algorithm 2.

Table 1. Standard HRV features Time and frequency domains and different techniques used in the study.

Feature	References
Time Domain Analysis	
AVNN, SDNN, SDSD, RMSSD, NN50, NN20, pNN50, pNN20	[25] Boon *et al.* 2016
Poincaré Plot	
SD1, SD2, SDRate	[31] Yu *et al.* 2012
Lomb–Scargle Periodogram	
lsULF, lsVLF, lsLF, lsHF, lsLFHF	[32] Lomb. 1976
Geometrical Method	
rrTri, TINN	[1] García *et al.* 2017
Nonlinear Analysis	
SampEn	[9] Mohebbi *et al.* 2012
Detrended Fluctuation Analysis	
DFA1, DFA2	[25] Boon *et al.* 2016
Bispectral Analysis	
Mave, Pe, P1, P2	[10] Acharya *et al.* 2006
H1, H2, H3, H4	[9] Mohebbi *et al.* 2012
MaveROI	[31] Yu *et al.* 2012
MaveLL, MaveLH, MaveHH	[25] Boon *et al.* 2016
PaveROI	[31] Yu *et al.* 2012
PaveLL, PaveLH, PaveHH, P1ROI, P1LL, P1LH, P1HH, P2ROI, P2LL, P2LH, P2HH, H1ROI, H1LL, H1LH, H1HH, H2ROI, H2LL, H2HH, H3ROI, H3LL, H3HH, H4ROI, H4LL, H4HH	[25] Boon *et al.* 2016
Z1ROI, Z2ROI, Z1LL, Z2LL, Z1LH, Z2LH, Z1HH, Z2HH	[31] Yu *et al.* 2012
Autoregressive Modeling	
arLF, arHF, arLFHF	[10] Acharya *et al.* 2006
Fast Fourier Transform	
fftLF, fftHF, fftLFHF	[19] Narin *et al.* 2018
Wavelet Packet Transform	
waveLF, waveHF, waveLFHF, entLF, entHF	[19] Narin *et al.* 2018

Algorithm 2. Time-domain analysis.

```
winHRV(t) // Each of the windows of HRV signal
AVNN ← mean(winHRV(t))
SDNN ← std(winHRV(t))
SDSD ← std(winHRV(t)-winHRV(t-1))
RMSSD ← rms(winHRV(t)-winHRV(t-1))
NN50 ← sum((winHRV(t)-winHRV(t-1)) > 50/1000)
NN20 ← sum((winHRV(t)-winHRV(t-1)) > 20/1000)
pNN50 ← NN50/length(winHRV(t))
pNN20 ← NN20/length(winHRV(t))
```

The Poincaré plot method is a graph of each RR interval versus immediately following the RR interval. This graph provides detailed beat-to-beat information on heart behavior [7, 33] and is very useful as a predictor of heart disease and dysfunction [10]. The features extracted by this method are based on the instantaneous beat-to-beat interval variability (SD1), the continuous long-term RR interval variability (SD2), and the SD1/SD2 ratio (SDRate) [34], as shown in Algorithm 3.

Algorithm 3. Poincare plot.

```
SD1 ← sqrt(1/2 * SDSD^2)
SD2 ← sqrt(2*SDNN^2 - 1/2 * SDSD^2)
SDRate ← SD1/SD2
```

The Lomb-Scargle periodogram is a method used to calculate power spectral density (PSD) without the need for preprocessing and is much more accurate than FFT methods [35]. To perform feature extraction, this method is applied in the 4 main frequency bands in an HRV signal: the ultralow-frequency band (ULF) between 0 Hz and 3.3 mHz, the very-low-frequency band (VLF) between 3.3 mHz and 40 mHz, the low-frequency band (LF) between 40 mHz and 150 mHz and the high-frequency band (HF) between 150 mHz and 400 mHz. As an additional feature, the LF/HF ratio is also calculated [19, 36], as shown in Algorithm 4.

Algorithm 4. Obtaining PSD by Lomb-Scargle periodogram method.

```
plomb // Lomb-Scargle Periodogram method
PSD(f) ← plomb(winHRV)

lsULF ← sum(PSD(f>=0 and f<0.0033))
lsVLF ← sum(PSD(f>=0.0033 and f<0.004))
lsLF ← sum(PSD(f>=0.004 and f<0.015))
lsHF ← sum(PSD(f>=0.015 and f<0.4))
lsLFHF ← lsLF/lsHF
```

2.3.2. *Extracted features from resampled HRV signal*

According to the geometric method, the histogram of the HRV signal was obtained, and from it, the HRV index (rrTri) and the triangular interpolation of the RR intervals (TINN) were calculated using Algorithm 5 [1,37,38].

Algorithm 5. Geometrical method.

```
winRes // Each of the resampled windows of HRV signal
fsHRV // HRV signal sample frequency

bin ← 1/fsHRV
hist ← histogram(winRes)
rrTri ← length(winRes)/max(hist)
tinn ← 2*length(winRes)*bin/max(hist)
```

From the nonlinear analysis, the sample entropy feature (SampEn) was extracted since it overcomes the limitations of Kolmogorov-Sinai (KS) entropy when working with real data [39], as shown in Algorithm 6. Where the length of two simultaneous data points (m) to be compared and the distance between said data points (r) were fixed according to the study done in [40].

Algorithm 6. Sample entropy method.

```
sampen // Sample entropy method

m ← 2
r ← 0.2*std(winRes)
sampEn ← sampen(winRes,m,r)
```

Detrended fluctuation analysis (DFA) is used to extract nonlinear characteristics from the HRV signal. It is a measure that quantifies the fractal scale properties of short RR intervals [13]. It is calculated by

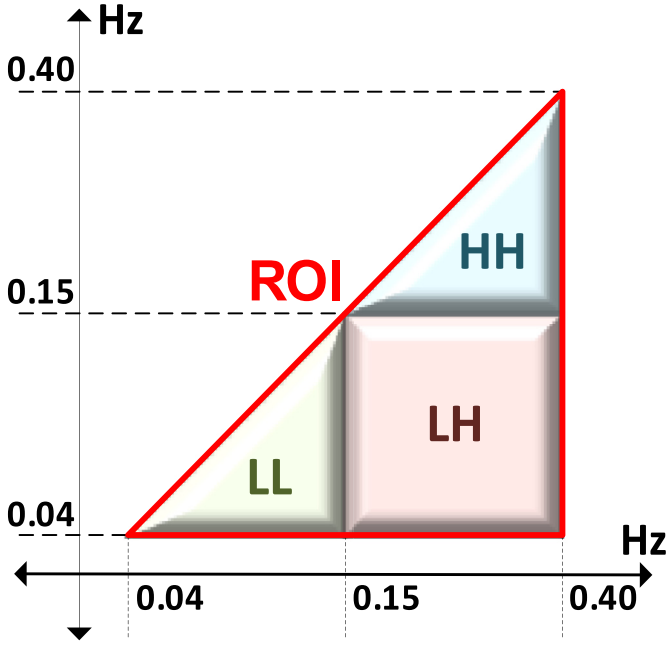

Figure 5. Region of interest (ROI) of the bispectrum and its 3 subdivisions.

