Using PPG Signals and Wearable Devices for Atrial Fibrillation Screening

Chengming Yang, César Veiga, Juan J. Rodríguez-Andina, *Senior Member, IEEE*, José Fariña, *Member, IEEE*, Andrés Iñiguez, and Shen Yin, *Senior Member, IEEE*

***Abstract*—Cardiovascular diseases are the primary cause of deaths in the world. Atrial fibrillation (AF) is the most common type of cardiac arrhythmia. Due to its high prevalence and associated risks, early detection of AF is an important objective for healthcare systems worldwide. The growing demand for medical assistance implies increased expenses, which could be limited by implementing ambulatory monitoring techniques based on wearable devices, thus reducing the number of people requiring observation in hospitals. One of the main challenges in this context is related to the large amount of data from patients to be analyzed, which points to the suitability of using computational intelligence techniques for it. The selection of the features to be extracted from data plays a key role in order for any classifier of heart rhythm to provide good results in this regard. This paper demonstrates that it is possible to achieve an accurate detection of AF using a very low number of relatively simple features extracted from photoplethysmographic signals, enabling the use of affordable wearable devices (with scarce processing and data storage resources) with this purpose over long periods of time. This fact has been validated in experiments using data from real patients under medical supervision.**

***Index Terms*—Atrial fibrillation, photoplethysmography, ambulatory screening, feature selection, wearable devices.**

I. Introduction

WITH population aging, the incidence of heart diseases is very high [1], threatening health at a worldwide scale. According to the World Health Organization, cardiovascular diseases are the primary cause of death in Western countries [2], and reasonable predictions [3] state that the situation will become worse. In this scenario, a growing demand for medical assistance implies an increase in the number of people requiring assistance in hospitals (and, in turn, of cost), which can be mitigated by implementing ambulatory monitoring techniques.

Atrial fibrillation (AF) is the most common type of cardiac arrhythmia [4]. As such, it has received a lot of attention from the research community. The analysis of heart rhythm to detect cardiac pathologies such as AF has been usually based on electrocardiogram (ECG) signals [5]-[9] recorded with Holter monitors. An important drawback of those approaches is that the (expensive) monitoring devices can only be attached to patients' bodies in hospital by specialized personnel. In the case of implantable devices, their invasive nature adds to the high cost. Hence, it is very inconvenient and costly to use these approaches for the analysis of heart rhythms aiming at early detection of heart diseases at a massive population scale, as increasingly required.

An alternative solution to ECG that is becoming highly popular is the use of photoplethysmography (PPG) signals, which can be acquired using wearable or portable devices [10] without the need for specialized personnel, opening the door for affordable ambulatory monitoring of patients at a massive population scale. Although ECG heart rate measurements are more accurate than PPG ones (0.27 vs 1.75 beats per minute in the experiments reported in [11] during physical activity), PPG measurements are accurate enough for screening purposes. Zhang *et al.* [12] proposed a general framework to estimate heart rate during fast running. Alqaraawi *et al.* [13] proposed a probabilistic approach based on Bayesian learning to estimate heart rate variability from PPG signals. Dao *et al.* [14] presented a robust motion artifact detection algorithm to estimate heart rates from PPG signals. Corino *et al.* [15] analyzed blood volume pulse signals and developed an algorithm for discriminating AF from normal sinus rhythm (NSR) or other arrhythmias. They used the sequential forward floating search algorithm for feature selection and a k-nearest neighbor classifier, achieving an accuracy around 90% for AF and NSR classification. Bonomi *et al.* [16] proposed the use of a wristband wearable device for AF detection in an ambulatory context. They store PPG signals acquired at 125 samples per second for later analysis and classification using Hidden Markov models, achieving an accuracy higher than 93%. In this context, reducing the amount of data to be saved is very important, since in this way patients could be monitored for longer periods of time. This in turn would increase the ability of wearable-based systems to detect the so-

Manuscript received August 2, 2018; revised November 13, 2018; accepted December 9, 2018.

Ch.Yang (corresponding author) and Sh.Yin are with the Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, Harbin 150001, China (email: chmyang@yeah.net, shen.yin2011@googlemail.com).

C.Veiga is with Instituto de Investigación Sanitaria Galicia Sur, Vigo 36312, Spain (email: Cesar.Veiga.Garcia@sergas.es).

J.J.Rodríguez-Andina and J.Fariña are with the Department of Electronic Technology, University of Vigo, Vigo 36310, Spain (email: {jjrdguez, jfarina}@uvigo.es).

A.Iñiguez is with the Cardiology Department, Álvaro Cunqueiro Hospital, Vigo 36312, Spain (email: Andres.Iniguez.Romo@sergas.es).

called paroxysmal AF. This type of AF is silently affecting an increasing number of individuals, so its detection is of paramount importance for health systems all over the world. Since paroxysmal AF is inherently intermittent, achieving longer monitoring times is a fundamental requirement for any ambulatory detection system.

Because of the huge amount of data involved in heart rhythm analysis (particularly in an ambulatory context, where data must be collected over long periods of time to enable paroxysmal AF detection), computational intelligence techniques are very suitable to carry it out. The selection of features to be extracted from data plays a key role in order for any classifier to provide good results [17]. In the frequency domain, feature extraction based on wavelet transform is widely used for signal analysis and processing [14][18][19]. Previously reported works use different sets of time- and frequency-domain features, but to the best of authors' knowledge an analysis of the most suitable ones for heart rhythm classification from PPG signals captured using wearable devices has not been reported in the literature.

The main contribution of this article is to demonstrate the feasibility of obtaining very good AF classification accuracy with simple and cheap wearable devices, enabling affordable AF monitoring over long periods of time in an ambulatory context at a wide population scale. For this, the article identifies the most suitable frequency- (wavelet-based) and time-domain features targeting PPG-based heart rhythm analysis for AF detection with wearable devices. Thanks to this, it demonstrates that a very accurate detection of AF can be achieved using a very low number of relatively simple features extracted from PPG signals. It is very important to highlight that, in the target ambulatory context, the goal is not to achieve the maximum possible classification accuracy, but good enough accuracy for screening purposes with devices as simple and affordable as possible. It is also very important to note that, as demonstrated by the experimental results in Section IV, the achieved classification accuracy is better than that reported in recent works, such as [15]. With regard to the results in [16], the proposed approach provides similar accuracy but much less data have to be saved in the wearable device, so monitoring times can be much longer. Therefore, the present paper provides the most efficient solution to date for ambulatory monitoring and detection of AF (including paroxysmal AF).

Using data obtained from real patients under medical supervision, several sets of experiments are conducted. First, the best wavelet transform configuration parameters and time slot sizes PPG signals can be split in are determined. Then it is shown that features can be listed according to their relevance regardless of the classifier used and that different tradeoffs can be defined between the number of features analyzed and the classification accuracy achieved. Given the limited amount and complexity of resources available in wearable devices, such tradeoffs have a major impact on the practical applicability of PPG signal monitoring for ambulatory AF screening. Finally, it is shown that classification accuracy over 90% can be achieved using a very reduced number of simple features requiring only basic computations to be performed, enabling the use of affordable wearable devices (with scarce resources) with this purpose over long periods of time.

The remainder of the paper is organized as follows. Section II briefly reviews the fundamentals of PPG signals in wearable devices and wavelet transform, and describes how the most relevant features can be identified. The methodology followed is described in Section III. Experimental results supporting the claimed contributions of the paper are presented and discussed in Section IV. Finally, Section V summarizes the conclusions.

II. Background

A. PPG signals in wearable devices

PPG is a simple, low-cost optical technique that allows volumetric measurements of an internal biological part to be carried out by illuminating the skin and measuring changes in light absorption or reflection [20]. It can be used to detect volume changes in blood vessels caused by heart beats. The period of the PPG signal corresponds to that of heart beats, and thus heart rate can be estimated from it.

The location of the sensor significantly influences the quality of the signals. Sensing in the fingers, as done e.g. in intensive care units, provides better signal quality, but is not considered in this work because of its bad usability in an ambulatory context, where patients have to wear the sensing device for long, continuous periods of time (during "normal" life). Sensing in the wrist is a suitable alternative in this context. Both wrist surfaces (ventral or dorsal) are usable, each one having its own pros and cons. The ventral position provides in general better signal quality, but signal-to-noise ratio is greatly degraded in the presence of patients' movements. Therefore, in this work the dorsal position has been used.

PPG signals may be obtained from pulse oximeters, which can be easily integrated in wearable wristband devices. Fig. 1 shows the one used in this work, based on an AFE4403 evaluation module from TI (whose block diagram is also shown in Fig. 1) and custom middleware specifically developed to configure the PPG data acquisition process. Blocks inside the dashed area are the ones used when monitoring patients. The inertial measurement unit (IMU) is not used in this work, and the USB port circuitry is only used when connected to a computer for configuration or for downloading data. Sample actual waveforms measured from real patients in both AF and NSR states are depicted in Fig. 2.

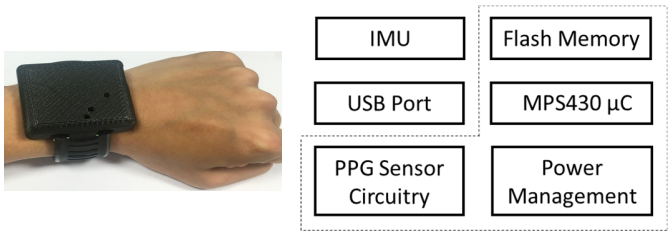

Fig. 1. Wearable device used in this work (left); block diagram (right).

B. Wavelet transform

The wavelet transform is a well-known analysis method in signal processing [21]. It is based on a generating function,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2018.2889614, IEEE Transactions on Industrial Electronics

called mother wavelet, which allows the signal under analysis to be projected in a set of frequency bands. The results of the analysis vary significantly depending on the mother wavelet function used, so its selection is of paramount importance. This is an issue for which no systematic solution exists, and it is typically based on previous experience or on experimental validation, like in the case of this work (Section IV.A).

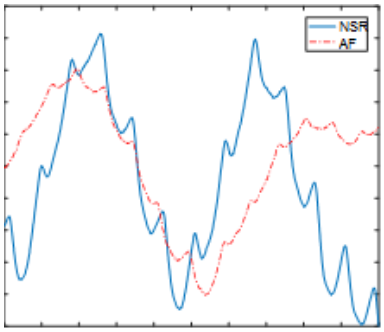

Fig. 2. PPG signals for 10 seconds of NSR and AF rhythms in one of the real patients monitored in this work.

Using the wavelet transform, a multi-resolution analysis can be performed. Information in a signal can be considered to be the superimposition of coarse (approximate) information (obtained by low-pass filtering it) and detail information (obtained by high-pass filtering it). Detail information can be in turn decomposed in a similar way, and the process repeated, obtaining successive levels of decomposition, as shown in Fig. 3, where *cAj* and *cDj* are the approximate and detail coefficients of the decomposition in level *j*, respectively.

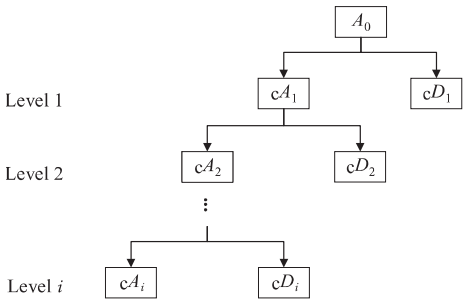

Fig. 3. Coefficients of multilevel wavelet decomposition.

C. SVM-RFE

As stated in Section I, computational intelligence techniques are very suitable for analyzing PPG signals targeting heart rhythm classification. It is important to note that this work evaluates features, *not* classifiers. With this purpose, the well-known, very popular Support Vector Machines (SVMs) [22] are used to separate PPG signals corresponding to NSR (healthy condition) from those corresponding to AF (pathological condition). SVMs generate separators (hyperplanes) for input samples, either in their original space or using the so-called kernel functions. For the sake of simplicity, a linear kernel is first used. The generality of the results obtained is confirmed by using other two kernels (polynomial and radial basis function, RBF), as discussed in Section IV-D.

In order for the effectiveness of different signal features to be evaluated, the SVM Recursive Feature Elimination (SVM-RFE) technique (Table I), which was proposed in a work also dealing with a medical application [23], is used. The goal is to select the minimum set of features that achieve optimal classification accuracy [24]. SVM-RFE allows features to be listed in descending order of contribution to classification accuracy (what is called ranking list) using a linear kernel. In each iteration, the last feature on the ranking list is eliminated. In the first iterations, deleted features are typically noisy, redundant, or irrelevant, whose use leads to non-optimal classification accuracy, mainly because of overfitting of the classifier. The process is finished when classification accuracy decreases after elimination of the last feature on the remaining list. Hence, the best possible classification accuracy is achieved with the minimum possible number of (relevant) features.