Algorithm 7. In this study, HRV is divided in windows of 20-samples length, then, DFA applies a linear regression (Order 1) to find and eliminate the local trend.

Algorithm 7. Detrended fluctuation analysis method.

```
DFA_fun // Detrended fluctuation analysis method
samples ← 20
order ← 1

dfa1, dfa2 ← DFA_fun(winRes,samples,order);
```

2.3.3. Extracted features from resampled and detrended HRV signals

Higher-order spectral analysis (HOS) has been used to estimate the bispectrum in recent research based on HRV analysis [15]. A region of interest (ROI) between frequencies of 40 mHz and 400 mHz is identified in this bispectrum. This region is subdivided into 3 smaller regions: the low-low- frequency region (LL), the low-high-frequency region (LH), and the high-high-frequency region (HH). Figure 5 shows these regions.

The HRV characteristics in the frequency domain are based on the analysis of the PSD obtained from different algorithms, such as FFT and wavelet packet transform (WPT) [41]. Spectral analysis tends to relate variations in frequency bands with physiological modular effects. WPT analysis in HRV is used to separate the signal by amplitude and scaling to simultaneously analyze the time and frequency domains [28]. Based on [42, 43], this paper used a Daubechies (DB4) mother wavelet with a scale of 7 for this process. Algorithm 8 shows how to obtain the bispectral features.

Algorithm 8. Bispectral analysis.

```
winSta // Each of the resampled and detrended windows of HRV signal
bispectral // Bispectral method

biSpec(f1,f2) ← bispectral(winSta)

// Non-redundadnt region
OmegaMag ← abs(biSpec(f2<=abs(f1) and f2>=0))
OmegaPhs ← angle(biSpec(f2<=abs(f1) and f2>=0))

// Region of interest (ROI)
ROIMag ← abs(biSpec(f1>=0.04 and f1<=0.4 and f2>=0.04 and f2<=0.4 and f2<=f1))
LLMag ← abs(biSpec(f1>=0.04 and f1<=0.15 and f2>=0.04 and f2<=0.15 and f2<=f1))
LHMag ← abs(biSpec(f1>0.15 and f1<=0.4 and f2>=0.04 and f2<=0.15))
HHMag ← abs(biSpec(f1>0.15 and f1<=0.4 and f2>0.15 and f2<=0.4 and f2<=f1))

Mave ← mean(OmegaMag)

N ← length(biSpec)
for n from 0 to N-1
   p(n+1) ← mean(OmegaPhs >= -pi+2*pi*n/N and OmegaPhs < -pi+2*pi*(n+1)/N)

Pe ← -sum(p*log(p))

pn ← OmegaMag/sum(OmegaMag)
P1 ← -sum(pn*log(pn))

pn ← OmegaMag^2/sum(OmegaMag)^2
P2 ← -sum(pn*log(pn))

H1 ← sum(log(OmegaMag))

logDiag ← log(abs(biSpec(f2==f1 and f2>=0)))
H2 ← sum(logDiag);

k ← (1 to length(logDiag))
H3 ← sum(k*(logDiag))

H4 ← sum((k-H3)^2*logDiag)

MaveROI ← mean(ROIMag)
MaveLL ← mean(LLMag)
MaveLH ← mean(LHMag)
MaveHH ← mean(HHMag)
PaveROI ← mean(ROIMag^2)
PaveLL ← mean(LLMag^2)
PaveLH ← mean(LHMag^2)
PaveHH ← mean(HHMag^2)

pn ← ROIMag/sum(ROIMag)
P1ROI ← -sum(pn*log(pn))

pn ← LLMag/sum(LLMag)
P1LL ← -sum(pn*log(pn))

pn ← LHMag/sum(LHMag)
P1LH ← -sum(pn.*log(pn))

pn ← HHMag/sum(HHMag)
P1HH ← -sum(pn*log(pn))

pn ← ROIMag^2/sum(ROIMag)^2
P2ROI ← -sum(pn*log(pn))

pn ← LLMag^2/sum(LLMag)^2
P2LL ← -sum(pn*log(pn))
```

```
pn ← LHMag^2/sum(LHMag)^2
P2LH ← -sum(pn*log(pn))

pn ← HHMag^2/sum(HHMag)^2
P2HH ← -sum(pn*log(pn))

H1ROI ← sum(log(ROIMag))
H1LL ← sum(log(LLMag))
H1LH ← sum(log(LHMag))
H1HH ← sum(log(HHMag))

logROIDiag ← log(abs(biSpec(f1>=0.04 and f1<=0.4 and f2>=0.04 and f2<=0.4 and f2==f1)))
H2ROI ← sum(logROIDiag)

logLLDiag ← log(abs(biSpec(f1>=0.04 and f1<=0.15 and f2>=0.04 and f2<=0.15 and f2==f1)))
H2LL ← sum(logLLDiag)

logHHDiag ← log(abs(biSpec(f1>0.15 and f1<=0.4 and f2>0.15 and f2<=0.4 and f2==f1)))
H2HH ← sum(logHHDiag)

kROI ← (1 to length(logROIDiag))
H3ROI ← sum(kROI*(logROIDiag))

kLL ← (1 to length(logLLDiag))
H3LL ← sum(kLL*(logLLDiag))

kHH ← (1 to length(logHHDiag))
H3HH ← sum(kHH*(logHHDiag))

H4ROI ← sum((kROI-H3ROI)^2*logROIDiag)
H4LL ← sum((kLL-H3LL)^2*logLLDiag)
H4HH ← sum((kHH-H3HH)^2*logHHDiag)

rows // Number of rows
columns // Number of columns
iROI ← 1 to rows(ROIMag)
jROI ← 1 to columns(ROIMag)
Z1ROI ← sum(iROI*ROIMag)/sum(ROIMag)
Z2ROI ← sum(jROI*ROIMag)/sum(ROIMag)

iLL ← 1 to rows(LLMag)
jLL ← 1 to columns(LLMag)
Z1LL ← sum(iLL*LLMag)/sum(LLMag)
Z2LL ← sum(jLL*LLMag)/sum(LLMag)

iLH ← 1 to rows(LHMag)
jLH ← 1 to columns(LHMag)
Z1LH ← sum(iLH*LHMag)/sum(LHMag)
Z2LH ← sum(jLH*LHMag)/sum(LHMag)

iHH ← 1 to rows(HHMag)
jHH ← 1 to columns(HHMag)
Z1HH ← sum(iHH*HHMag)/sum(HHMag)
Z2HH ← sum(jHH*HHMag)/sum(HHMag)
```

The autoregressive modeling, FFT, and WPT methods were applied to the LF and HF frequency bands used in the Lomb-Scargle periodogram. ULF and VLF were not used in these methods since detrending eliminates the information of these frequency bands. Each of these methods is calculated using Algorithm 9, Algorithm 10, and Algorithm 11. The autoregressive model uses 16 coefficients (order 16) to calculate the power spectral density of the HRV signal [44]. On the other hand, WPT method uses 10 decomposition levels (nPack = 10) and a mother wavelet daubechies 6 [17].