TABLE I
THE SVM-RFE ALGORITHM [23]

Input:
Training samples: $X_0 = [X_1, X_2, ..., X_k, ..., X_l]^T$
Class labels: $y_0 = [y_1, y_2, ..., y_k, ..., y_l]^T$
Initialize:
Surviving features: S = [1, 2, ... , n]
Features ranking list: r = []
Repeat until s = []
Restrict training samples to good feature indices: $X = X_0(:,s)$
Train the classifier: α = SVM − train(X, y)
Calculate weight vector of dimension length(s): $\omega = \sum_k \alpha_k \cdot y_k \cdot x_k$
Calculate the ranking criteria: $c_i = (\omega_i)^2$, for all i
Identify feature with the least significant ranking: f = argmin(c)
Update feature ranking list: r = [s(f), r]
Remove the identified feature: s = s(1 : f − 1,f + 1 : length(s))
end
Output:
Ranking list: r

D. Feature relevance evaluation

The evaluation of the effectiveness of the different features must be based on suitable metrics [25][26]. In this case, the sizes of the data sets to be analyzed (corresponding to NSR and AF, respectively) are large enough and balanced, and a few incorrect classifications will not affect the quality of the classifier. This is different from some applications where data sets are highly unbalanced (such as cybersecurity, where most data that can be gathered correspond to no-attack situations). Therefore, in this case a simple approach based on the computation of true / false positive / negative classification results can be used. True positives and negatives (TP / TN) correspond to the cases where the predicted class matches the actual (true) class of the instances and false positives and negatives (FP / FN) to the opposite cases. Since the target here is the detection of AF, TP correspond to the cases where an AF instance is correctly (true) classified as such (positive to AF) and TN to the cases where NSR is correctly (true) classified as such (negative to AF). Similarly, FP correspond to the cases where a NSR instance is incorrectly (false) classified as AF (positive to AF) and FN to the cases where an AF instance is incorrectly (false) classified as NSR (negative to AF).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2018.2889614, IEEE Transactions on Industrial Electronics

From these definitions, the classification accuracy can be computed as:

$$A_C = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

The sensitivity or true positive rate (TPR), which corresponds here to the percentage of AF instances correctly identified, can be computed as:

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

and the specificity or true negative rate (TNR), which corresponds here to the percentage of NSR instances correctly identified, as:

$$TNR = \frac{TN}{TN + FP} \quad (3)$$

III. Methodology

The methodology followed to obtain the experimental results presented and discussed in Section IV corresponds to the processing pipeline depicted in Fig. 4.

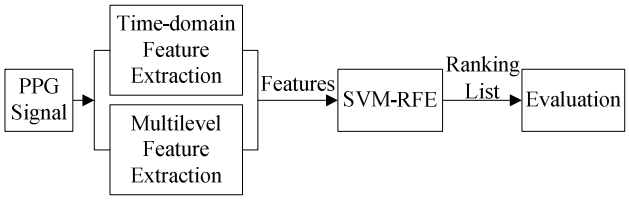

Fig. 4. The processing pipeline.

The features analyzed in this work have been obtained from the 9 statistical parameters most frequently used in related works, i.e., the well-known median, mean, standard deviation, and variance, plus the five following ones:

- Shannon entropy: $SE = -\sum_{x=1}^{n} p(x) \cdot log_2[p(x)]$ (4)
- Energy: $E = \sum_{x=1}^{n} [p(x)]^2$ (5)
- Contrast: $CON = \sum_{x=1}^{n} (1-x)^2 \cdot p(x)$ (6)
- Inverse different moment: $IDM = \sum_{x=1}^{n} \frac{1}{1+(1-x)^2} \cdot p(x)$ (7)
- Homogeneity: $HOM = \sum_{x=1}^{n} \frac{1}{1+|1-x|} \cdot p(x)$ (8)

In these equations p(x), $1 \leq x \leq n$, are either raw PPG signal samples (resulting in time-domain features) or the coefficients of multilevel wavelet transforms applied to the raw signals (resulting in frequency-domain features).

As shown in Fig. 4, frequency-domain features are obtained from the Multilevel Feature Extraction block (time-domain features may be considered 'level 0', but for the sake of clarity they are considered separately). Multilevel feature extraction is performed as described in Table II, and it can be configured to work with raw data divided in time slots of different sizes and to use different wavelet families and decomposition levels, as discussed in Sections IV-B and IV-A, respectively.

The usefulness of the different features is analyzed in terms of classification accuracy, using the SVM-RFE technique (described in Section II-C), several classifiers, and the metrics discussed in Section II-D.

TABLE II
Procedure for Multilevel Feature Extraction

Input: PPG Signal: *p(x)*
Initialize:
  Set parameters:
    Time slot: *t* → divide *p(x)* in time slots of duration t
    Wavelet family: select the wavelet function to be used
    Decomposition level: *n*
  For each time slot:
    For k = 1 to n:
      Extract the 9 frequency-domain features from the coefficients of the current level of the wavelet transform
Output: Feature vector: *f*

IV. Experimental Results

Experiments have been conducted to evaluate the usefulness (contribution to heart rhythm classification between NSR and AF) of the different time- and frequency-domain features considered. As discussed in Section II-C, for the sake of simplicity a linear kernel SVM has been first chosen as classifier. It is important to emphasize again that the main objective of this work is to evaluate features, *not* classifiers.

Data obtained from 11 real patients in hospital (with their informed consent and after approval of the ethics committee) have been used in all experiments. These patients were 75% males, age 63±12 years, and around 20% had a reduced (<45%) left ventricle ejection fraction. They entered hospital in AF state, and were recovered to NSR state by means of electrical cardioversion [27]. Since they were continuously monitored by ECG equipment under medical supervision and, simultaneously, with the wearable device including the PPG sensor (Fig. 1), it was possible to identify the parts of the PPG signals corresponding to AF and NSR respectively.

The total times each patient was in NSR and AF states during the experiments are listed on Table III. A "global" patient was also created by merging data from all actual patients together, to obtain a dataset not biased by the specific characteristics of individuals. For each patient (including the "global" one) PPG signals were split in two separate datasets (including both AF and NSR data on each of them) to extract features for training and testing, respectively. Training and testing datasets were obtained by randomly taking 75% of the time slots for training and 25% for testing. It can be seen in Table III that, considering 10s intervals, the amount of time slots to be analyzed for both AF and NSR states is large enough for the metrics considered for feature evaluation (Section II-D) to be valid in this context.