Algorithm 9. Autoregressive modeling.

```
pburg // Autoregressive PSD estimate using Burg's method
order ← 16

arPSD(f) ← pburg(winSta,order,fsHRV)
arLF ← sum(arPSD(f>0.04 and f<=0.15))
arHF ← sum(arPSD(f>0.15 and f<=0.40))
arLFHF ← arLF/arHF;
```

Algorithm 10. Fourier transform.

```
fft // Fast Fourier transform method

fftWin(f) ← fft(winSta)
fftPSD(f) ← abs(fftWin)^2
fftLF ← sum(fftPSD(f > 0.04 and f <= 0.15))
fftHF ← sum(fftPSD(f > 0.15 and f <= 0.40))
fftLFHF ← fftLF/fftHF
```

Algorithm 11. Wavelet transform.

```
wpt // Wavelet package transform method
nPack ← 10 // Number of packages

waveWin(f,t) ← wpt(winSta, "daubechies6")
wavePSD(f,t) ← abs(waveWin)^2
waveLF ← sum(wavePSD(f>0.04 and f<=0.15,t))
waveHF ← sum(wavePSD(f>0.15 and f<=0.40,t))
waveLFHF ← waveLF/waveHF

entLF ← -sum(wavePSD(f>0.04 and f<=0.15,t) *log(wavePSD(f>0.04 and f<=0.15,t)))
entHF ← -sum(wavePSD(f>0.15 and f<=0.40,t) *log(wavePSD(f>0.15 and f<=0.40,t)))
```

2.4. Data analysis

In this stage, some of the extracted features are removed and prepared to be delivered to a classifier.

Reducing computational cost and reducing classifier dimensionality are two of the benefits of eliminating features. Furthermore, the elimination process seeks to obtain features that contain the most relevant information to classify the data. In this paper, three methods were used to reduce the characteristics in the following order: near-zero value, correlation, and recursive feature elimination.

First, the values are standardized to facilitate the learning process and normalize the scale in all dimensions.

Subsequently, data is rounded to two digits to facilitate obtaining unique values, which are the same data but with no repeating values. Rounding allows small differences between data to be eliminated, thus, reduces the number of unique values and increases the effectiveness of the near-zero variance elimination.

2.4.1. Eliminating features with near-zero variance

A feature whose variance is zero or contains highly repeating values has no contribution to the classification process. It is possible to find these features by calculating the number of unique values and comparing how many times these values are repeated. Using the method proposed in [45], a feature is eliminated if it meets both of the following conditions:

- The percentage of unique values is less than 10%.
- The rate between the value that is repeated the most and the second value that is repeated the most is greater than 19.

2.4.2. Eliminating features with high correlation

The fact that two or more features are correlated implies that they contain redundant information. To avoid this situation, the correlation matrix between all features was calculated. Each pair of features that had a very strong correlation, that is a value greater than 0.9 or less than -0.9 [46] was analyzed, and the feature that had a higher index calculated according to equation (1) was eliminated.

$$index_x = \sum_{y \neq x} (\rho_{x,y})^2 \quad (1)$$

where x is each of the features in the correlated pair, y is each of the features of the database, and $\rho_{x,y}$ is the Pearson correlation coefficient between x and y.

2.4.3. Recursive feature elimination

Recursive feature elimination (RFE) recursively evaluates subsets of features and finds the importance of each feature individually. This allows us to retain independent features and remove features that have a low impact on improving accuracy [47].

RFE has 2 main stages: feature subset selection and classification using this subset. There are different combinations of methods applicable to these stages. In this paper, backward selection, genetic algorithm, analysis of variance (ANOVA), and non-dominated sorting genetic algorithm (NSGA-III) were used for the selection of features, and random forest, conditional random forest, k-nearest neighbor (KNN), and support vector machine (SVM) were used for classification. To ensure the independence of the partitioning of data, 10-fold cross-validation was used to evaluate the results. The partitioning was the same for all methods. Table 2 summarizes the methods used.

The feature selection aims to find the most relevant features of a problem. It improves computational speed and prediction accuracy [48]. In this study, the feature selection algorithms of 'caret' package version 6.0–88 in R software is used: Backwards selection, genetic algorithm and anova, and the random forest classifier, which works well with high-dimension problems and identifies strong predictors of a specific result without making assumptions about a underlying model [48].

Furthermore, we extend the study using NSGA-III from the package 'mlr' version 2.19.0 and as classifiers: Contidional Random Forest, KNN and SVM. These machines build a classification model based on previous features and have been successfully applied in previous clinical studies [49].

Each of the aforementioned methods was evaluated using 3 performance metrics: sensitivity (SN), specificity (SP), and accuracy (ACC), as shown in equations (2), (3), and (4), respectively. These metrics are widely used and allow comparing the probability of success of the proposed method with previously published works.

$$SN = \frac{TP}{TP + FN} \quad (2)$$

Table 2. Different methods and parameters for optimal feature selection and classification.

Feature selection	Classification	Evaluation
Backwards Selection	Random Forest Number of trees = 500	Cross Validation 10-Fold (CV-10)
Genetic Algorithm Population size = 100 Max generations = 50 Crossover probability = 0.8 Mutation probability = 0.1		
ANOVA		
NSGA-III Population size = 100 Max generations = 50 Crossover probability = 0.8 Mutation probability = 0.1	Conditional Random Forest Number of trees = 500	
	KNN	
	SVM k = 1 Kernel = Radial σ^2 = Number of features	

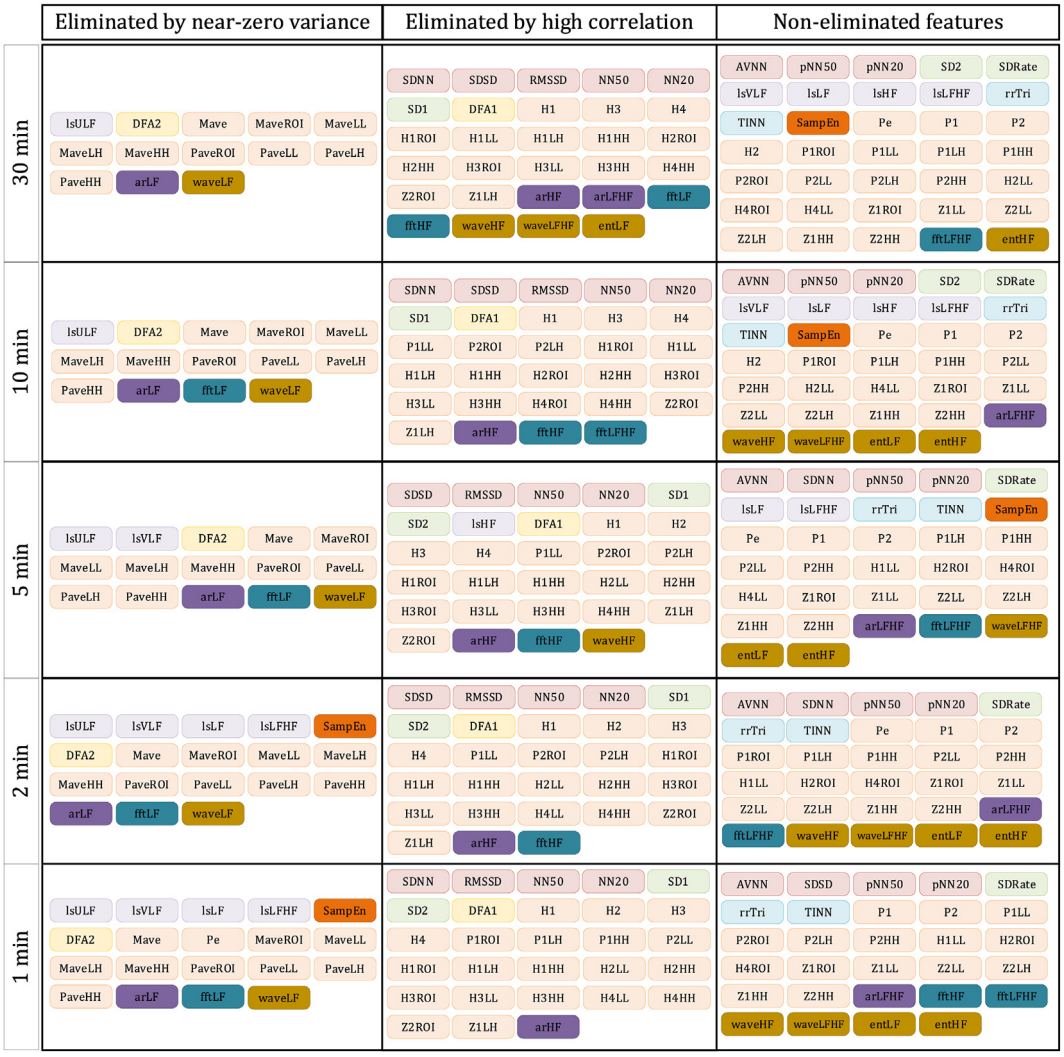

Figure 6. Elimination of features in group 1 using near-zero variance and high correlation methods.

$$SP = \frac{TN}{TN + FP} \tag{3}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

where *TP* is the true positive value, *TF* is the true negative value, *FP* is the false positive value, and *FN* is the false negative value.

2.5. Ethical statement

The work covered in this paper has been carried out using the AFPDB database from PhysioNet [22]. All medical data included in this database is publicly available and approved to be freely used and shared.

We confirm that this paper complies with any ethical conditions established by Physionet and the database authors.

3. Results and analysis

Once the 77 features were obtained, the near-zero variance and high correlation method were applied. The results obtained show that depending on the length of the window, some features become more or less relevant. Figure 6 and Figure 7 graphically show the results obtained for groups 1 and 2, respectively.

In group 1, it can be seen that regardless of the window length, 15 features contain a large amount of information; therefore, they were not eliminated due to low variance or high correlation.

In contrast, 13 features (lsULF, dfa2, arLF, waveLF, Mave, MaveROI, MaveLL, MaveLH, MaveHH, PaveROI, PaveLL, PaveLH, and PaveHH) were eliminated due to low variance regardless of the window length. Likewise, 20 features (RMSSD, NN50, NN20, SD1, DFA1, arHF, H1, H3, H4, H1ROI, H1LL, H1LH, H1HH, H2HH, H3ROI, H3LL, H3HH, H4HH, Z2ROI, and Z1LH) were highly correlated with other features and were thus eliminated regardless of the window length.

In group 2, 17 features were not eliminated by near-zero variance or by high correlation regardless of the window length. In contrast, the 6 features dfa2, lsULF, PaveROI, PaveLL, PaveLH, and PaveHH were eliminated by near-zero variance in all window lengths. In the same way, the 20 features SDSD, RMSSD, NN50, pNN20, SD1, SD2, DFA1, H1, H3, H4, P1LH, H1ROI, H1LL, H1LH, H1HH, H2LL, H2HH, H3HH, H4HH, and Z1LH were eliminated by correlation across all window lengths.

Based on these results, 12 features (AVNN, pnn50, rrTri, TINN, SDRate, entHF, P1, P2, Z1LL, Z2LL, Z2LH, and Z2HH) contain the most information and can help to correctly predict AFP. In contrast, 21 features (RMSSD, NN50, SD1, DFA1, DFA2, lsULF, PaveROI, PaveLL, PaveLH, PaveHH, H1, H3, H4, H1ROI, H1LL, H1LH, H1HH, H2HH, H3HH, H4HH, and Z1LH) contain reduced or redundant information and should not be used for the prediction of AFP.