TABLE III
Times in NSR and AF States of the 11 Patients Monitored at Hospital

Patient	NSR state	AF state
1	3880 seconds	2800 seconds
2	5050 seconds	1330 seconds
3	6410 seconds	16360 seconds
4	2850 seconds	1340 seconds
5	1890 seconds	910 seconds
6	7080 seconds	2160 seconds
7	5850 seconds	5140 seconds
8	560 seconds	4860 seconds
9	3030 seconds	7770 seconds
10	4530 seconds	5900 seconds
11	6490 seconds	36920 seconds

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2018.2889614, IEEE Transactions on Industrial Electronics

Four sets of experiments have been performed. First, the wavelet family and decomposition level that provide the best features for PPG signal analysis with heart rhythm classification purposes have been identified (Section IV-A). Using this wavelet family and decomposition level, a second set of experiments has allowed the time slot size providing the best results in this context to be determined (Section IV-B). After that, using the wavelet family, decomposition level, and time slot size identified as the best in the previous experiments, the usefulness of the different time- and frequency-domain features considered has been evaluated, and the feasibility of achieving different interesting tradeoffs between number of features and classification accuracy has been proven (Section IV-C). Finally, to demonstrate the generality of the results obtained, the third set of experiments has been repeated using other two classifiers, namely polynomial and RBF kernel SVMs (Section IV-D).

A. Analysis of different wavelet families

As mentioned above, the first set of experiments was conducted with the objective of identifying the wavelet family and decomposition level that provide the best features for PPG signal analysis with heart rhythm classification purposes. Different wavelet families require different filter lengths, which affects the maximum achievable decomposition level. The following wavelet families (and filter lengths) were analyzed: Haar (filter length 2), Db4 (length 8), Sym4 (length 8), Coif4 (length 24), Bior4.4 (length 10), Rbio4.4 (length 10), and Dmey (length 62). For feature extraction, raw PPG signals were sampled at 100 Hz with 3-Byte resolution and split in 10s time slots, resulting in data length of 1,000 samples per slot. The maximum decomposition level for each family is:

$$max\_level = int\left(\log_2\left[\frac{data\_length}{(filter\_length - 1)}\right]\right) \qquad (9)$$

which yields level 9 for Haar, 7 for Db4 and Sym4, 6 for Bior4.4 and Rbio4.4, 5 for Coif4, and 4 for Dmey. Features were extracted from wavelet coefficients for all levels, from 1 to the maximum possible for each family.

As shown in Fig. 3, for n-level decomposition, n+1 sets of coefficients are obtained, namely cA_n and cD_1 to cD_n. For each set of coefficients, the 9 features listed in Section III are extracted, resulting in a feature set of (n + 1) × 9 elements. One such set is obtained for every 10s time slot of the PPG signal. After that, the SVM-RFE algorithm is applied to obtain the ranking lists, determining classification accuracy with a linear kernel SVM. The results of the experiment are shown in Table IV, where those for each patient have been obtained by training the classifier with the features extracted from part of his / her own PPG signals and then testing it with the remaining features from that patient (i.e., obtaining the best possible classifier for each patient), and choosing the optimal feature set determined by means of SVM-RFE (consisting of F features, which may be different for each classifier). The fact that in most cases F < (n + 1) × 9 (in many cases <<) is an indication of the presence of noisy, redundant, or irrelevant features that cause overfitting of the classifier and must therefore be removed for classification accuracy to be maximized, as explained in Section II-C.

From a practical viewpoint, the most significant results are those for the "global" patient, since they correspond to a variety of individuals (as it will happen in a real medical context) and, therefore, they are not biased by specific characteristics of each patient. In any case, by analyzing the performance of the different wavelet families for each patient and for all of them combined, from Table IV it is clear that in nearly all cases the best option is to use the Haar wavelet at the maximum possible decomposition level. Hence, the choice of the best wavelet family and decomposition level is clear, regardless of the scenario (single patients or "global" one): Haar wavelet, decomposed until the maximum possible level (9 in this case). Note that results for the "global" patient are worse than for individual ones, which highlights the fact that the characteristics of NSR and AF signals are not exactly the same for all patients, so when analyzing them together classification accuracy is degraded with regard to that achieved by considering each patient separately (but, again, this is the realistic scenario in practice).

It is also interesting to note that the best results are obtained in most of the cases for the maximum decomposition level of the different wavelet families.

A. Analysis of different time slots

The second set of experiments is intended to determine the time slot size (to split PPG signals in time slots) with which the best classification results are achieved, while complying

TABLE IV
CLASSIFICATION ACCURACY (%) OF EACH PATIENT FOR HIS / HER OPTIMAL FEATURE SET (F FEATURES) AND BEST DECOMPOSITION LEVEL (L) OF EACH WAVELET FAMILY

Patient	Bior4.4			Coif4			Db4			Dmey			Haar			Rbio4.4			Sym4		
	F	L	Ac	F	L	Ac	F	L	Ac	F	L	Ac	F	L	Ac	F	L	Ac	F	L	Ac
1	60	6	92.22	21	2	88.62	57	6	94.01	36	3	88.62	66	9	**98.20**	63	6	91.62	54	5	94.01
2	63	6	98.13	45	4	98.75	72	7	98.75	45	4	98.75	15	9	**99.38**	63	6	98.13	69	7	98.75
3	36	5	**100.0**	15	3	**100.0**	27	6	**100.0**	36	4	**100.0**	21	9	**100.0**	12	4	**100.0**	18	4	**100.0**
4	21	6	**100.0**	18	5	**100.0**	24	7	**100.0**	42	4	**100.0**	15	9	**100.0**	21	6	**100.0**	48	7	**100.0**
5	9	1	95.74	33	3	95.71	42	6	98.57	45	4	95.71	6	9	**100.0**	63	6	**100.0**	60	6	**100.0**
6	60	6	85.71	36	3	85.71	36	7	85.71	6	1	85.71	6	7	85.71	54	5	**86.15**	3	3	85.71
7	18	6	**100.0**	15	5	**100.0**	18	7	**100.0**	6	4	**100.0**	3	9	**100.0**	18	6	**100.0**	27	7	**100.0**
8	21	6	**100.0**	18	4	**100.0**	18	7	**100.0**	6	4	**100.0**	3	9	**100.0**	21	6	**100.0**	69	7	**100.0**
9	63	6	87.78	45	5	85.56	72	7	88.89	42	4	83.33	57	7	88.52	63	6	**93.70**	72	7	90.74
10	51	5	84.67	54	5	87.36	63	6	92.34	45	4	81.99	45	5	**96.17**	63	6	91.19	66	7	89.27
11	27	2	83.89	15	1	83.70	45	4	84.44	15	1	83.52	75	9	**90.33**	33	3	84.35	18	1	83.89
Global	63	6	76.64	54	5	72.33	72	7	78.30	45	4	66.65	78	9	**79.97**	63	6	77.84	66	7	78.01