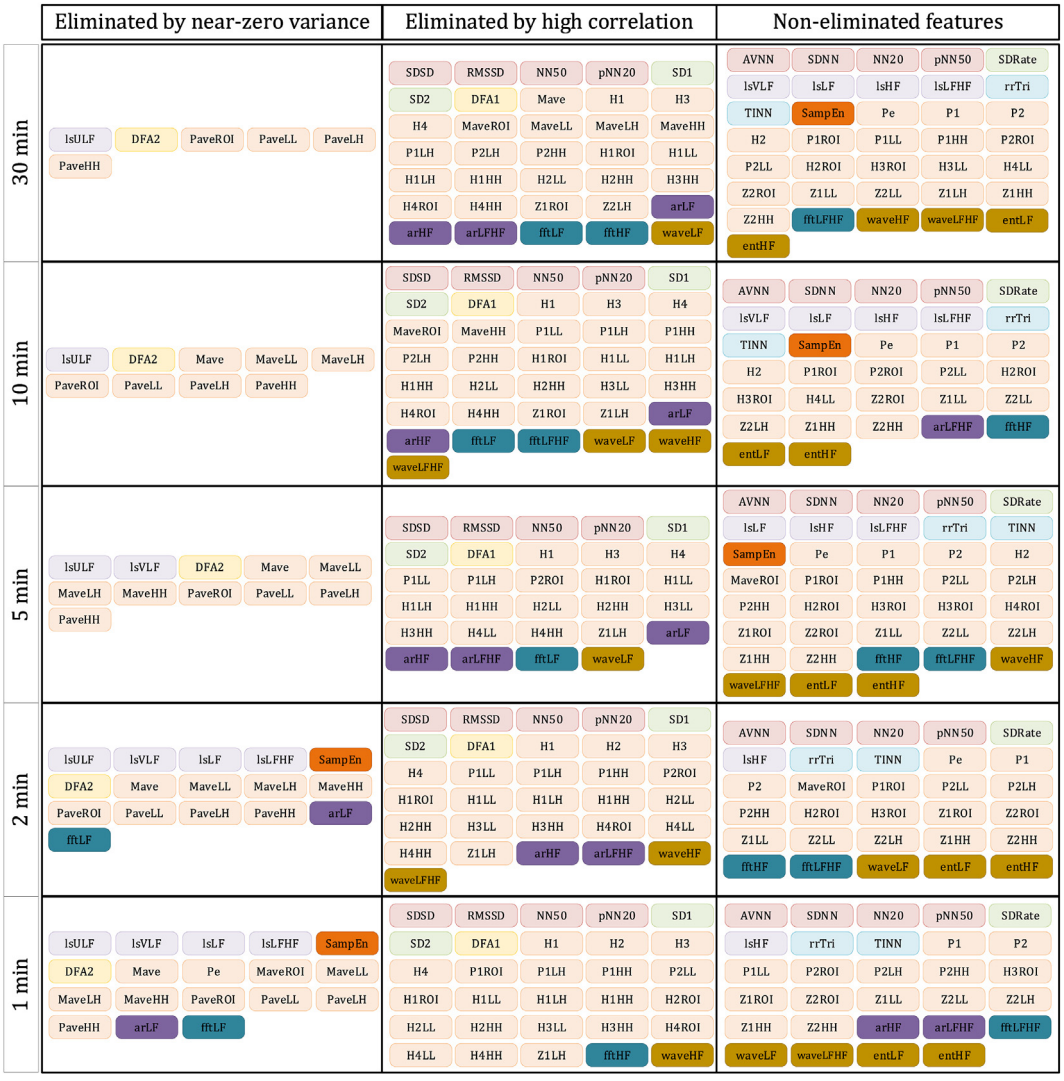

Figure 7. Elimination of features in group 2 using near-zero variance and high correlation methods.

After an exhaustive elimination of features, between 29 and 34 of them remain depending on the window length. Some of these features were further eliminated using the recursive feature elimination method. Table 3 and Table 4 show the obtained optimal set of features and the classification accuracy.

The highest accuracy was 89.01% for a window length of 2 min and 93.24% for a window length of 5 min.

According to these results, the best combination of algorithms for the selection of the optimal set of features was NSGA-III + KNN for group 1 and a 5-minute window length and backwards selection + random forest for group 2 and a 2-minute window length.

In group 1, the highest precision was obtained using the 6 features AVNN, SDNN, pNN50, pNN20, lsLF, and TINN. This result shows that time-domain analysis has a great impact on predicting a PAF event.

In group 2, the highest precision was obtained using the 9 features AVNN, NN20, pNN50, TINN, rrTri, SDNN, fftHF, MaveROI, and SDRate. As with group 1, time-domain analysis is very important in predicting PAF events, but in this case, the geometrical method also has a great impact.

According to these results, fftHF, MaveROI, and SDRate features are relevant to discriminate between distant PAF and close PAF. However, when including normal subjects, this feature's importance is lost. Therefore, the information contained in the high frequencies of the ECG signal and the rate of occurrence of ectopic beats vary considerably in normal subjects, but in people with fibrillation, it helps to predict when a PAF may occur.

In Table 5, this paper is compared to previous works in predicting a PAF event on the AFPDB. Five separate works used a 5-minute window length [30]. obtained a classification performance of 78.4% by using the P wave power spectral density [31]. achieved a classification performance of 72% using HRV power spectral density and premature atrial contractions (PACs). A very recent study in [13] reconsidered the problem using 5-minute HRV segments and obtained a classification performance with an accuracy of 87.7%.

The proposed methodology exceeds the results obtained by all of the methods mentioned before. The highest sensitivity and specificity for group 1 were 88.53% and 95.03%, respectively, using a 5-minute window. These results outperform previous studies with the same window length. On the other hand, for group 2 using a 2-minute window, a sensitivity of 88.00% and a specificity of 90.43% were obtained. Despite using a shorter window length, the results of the group are higher than in previous works except for [20], where they used a 10-minute window, and [9], where they used a 30-minute window.

Using a smaller window length reduces the amount of data that needs to be processed to obtain a classification of the signal and allows a PAF to be predicted more quickly than with a longer window length. In a real implementation, these advantages mean fewer data to store and process and timely medical decision making. On the other hand, reducing the

Table 3. Optimal set of HRV features for group 1.

Window length (minutes)	Features	SN (%)	SP (%)	ACC (%)
Backwards Selection + Random Forest + CV-10				
30	SDRate, P2LH	32.00	87.83	75.77
10	P2, SDRate, P1, pNN20, rrTri, TINN, Z1ROI, waveHF, sampEn, SD2	54.00	98.32	86.11
5	AVNN, P2, pNN20, SDNN, SDRate	62.80	95.95	87.59
2	AVNN, pNN20, SDNN, pNN50, rrTri, TINN, SDRate	62.71	95.53	88.00
1	AVNN, SDSD, pNN20, TINN, SDRate, rrTri, fftHF, pNN50, waveHF, arLFHF, H2ROI	62.00	96.88	88.31
Genetic Algorithm + Random Forest + CV-10				
30	pNN50, lsLF, rrTri, P2LH, P2HH, H4ROI, Z2LL, Z1HH, Z2HH	28.00	91.89	75.76
10	AVNN, pNN20, SD2, SDRate, rrTri, TINN, sampEn, arLFHF, P2, P2HH, Z2LH	55.00	97.98	87.15
5	AVNN, SDNN, pNN50, pNN20, SDRate, TINN, sampEn, fftLFHF, Z2HH	63.60	96.49	88.19
2	AVNN, SDNN, pNN50, pNN20, SDRate, TINN, fftLFHF, H2ROI, Z2HH	63.71	96.62	88.32
1	AVNN, SDSD, pNN50, pNN20, SDRate, lsHF, TINN, arLFHF, fftHF	68.62	95.90	89.01
Anova + Random Forest + CV-10				
30	SDRate, sampEn, Z2LH	16.00	91.89	72.73
10	SDRate, arLFHF, entLF, Pe, P1, P2, P1ROI, P1LH, P1HH, P2LL, P2HH, Z1ROI, Z1LL, Z2LL, Z2LH, Z1HH	23.00	96.30	77.83
5	SDRate, lsLF, rrTri, tinn, arLFHF, fftLFHF, entLF, P1, P2, P1ROI, P1LH, P1HH, H2ROI, H4ROI, Z1ROI, Z1LL	46.40	95.95	83.45
2	AVNN, pNN50, SDRate, arLFHF, fftLFHF, waveLFHF, entLF, entHF, P1, P2, P1ROI, P1LH, H2ROI, H4ROI	43.14	97.97	84.13
1	AVNN, PNN50, pNN20, SDRate, arLFHF, fftHF, fftLFHF, waveLFHF, entLF, entHF, P1, P2, P1LL, P2ROI, P2LH	40.21	97.74	83.21
NSGA-III + Conditional Random Forest + CV-10				
30	tinn, P2, P2LH, P2HH, H2LL, H4ROI	16.00	91.89	72.73
10	lsLFHF, rrTri, sampEn, P2, P1HH	38.17	97.93	82.14
5	AVNN, SDRate, sampEn, fftLFHF, H2ROI	42.44	96.70	82.84
2	AVNN, SDNN, pNN20, SDRate	50.61	96.16	84.63
1	AVNN, SDSD, TINN	54.87	96.38	85.90
NSGA-III + KNN + CV-10				
30	SDRate, sampEn, entHF, P1, P2, H2, P2LL, P2LH, H4ROI	55.00	89.11	80.88
10	AVNN, pNN50, pNN20, SD2, P2	80.68	93.11	89.66
5	**AVNN, SDNN, PNN50, pNN20, lsLF, TINN**	**88.53**	**95.03**	**93.24**
2	AVNN, SDNN, pNN20, SDRate, lsHF, P1LH	64.46	89.15	82.90
1	AVNN, SDSD, TINN	66.10	88.77	83.06
NSGA-III + SVM + CV-10				
30	pNN20, SDRate, P1, Z1HH	11.67	100	77.94
10	pNN50, pNN20, SDRate, entLF, P1, H2LL	19.95	99.33	79.11
5	AVNN, pNN50, pNN20, SDRate, entHF, P1, P1HH, H2ROI, H4ROI, Z2LH, Z1HH	22.69	97.87	78.81
2	AVNN, pNN50, pNN20, P1LH, H2ROI, H4ROI, Z1ROI, Z1LL, Z2LH, Z1HH	19.54	97.91	78.27
1	AVNN, TINN	8.49	98.71	75.94

Row in bold shows the solution with the highest accuracy for each group.

Table 4. Optimal set of HRV features for group 2.

Window length (minutes)	Features	SN (%)	SP (%)	ACC (%)
Backwards Selection + Random Forest + CV-10				
30	Pe, SDNN	56.00	72.00	64.00
10	Z2ROI, SDNN, pNN50, rrTri, P2, NN20, AVNN, TINN	80	83	81.50
5	AVNN, NN20, pNN50, SDNN, Z2ROI, rrTri, TINN, fftHF	85.6	86.4	86
2	**AVNN, NN20, pNN50, TINN, rrTri, SDNN, fftHF, MaveROI, SDRate**	**88.00**	**90.43**	**89.21**
1	AVNN, NN20, pNN50, SDNN, rrTri, TINN, arHF, SDRate, arLFHF	87.03	88.07	87.55
Genetic Algorithm + Random Forest + CV-10				
30	SDNN, lsLFHF, rrTri, TINN, waveHF, P1ROI, H4LL	48.00	56.00	52.00
10	AVNN, SDNN, NN20, pNN50, SDRate, rrTri, TINN, sampEn, P1ROI, H2ROI, H4LL, Z1LL, Z1HH, Z2HH	72.00	72.00	72.00
5	AVNN, SDNN, NN20, pNN50, SDRate, sampEn, P2, H2ROI, Z2ROI, Z1LL	83.20	84.80	84.00
2	AVNN, SDNN, NN20, pNN50, SDRate, TINN, fftHF, P2, H2ROI, Z2LH	84.86	86.57	85.71
1	AVNN, SDNN, NN20, pNN50, SDRate, rrTri, arHF, arLFHF, Z2ROI	85.59	85.86	85.72
Anova + Random Forest + CV-10				
30	pe	40.00	40.00	40.00
10	rrTri, TINN, entHF, P2ROI, H4LL, Z2HH	64.00	67.00	65.50
5	pNN50, SDRate, rrTri, TINN, sampEn, fftHF, waveHF, entHF, H2, P1HH, P2LH, H2ROI, H4ROI, Z2ROI	72.8	74.4	73.6
2	SDNN, NN20, pNN50, SDRate, rrTri, TINN, fftHF, MaveROI, H2ROI, Z1LL, Z2LL, Z2LH, Z2HH	78.29	82.14	80.21
1	SDNN, NN20, pNN50, SDRate, rrTri, TINN, arHF, waveLF, P1, P2, P2LH, H3ROI, Z1LL, Z2LL, Z2LH, Z1HH	74.62	76.41	75.52
NSGA-III + Conditional Random Forest + CV-10				
30	AVNN, pNN20, SDRate, fftLFHF	53.52	95.50	84.92
10	SDNN, rrTri, TINN, sampEn	63.60	76.27	70.5
5	AVNN, SDNN, NN20	76.43	82.48	79.00
2	AVNN, SDNN, NN20, pNN50, SDRate, rrTri, P2LH, Z2ROI	80.49	84.26	82.36
1	AVNN, SDRate, TINN	78.12	79.04	78.60
NSGA-III + KNN + CV-10				
30	AVNN, SDSD, pNN20, TINN, fftHF, P2ROI	71.68	90.35	85.60
10	AVNN, pNN50, lsHF, fftHF, P2, Z2ROI, Z1HH, Z2HH	69.51	82.03	76.00
5	rrTri, sampEn, Z2HH	59.00	63.42	60.8
2	AVNN, pNN50, TINN	83.11	84.10	83.64
1	AVNN, SDNN, pNN50, TINN	82.40	83.14	82.76
NSGA-III + SVM + CV-10				
30	AVNN, TINN, H4ROI, Z1ROI, Z2LH	10.76	99.14	76.82
10	AVNN, NN20, TINN, H2ROI, H4LL	66.72	70.88	68.00
5	AVNN, NN20, SDRate, TINN, P1ROI, P2LL	62.98	73.63	68.20
2	AVNN, NN20, pNN50 TINN, H2ROI	69.30	73.40	71.35
1	pNN50, entHF	41.37	73.59	57.41

Row in bold shows the solution with the highest accuracy for each group.

length of the window excessively affects the precision of the classification. In [23], they used a 1-minute window, obtaining a low precision of 68%; in the same way, in our work, the results obtained by 1-minute windows were lower in both group 1 and group 2.

Table 5. Proposed method compared to previous works.