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2018.2889614, IEEE Transactions on Industrial Electronics

with some practical restrictions. Time slots have to be short enough for the moments when arrhythmic events happen to be accurately identified, but as long as possible to reduce their number and, in turn, the amount of computations to be performed and of data to be stored in the wearable device. With these restrictions in mind, from the conclusions of the previous set of experiments, the Haar wavelet with maximum decomposition level has been used to extract features from data sets consisting of 5, 10, 15, and 20s of PPG signals, respectively. Table V shows the classification accuracy obtained for each patient and time slot size with a linear kernel SVM classifier and the optimal feature set (consisting of F features) as determined using SVM-RFE. It is clear that in most cases the best results are obtained when dividing PPG signals in 10s time slots (and they are very close to the best in most others). It may be argued that results for the "global" patient are slightly better using 15s time slots. However, as discussed in next section, results for individual patients must also be taken into account so, both aspects considered, using 10s time slots is the best option.

TABLE V
CLASSIFICATION ACCURACY (%) FOR DIFFERENT TIME SLOTS USING HAAR WAVELET WITH MAXIMUM DECOMPOSITION LEVEL

Patient	5 seconds		10 seconds		15 seconds		20 seconds	
	F	Ac	F	Ac	F	Ac	F	Ac
1	69	92.81	66	**98.20**	75	93.75	72	97.62
2	66	**100.0**	15	99.38	63	99.07	12	**100.0**
3	6	**100.0**	21	**100.0**	21	99.47	15	**100.0**
4	63	98.57	15	**100.0**	3	**100.0**	3	**100.0**
5	18	**100.0**	3	**100.0**	57	**100.0**	12	**100.0**
6	9	87.66	75	84.85	12	**88.31**	84	87.07
7	3	**100.0**	3	**100.0**	12	**100.0**	3	**100.0**
8	60	**100.0**	3	**100.0**	60	98.90	6	**100.0**
9	36	84.63	48	**87.41**	57	82.22	84	84.56
10	69	91.00	42	**95.79**	69	90.80	51	91.60
11	78	89.91	75	**90.33**	78	87.29	78	89.13
Global	72	78.47	78	79.97	69	**80.02**	84	77.96

B. Analysis of time- and frequency-domain features

This set of experiments consists of three parts: analysis of the effectiveness of time-domain features, of frequency-domain features, and of time- and frequency-domain features combined for classification of NSR and AF rhythms from PPG signals:

- Frequency-domain features have been obtained using the optimal parameters for multilevel feature extraction identified in the two first sets of experiments, namely 10s time slots, Haar wavelet, maximum decomposition level, and the optimal set of features for the "global" patient. As stated in Section IV-A, in a real medical context the significant results in terms of selecting the best possible classifier are those for the "global" patient. The best configuration found for this patient should be the one used in practice and, according to Table IV, it corresponds to the use of 78 features. These same 78 features are also used now to analyze data for the 11 individual patients (Table VI). Although this is not the optimal configuration for each of them, the reason for this is to check the different responses that may be obtained in practice from different individuals. It is worth noting that a high classification accuracy for the "global" patient cannot be the only figure of merit to be taken into account. It is also important that classification accuracy be good in practice for as many individual patients as possible (or, conversely, classification should fail in practice for as few individual patients as possible, in terms of the metrics described in Section II-D).

TABLE VI
SENSITIVITY, SPECIFICITY, AND CLASSIFICATION ACCURACY (%) FOR EACH PATIENT BASED ON 78 FREQUENCY-DOMAIN FEATURES

Patient	TPR (%)	TNR (%)	Ac (%)
1	97.80	97.37	97.60
2	99.19	97.22	98.75
3	98.21	100.0	99.47
4	100.0	100.0	100.0
5	100.0	100.0	100.0
6	97.25	34.69	83.98
7	100.0	100.0	100.0
8	100.0	100.0	100.0
9	69.35	90.38	85.56
10	92.31	97.45	95.40
11	50.28	98.23	90.24
Global	69.59	88.30	79.97

- As explained in Section III, the 9 time-domain features are extracted from raw signals, also split in 10s time slots. The results of using them for the classification of heart rhythms of the patients are listed on Table VII. It can be seen that they are worse than those obtained with frequency-domain features (Table VI).

TABLE VII
SENSITIVITY, SPECIFICITY, AND CLASSIFICATION ACCURACY (%) FOR EACH PATIENT BASED ON THE 9 TIME-DOMAIN FEATURES

Patient	TPR (%)	TNR (%)	Ac (%)
1	93.41	77.63	86.23
2	100.0	66.67	92.50
3	100.0	100.0	100.0
4	95.77	100.0	97.14
5	97.67	88.89	94.29
6	99.45	32.65	85.28
7	100.0	100.0	100.0
8	100.0	99.21	99.26
9	41.94	89.90	78.89
10	86.54	82.80	84.29
11	6.63	99.67	84.16
Global	52.06	77.02	65.90

- Finally, experiments have been conducted using time- and frequency-domain features combined. The best results, listed on Table VIII, have been obtained by using 96 features out of the 99 extracted (9 × 10 in the frequency domain plus 9 in the time domain). They are in general better than those obtained using frequency-domain features only, but differences are very small to be significant, moreover considering that 96 features have to be used, compared to 78 for frequency-domain analysis. For portable or wearable solutions, this is an important point to consider because of its implication on the processing power required.

Results on Tables VI-VIII clearly show that classification accuracies for individual patients are in general close to those obtained with the optimal set of features for each patient on Table IV. It can also be noticed that results are very good with a few exceptions. And, once more, it has to be emphasized

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2018.2889614, IEEE Transactions on Industrial Electronics

that the focus of this paper is on evaluating features, *not* classifiers.