Reference	Methods	Group	Window length (minutes)	Cross Validation	SN (%)	SP (%)	ACC (%)
[20] Chazal et al. 2001	Time Domain Analysis	Group 2	10	5-Fold	85	97	90.4
	Fast Fourier Transform	Group 2	5	5-Fold	81	75	78.4
[23] Hickey et al. 2002	Time Domain Analysis	Group 1	30	5-Fold	61	75	70
	Fast Fourier Transform	Group 1	10	5-Fold	65	75	72
		Group 1	5	5-Fold	62	77	72
		Group 1	1	5-Fold	60	72	68
[4] Zong et al. 2001	Time Domain Analysis	Group 1	30	Single-fold	-	-	80
[24] Thong et al. 2004	Time Domain Analysis	Group 1	30	Single-fold	84	88	86
[9] Mohebbi et al. 2012	Time Domain Analysis Poincaré Plot Nonlinear Analysis Autoregressive Modeling Fast Fourier Transform	Group 2	30	Single-fold	96.30	93.10	92.86
[25] Boon et al. 2016	Time Domain Analysis	Group 2	30	Single-fold	96.4	71.4	83.9
	Poincaré Plot	Group 2	30	10-fold	81.1	79.3	80.2
	Nonlinear Analysis	Group 2	10	Single-fold	75.1	54.3	69.6
	Autoregressive Modeling	Group 2	10	10-fold	58.5	81.1	68.9
	Fast Fourier Transform	Group 2	15	Single-fold	85.1	82.1	83.9
		Group 2	15	10-fold	77.4	81.1	79.3
[26] Boon et al. 2018	Time Domain Analysis Poincaré Plot Nonlinear Analysis Autoregressive Modeling Fast Fourier Transform	Group 2	5	10-fold	86.8	88.7	87.7
[19] Narin et al.2018	Time Domain Analysis	Group 1	5	10-fold	64	90.5	83.8
	Lomb–Scargle Periodogram	Group 2	5	10-fold	92	88	90
	Fast Fourier Transform Wavelet Packet Transform						
Proposed Method	Time Domain Analysis	Group 1	5	10-fold	88.53	95.03	93.24
	Poincaré Plot	Group 2	2	10-fold	88.00	90.43	89.21
	Lomb–Scargle Periodogram Geometrical Method Nonlinear Analysis Detrended Fluctuation Analysis Bispectral Analysis Autoregressive Modeling Fast Fourier Transform Wavelet Packet Transform						

4. Conclusion

HRV has proven to be an essential tool to predict PAF events and thereby to study the behavior of the sympathetic and parasympathetic function of sympathetic nerve activity.

In this study, a methodology was presented using the HRV signal, from which 77 features were selected based on a literature review of the majority of studies carried out in PAF event prediction. Features containing near-zero variance and high correlation were eliminated. In addition, 6 different techniques were used for recursive feature elimination, and the performance of the classifier was evaluated using 10-fold cross-validation.

Our method can predict a PAF event with 93.24% accuracy using a 5-minute window of an ECG signal or 89.21% accuracy using a 2-minute window of an ECG signal. These results were obtained for groups 1 and 2 using the AFPDB database from PhysioNet.

The proposed methodology exceeds the accuracy obtained by all of the methods consulted. The sensitivity obtained for group 1 was 88.53%, and the sensitivity of 95.03% was the highest.

The accuracy obtained for group 2 was 1% below the top 2 other methods. However, since this study uses a smaller window length, it has greater advantages than the methods consulted.

Another highlight of this work is the ability to reduce high-dimensional data from 77 to just 6 to 9 features. For group 1, the most important features were AVNN, SDNN, pNN50, pNN20, lsLF, and TINN. For group 2, the highest precision was obtained using AVNN, NN20, pNN50, TINN, rrTri, SDNN, fftHF, MaveROI, and SDRate. This result shows that time-domain analysis and geometrical methods have a great impact on predicting a PAF event.

This study uses features based only on the HRV signal. In future work, features based on the morphology of the ECG signal could be added, such as P-Wave and QR alternance analysis, the methodology could be extended to other cardiac pathologies, the hardware implementation of the propose methodology to create a real-time PAF detection and prediction device and expand the methodology including other ECG leads or with multiple leads at the same time.

Declarations

Author contribution statement

Henry Castro, Juan D Garcia-Racines & Alvaro Bernal-Norena: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research was supported by Dirección General de Investigaciones of Universidad Santiago de Cali under call No. 01-2021.

Data availability statement

No data was used for the research described in the article

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] C.A. García Martínez, A. Otero Quintana, X.A. Vila, M.J. Lado Touriño, L. Rodríguez-Liñares, J.M. Rodríguez Presedo, A.J. Méndez Penín, Heart Rate Variability Analysis with the R Package RHRV, Springer International Publishing, Cham, 2017.

[2] D. Lakkireddy, J. Pillarisetti, A. Patel, K. Boc, S. Bommana, Y. Sawers, S. Vanga, H. Sayana, W. Chen, J. Nath, J. Vacek, D. Lakkireddy, Evolution of paroxysmal atrial fibrillation to persistent or permanent atrial fibrillation: predictors of progression, J. Atr. Fibrillation 1 (2009) 388–394.

[3] S. Agewall, J. Camm, G. Barón Esquivias, W. Budts, S. Carerj, F. Casselman, A. Coca, R. De Caterina, S. Deftereos, D. Dobrev, J.M. Ferro, G. Filippatos, D. Fitzsimons, B. Gorenek, M. Guenoun, S.H. Hohnloser, P. Kolh, G.Y.H. Lip, A. Manolis, J. McMurray, P. Ponikowski, R. Rosenhek, F. Ruschitzka, I. Savelieva, S. Sharma, P. Suwalski, J. Luis Tamargo, C.J. Taylor, I.C. Van Gelder, A.A. Voors, S. Windecker, J. Luis Zamorano, K. Zeppenfeld, P. Kirchhof, S. Benussi, D. Kotecha, A. Ahlsson, D. Atar, B. Casadei, M. Castellá, H.-C. Diener, H. Heidbuchel, J. Hendriks, G. Hindricks, A.S. Manolis, J. Oldgren, B. Alexandru Popescu, U. Schotten, B. Van Putte, P. Vardas, Guía ESC 2016 sobre el diagnóstico y tratamiento de la fibrilación auricular, desarrollada en colaboración con la EACTS, Rev. Española Cardiol. 70 (2017), 50.e1-50.e84.

[4] W. Zong, R. Mukkamala, R.G. Mark, A methodology for predicting paroxysmal atrial fibrillation based on ECG arrhythmia feature analysis, Comput. Cardiol. (2001) 125–128.

[5] P. Langley, D. Di Bernardo, J. Allen, E. Bowers, F.E. Smith, S. Vecchietti, A. Murray, Can paroxysmal atrial fibrillation be predicted? Comput. Cardiol. (2001) 121–124.

[6] Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task force of the European society of cardiology and the North American society of pacing and electrophysiology, Eur. Heart J. 17 (1996) 354–381. http://www.ncbi.nlm.nih.gov/pubmed/8737210.

[7] P.W. Kamen, H. Krum, A.M. Tonkin, Poincare plot of heart rate variability allows quantitative display of parasympathetic nervous activity in humans, Clin. Sci. 91 (1996) 201–208.

[8] P.W. Kamen, A.M. Tonkin, Application of the Poincaré plot to heart rate variability: a new measure of functional status in heart failure, Aust. N. Z. J. Med. 25 (1995) 18–26.

[9] 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 Progr. Biomed. 105 (2012) 40–49.

[10] U.R. Acharya, K.P. Joseph, N. Kannathal, C.M. Lim, J.S. Suri, Heart rate variability: a review, Med. Biol. Eng. Comput. 44 (2006) 1031–1051.

[11] 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.

[12] C.-K. Peng, S. Havlin, J.M. Hausdorff, J.E. Mietus, H.E. Stanley, A.L. Goldberger, Fractal mechanisms and heart rate dynamics, J. Electrocardiol. 28 (1995) 59–65.

[13] 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.