TABLE VIII
SENSITIVITY, SPECIFICITY, AND CLASSIFICATION ACCURACY (%) FOR EACH PATIENT USING 96 FEATURES (TIME- AND FREQUENCY-DOMAIN)

Patient	TPR (%)	TNR (%)	Ac (%)
1	98.90	98.68	98.80
2	99.19	97.22	98.75
3	98.81	100.0	99.65
4	100.0	100.0	100.0
5	100.0	100.0	100.0
6	96.70	36.73	83.98
7	100.0	100.0	100.0
8	100.0	100.0	100.0
9	69.35	91.35	86.30
10	94.23	96.82	95.79
11	51.93	98.12	90.42
Global	70.10	88.61	80.37

Regarding the specific features to be used, they can be identified in Fig. 5, which shows the position on ranking lists of the different time- and frequency-domain features when training and testing the classifier with data from the "global" patient. The **lower** the position on the vertical axes of the figure, the **more relevant** the feature (the more it contributes to classification). Threshold lines separate the features actually used from those discarded. As mentioned above, 78 features are used when only frequency-domain ones are considered, whereas 96 are used when both time- and frequency-domain ones are considered. Energy-, variance- and contrast-based features (e.g., those resulting from computing the energy, the variance, and the contrast, respectively, of PPG signal samples and of the coefficients of the wavelet transform) are the most relevant in both scenarios, whereas Shannon entropy-related are clearly the less relevant. Actually, Shannon entropy is totally discarded when using only frequency-domain features.

If, after reaching maximum classification accuracy, the ranking list is further reduced, in turn reducing processing and data storage requirements of the wearable device, this would be at the expense of decreased classification accuracy. It is interesting to analyze if this approach would lead to suitable tradeoffs. Results obtained by such feature elimination beyond the maximum accuracy point when using frequency-domain features are shown in Fig. 6. It is important to recall that the training and testing data sets are randomly selected in each experiment for a given feature set, as previously explained. Due to this the curve shows small peaks at some points. It can be noticed that by using 75 features, accuracy is negligibly degraded (-0.06%). Using 63 features (≈20% reduction) yields a 1.5% decrease in accuracy, and with 36 features (a very significant ≈54% reduction) accuracy decreases just 3%.

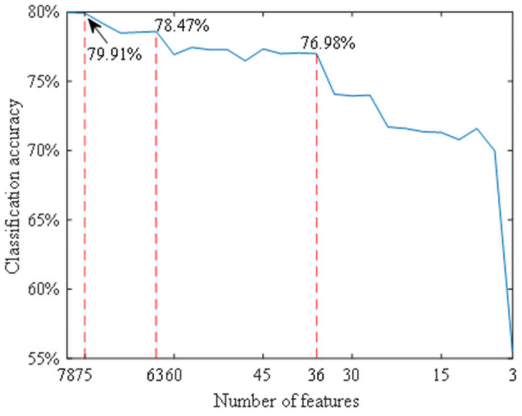

Fig. 6. Evolution of classification accuracy for the "global" patient when reducing the number of frequency-domain features beyond the maximum point.

C. Generality of the Results

To demonstrate the validity and usefulness for other classifiers of the same ranking lists obtained (in a simple way) with a linear kernel SVM, hence showing the general practical applicability of the feature selection approach followed, the analysis described in Section IV-C was repeated using polynomial (degree 3) and RBF (gamma 0.7) kernel SVMs.

The results obtained are summarized in Figs. 7 and 8. A first important conclusion is that the same evolution pattern can be identified in the accuracy of the three classifiers as the number of features decreases. First, classification accuracy increases as overfitting effects are removed. This has little influence for the linear kernel SVM, more noticeable for the polynomial kernel SVM, and very significant for the RBF kernel SVM. After that, variations in classification accuracy

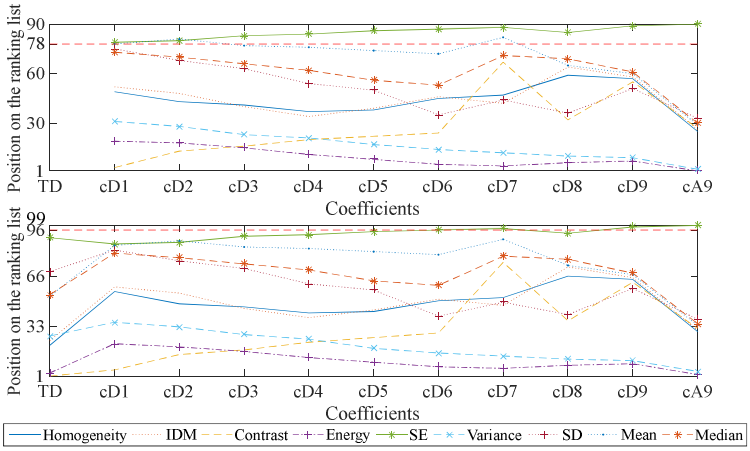

Fig. 5. Position on the ranking list of the features for the "global" patient, using frequency-domain features (top) and time- and frequency-domain features (bottom).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2018.2889614, IEEE Transactions on Industrial Electronics

are small, and finally it decreases sharply, as the most significant features are removed. These trends can be clearly identified in the curves, despite the peaks due to the aforementioned random selection of training and testing data sets for each feature set.

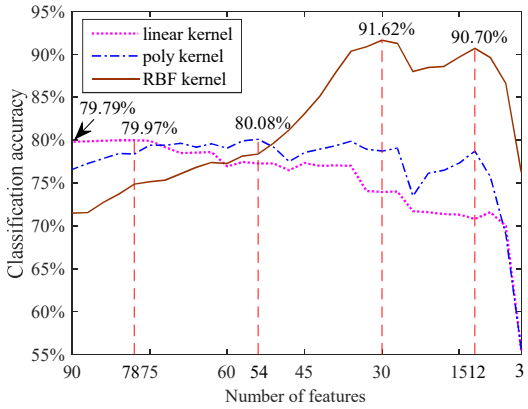

Fig. 7. Classification accuracy for the "global" patient using frequency-domain features and three different classifiers.

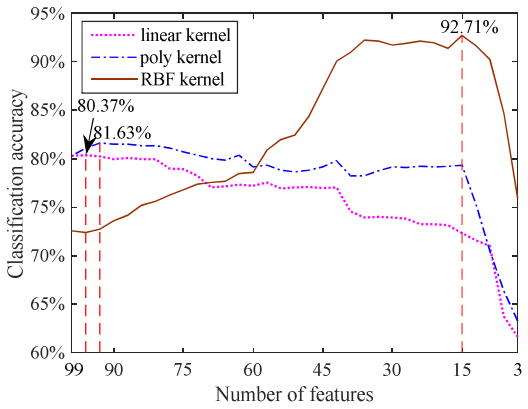

Fig. 8. Classification accuracy for the "global" patient using time- and frequency-domain features combined and three different classifiers.

A second important conclusion is that the most relevant features, which are energy-, variance-, and contrast-based, are easy to compute with simple processors, since they only require addition, subtraction, and multiplication operations. Hence, it is possible to extract them in the wearable device itself.

A third important conclusion is that the relatively flat region of the curves allows a significant number of features to be removed with little reduction in classification accuracy. Although the aim here is not to evaluate classifiers, it is clear that the RBF kernel SVM is the one providing the best results for the target application among the three tested. Just then focusing on it, it can be seen that by using only 12 frequency-domain features (Fig. 7) or 15 time- and frequency-domain features (Fig. 8), accuracy is close to the maximum and above 90%, better than that reported in [15] and very similar to that in [16]. In spite of the similar accuracy achieved, the proposed approach is more suitable for ambulatory screening than that in [16] because it requires much less data to be stored in the wearable device. Although some specific details about data acquisition are not provided in [16], still a fair comparison can be made in terms of data storage requirements. In [16], PPG signals are acquired at 125 samples per second and stored in the wearable device. Assuming just a minimum 1-Byte resolution, 125 Bytes are saved per second. In our case, 12 or 15 3-Byte features would be saved each 10s (the time slot size used for feature extraction), which means between 3.6 and 4.5 Bytes per second (between 34 and 27 times less data than in [16]). Therefore, if wearable devices with the same memory capacity were used, the proposed approach would allow much longer monitoring times to be achieved. As stated in Section I, this is a very important feature in an ambulatory context. Therefore, this work provides the most efficient solution to date for ambulatory AF monitoring and detection.

The flash memory block in Fig. 1 includes two 128 Mb chips. As mentioned in Section IV-A, 3-Byte values are sampled at 100Hz, which means PPG samples for nearly 30 hours can be stored. By extracting features in the wearable device (keeping 3-Byte resolution), using the 12 frequency-domain ones mentioned above, only 3.6 Bytes have to be saved per second. This means that data for 2,469 hours (or 102 days) could be stored. The interesting possibility of also implementing the classification algorithm in the wearable device and saving only data identified as corresponding to AF has been discarded because of two reasons. The main one is that it implies providing the device with enough computing power for the algorithm to be executed on it in real time, which conflicts with the goal of using devices as simple and cheap as possible. The second reason is that by recording massive amounts of data from real patients a better knowledge base can be obtained.

The microcontroller in the wearable device includes 128kB of on-chip flash memory. The (non-optimized) program code (stored in on-chip flash memory) requires 45kB, which leaves 83 kB for data storage, where features from 6.5 hours of operation can be accommodated. If a microcontroller with 256kB of on-chip flash memory is used, data for nearly 17 hours could be stored. By optimizing the code, close to 1-day data storage could be achieved. This would allow the flash memory chips to be removed from the circuit, significantly reducing its size as well as power consumption, which are very important issues in wearable devices.

It is thus clear that accurate and affordable AF screening using PPG signals and wearable devices is practically feasible if the right feature set is selected. The fact that only basic computations and a reduced number of features are required implies that the processing and data storage requirements available in relatively simple wearable devices are enough for the screening algorithms to be implemented in them.

V. Conclusions

The most relevant features for AF detection using PPG signals have been identified to be those based on energy, variance, and contrast, both in the time and frequency domains. It has also been found that the best results are obtained splitting PPG signals in 10s time slots and, in the frequency domain, using the Haar wavelet family with maximum decomposition level. In addition, it has been shown

that tradeoffs can be identified between number of features analyzed (which impacts the processing power and data storage resources required) and classification accuracy achieved.

The validity and generality of these conclusions have been confirmed by results obtained with three different classifiers and reinforced by the fact that data have been captured in a real-life scenario, from patients under medical supervision. These data have been analyzed for each patient and also for a "global" one created by combining data from all of them. In this way, a dataset not biased by the specific characteristics of individual patients (and, therefore, more representative of what the real situation in a medical context is) was obtained.

Individual analyses confirmed that the optimal configuration of the analysis for the "global" patient also provides good results for individual ones, with a few exceptions that require further investigation. These may be due to poor quality of the captured PPG signals or to the presence in them of heart rhythms others than AF and NSR.

By using a small feature set whose analysis requires only basic operations to be computed a very good AF detection accuracy can be achieved, so it is feasible to implement such analysis in resource-scarce, wearable devices, enabling affordable AF screening at a massive population scale in an ambulatory context, therefore contributing to the quality and sustainability of health systems worldwide. By comparison with recently published works in terms of classification accuracy and monitoring times, it has been demonstrated that the proposed approach is the most efficient solution to date for ambulatory AF monitoring and detection.

It is even possible to simplify the structure of the device, using only on-chip processor memory, therefore obtaining a simpler, smaller (hence more ergonomic) implementation with less power consumption.

ACKNOWLEDGMENT

The authors thank Dr. Enrique Garcia (whose patients were enrolled in this study) and Daniel Rivera, for their valuable suggestions to carry out the analyses presented in this article.

REFERENCES

[1] A.Vahanian *et al.*, "Guidelines on the management of valvular heart disease", *European Heart Journal*, vol. 28, no. 2, pp. 230-268, Jan. 2007.

[2] Online: http://www.who.int/cardiovascular_diseases/about_cvd/en/

[3] B.B.Hughes *et al.*, "Projections of global health outcomes from 2005 to 2060 using the International Futures integrated forecasting model", *Bulletin of the World Health Organization*, vol. 89, no. 7, pp. 478-486, Jul. 2011.

[4] P.Kirchhof *et al.*, "2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS", *European Heart Journal*, vol. 37, no. 38, pp. 2893-2962, Oct. 2016.

[5] V.P.Rachim and W.Y.Chung, "Wearable noncontact armband for mobile ECG monitoring system". *IEEE Trans. Biomedical Circuits and Systems*, vol. 10, no. 6, pp. 1112-1118, Dec. 2016.

[6] M.Meo, V.Zarzoso, O.Meste, D.G.Latcu, and N.Saoudi, "Spatial variability of the 12-lead surface ECG as a tool for noninvasive prediction of catheter ablation outcome in persistent atrial fibrillation", *IEEE Trans. Biomedical Engineering*, vol. 60, no. 1, pp. 20-27, Jan. 2013.

[7] C.T.Lin *et al.*, "An intelligent telecardiology system using a wearable and wireless ECG to detect atrial fibrillation", *IEEE Trans. Information Technology in Biomedicine*, vol. 14, no. 3, pp. 726-733, May 2010.