[14] J. Jamsek, A. Stefanovska, P.V.E. McClintock, Nonlinear cardio-respiratory interactions revealed by time-phase bispectral analysis, Phys. Med. Biol. 49 (2004) 4407–4425.

[15] D. Ge, N. Srinivasan, S.M. Krishnan, Cardiac arrhythmia classification using autoregressive modeling, Biomed. Eng. Online 1 (2002).

[16] K. Tanaka, A.R. Hargens, Wavelet packet transform for R-R interval variability, Med. Eng. Phys. 26 (2004) 313–319.

[17] H.M. Tun, Analysis of heart rate variability based on quantitative approach, MOJ Proteomics Bioinform. 7 (2018).

[18] G.B. Moody, A.L. Goldberger, S. McClennen, S.P. Swiryn, Predicting the onset of paroxysmal atrial fibrillation: the computers in cardiology challenge 2001, Comput. Cardiol. (2001) 113–116.

[19] A. Narin, Y. Isler, M. Ozer, M. Perc, Early prediction of paroxysmal atrial fibrillation based on short-term heart rate variability, Phys. Stat. Mech. Appl. 509 (2018) 56–65.

[20] P. de Chazal, C. Heneghan, Automated assessment of atrial fibrillation, Comput. Cardiol. 28 (2001) 117–120 (Cat. No.01CH37287), IEEE, 2001.

[21] M. Carrara, L. Carozzi, T.J. Moss, M. De Pasquale, S. Cerutti, M. Ferrario, D.E. Lake, J.R. Moorman, Heart rate dynamics distinguish among atrial fibrillation, normal sinus rhythm and sinus rhythm with frequent ectopy, Physiol. Meas. 36 (2015) 1873–1888.

[22] PhysioNet, PAF Prediction Challenge Database, 2001.

[23] B. Hickey, C. Heneghan, Screening for paroxysmal atrial fibrillation using atrial premature contractions and spectral measures, in: Comput. Cardiol., IEEE, 2002, pp. 217–220.

[24] 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.

[25] K.H. Boon, M. Khalil-Hani, M.B. Malarvili, C.W. Sia, Paroxysmal atrial fibrillation prediction method with shorter HRV sequences, Comput. Methods Program. Biomed. 134 (2016) 187–196.

[26] K.H. Boon, M. Khalil-Hani, M.B. Malarvili, Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III, Comput. Methods Progr. Biomed. 153 (2018) 171–184.

[27] G.D. Clifford, L. Tarassenko, Quantifying errors in spectral estimates of HRV due to beat replacement and resampling, IEEE Trans. Biomed. Eng. 52 (2005) 630–638.

[28] F.K. Shafqat, S.S.K. Pal, T.P.A. Kyriacou, Evaluation of two detrending techniques for application in heart rate variability, Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc. (2007) 267–270.

[29] L. Li, C. Liu, K. Li, C. Liu, Comparison of detrending methods in frequency domain analysis of R-R interval series, Appl. Mech. Mater. 128–129 (2012) 1359–1362.

[30] J.T. VanderPlas, Understanding the lomb–scargle periodogram, Astrophys. J. Suppl. 236 (2018) 16.

[31] S.N. Yu, M.Y. Lee, Bispectral analysis and genetic algorithm for congestive heart failure recognition based on heart rate variability, Comput. Biol. Med. 42 (2012) 816–825.

[32] N.R. Lomb, Least-squares frequency analysis of unequally spaced data, Astrophys. Space Sci. 39 (1976) 447–462.

[33] M.A. Woo, W.G. Stevenson, D.K. Moser, R.B. Trelease, R.M. Harper, Patterns of beat-to-beat heart rate variability in advanced heart failure, Am. Heart J. 123 (1992) 704–710.

[34] M. Brennan, M. Palaniswami, P. Kamen, Do existing measures of Poincareé plot geometry reflect nonlinear features of heart rate variability? IEEE Trans. Biomed. Eng. 48 (2001) 1342–1347.

[35] G.D. Clifford, Signal Processing Methods for Heart Rate Variability, University of Oxford, 2002.

[36] K.C. Bilchick, R.D. Berger, Heart rate variability, J. Cardiovasc. Electrophysiol. 17 (2006) 691–694.

[37] T.G. Farrell, Y. Bashir, T. Cripps, M. Malik, J. Poloniecki, E.D. Bennett, D.E. Ward, A.J. Camm, Risk stratification for arrhythmic events in postinfarction patients based on heart rate variability, ambulatory electrocardiographic variables and the signal-averaged electrocardiogram, J. Am. Coll. Cardiol. 18 (1991) 687–697.

[38] L.C.M. Vanderlei, C.M. Pastre, I.F. Freitas Júnior, M.F. de Godoy, Índices geométricos de variabilidade da frequência cardíaca em crianças obesas e eutróficas, Arq. Bras. Cardiol. 95 (2010) 35–40.

[39] J.S. Richman, J. Randall Moorman, J. Randall, M. Physi, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol. Heart Circ. Physiol. 278 (2000) 2039–2049.

[40] J.M. Yentes, N. Hunt, K.K. Schmid, J.P. Kaipust, D. McGrath, N. Stergiou, The appropriate use of approximate entropy and sample entropy with short data sets, Ann. Biomed. Eng. 41 (2013) 349–365.

[41] P. Laguna, G.B. Moody, R.G. Mark, Power spectral density of unevenly sampled data by least-square analysis: performance and application to heart rate signals, IEEE Trans. Biomed. Eng. 45 (1998) 698–715.

[42] C. Torrence, G.P. Compo, A practical guide to wavelet analysis, Bull. Am. Meteorol. Soc. 79 (1998) 61–78.

[43] U. Wiklund, M. Akay, U. Niklasson, Short-term analysis of heart-rate variability by adapted wavelet transforms, IEEE Eng. Med. Biol. Mag. 16 (1997).

[44] A. Boardman, F.S. Schlindwein, A.P. Rocha, A. Leite, A study on the optimum order of autoregressive models for heart rate variability, Physiol. Meas. 23 (2002) 325–336.

[45] M. Kuhn, K. Johnson, Applied Predictive Modeling, 2013.

[46] M.M. Mukaka, Statistics corner: a guide to appropriate use of correlation coefficient in medical research, Malawi Med. J. 24 (2012) 69–71. https://www.ajol.info/index.php/mmj/article/view/81576.

[47] X.W. Chen, J.C. Jeong, Enhanced recursive feature elimination, in: Proc. – 6th Int. Conf. Mach. Learn. Appl. ICMLA 2007, 2007, pp. 429–435.

[48] R.-C. Chen, C. Dewi, S.-W. Huang, R.E. Caraka, Selecting critical features for data classification based on machine learning methods, J. Big Data. 7 (2020) 52.

[49] E.M. Senan, M.H. Al-Adhaileh, F.W. Alsaade, T.H.H. Aldhyani, A.A. Alqarni, N. Alsharif, M.I. Uddin, A.H. Alahmadi, M.E. Jadhav, M.Y. Alzahrani, Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques, J. Healthc. Eng. 2021 (2021) 1–10.