[8] O.Perlman, A.Katz, G.Amit, and Y.Zigel, "Supraventricular tachycardia classification in the 12-lead ECG using atrial waves detection and a clinically based tree scheme", *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 6, pp. 1513-1520, Nov. 2016.

[9] J.Lee, D.D.McManus, S.Merchant, and K.H.Chon, "Automatic motion and noise artifact detection in holter ECG data using empirical mode decomposition and statistical approaches", *IEEE Trans. Biomedical Engineering*, vol. 59, no. 6, pp. 1499-1506, Jun. 2012.

[10] P.-Y.Chiang, P.C.-P. Chao, D.-Ch.Tarng, and Ch.-Y.Yang, "A novel wireless photoplethysmography blood-flow volume sensor for assessing arteriovenous fistula of hemodialysis patients", *IEEE Trans. Industrial Electronics*, vol. 64, no. 12, pp. 9626-9635, Dec. 2017.

[11] D.Jarchi and A.J.Casson, "Towards photoplethysmography-based estimation of instantaneous heart rate during physical activity", *IEEE Trans. Biomedical Engineering*, vol. 64, no. 9, pp. 2042-2053, Sep. 2017.

[12] Z.Zhang, Z.Pi, and B.Liu, "TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise", *IEEE Trans. Biomedical Engineering*, vol. 62, no. 2, pp. 522-531, Feb. 2015.

[13] A.Alqaraawi, A.Alwosheel, and A.Alasaad, "Heart rate variability estimation in photoplethysmography signals using Bayesian learning approach", *Healthcare Technology Letters*, vol. 3, no. 2, pp. 136-142, Jun. 2016.

[14] D.Dao *et al.*, "A robust motion artifact detection algorithm for accurate detection of heart rates from photoplethysmographic signals using time-frequency spectral features", *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 5, pp. 1242-1253, Sep. 2017.

[15] V.D.A.Corino *et al.*, "Detection of atrial fibrillation episodes using a wristband device", *Physiological Measurement*, vol. 38, no. 5, pp. 787-799, May 2017.

[16] A.G.Bonomi *et al.*, "Atrial fibrillation detection using a novel cardiac ambulatory monitor based on photo-plethysmography at the wrist", *Journal of the American Heart Association*, vol. 7, no. 15, pp. 1-17, Aug. 2018.

[17] T.W.Rauber, F.Boldt, and F.M.Varejao, "Heterogeneous feature models and feature selection applied to bearing fault diagnosis", *IEEE Trans. Industrial Electronics*, vol. 62, no. 1, pp. 637-646, Jan. 2015.

[18] B.M.Ebrahimi, M.J.Roshtkhari, J.Faiz, and S.V.Kathami, "Advanced eccentricity fault recognition in permanent magnet synchronous motors using stator current signature analysis", *IEEE Trans. Industrial Electronics*, vol. 61, no. 4, pp. 2041-2052, Apr. 2014.

[19] A.S.S.Vasan, B.Long, and M.Pecht, "Diagnostics and prognostics method for analog electronic circuits", *IEEE Trans Industrial Electronics*, vol. 60, no. 11, pp. 5277-5291, Nov. 2013.

[20] R.Yousefi, M.Nourani, S.Ostadabbas, and I.Panahi, "A motion-tolerant adaptive algorithm for wearable photoplethysmographic biosensors", *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 670-681, Mar. 2014.

[21] P.R.Haddad and A.N.Akansu, *Multiresolution signal decomposition: Transforms, subbands, and wavelets*, Academic Press, 1992.

[22] V.Kecman, *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*, MIT Press, 2001.

[23] I.Guyon, J.Weston, S.Barnhill, and V.Vapnik, "Gene selection for cancer classification using support vector machines", *Machine Learning*, vol. 46, no. 1-3, pp. 389-422, Jan. 2002.

[24] G.H.John, R.Kohavi, and K.Pfleger, "Irrelevant features and the subset selection problem", in *Proc. 11th Int. Conf. on Machine Learning*, pp. 121-129, 1994.

[25] H.J.Carey, M.Manic, and P.Arsenovic, "Epileptic spike detection with EEG using artificial neural networks", in *Proc. 9th IEEE Int. Conf. on Human System Interactions (HSI)*, pp. 89-95, 2016.

[26] K.Amarasinghe, P.Sivils, M.Manic, "EEG feature selection for thought driven robots using evolutionary algorithms", in *Proc. 9th IEEE Int. Conf. on Human System Interactions (HSI)*, pp. 355-361, 2016.

[27] P.M.Zoll, A.J.Linenthal, W.Gibson, M.H.Paul, and L.R.Norman, "Termination of ventricular fibrillation in man by externally applied electric countershock", *New England Journal of Medicine*, vol. 254, no. 16, pp. 727-732, Apr. 1956.

Chengming Yang received the B.E. degree from Northeast Agricultural University, Harbin, China, in 2010, and the M.E. degree from Kunming University of Science and Technology, Kunming, China, in 2013, both in in agricultural mechanization engineering. He is working towards the Ph.D. degree in control science and engineering at Harbin Institute of Technology.
His research interests include fault diagnosis, process monitoring and their applications to large-scale industrial processes, as well as medical screening based on wearable devices.

César Veiga received the M.Sc. and Ph.D. degrees in Physics from University of Santiago de Compostela, Spain, in 1993 and 1998, respectively. He is a senior researcher at Instituto de Investigación Sanitaria Galicia Sur, Vigo, Spain. He carried out research on biomedical engineering at the Institute for Systems Informatics and Safety (Italy), the Galician Supercomputing Center (CESGA), and some private companies. His research interests include processing tools and algorithms for analysis of biomedical signals and their application to cardiology.

Juan J. Rodríguez-Andina (M'00–SM'04) received the M.Sc. degree from Technical University of Madrid, Spain, in 1990, and the Ph.D. degree from University of Vigo, Spain, in 1996, both in electrical engineering.
He is an Associate Professor in the Department of Electronic Technology, University of Vigo. His research interests include the implementation of complex control and processing algorithms and intelligent sensors in embedded platforms.

José Fariña (M'04) received the M.Sc. and Ph.D. degrees in electrical engineering from University of Santiago de Compostela, Spain, in 1984 and 1989, respectively.
He is an Associate Professor in the Department of Electronic Technology, University of Vigo, Spain. His research interests include the implementation of complex control and processing algorithms and intelligent sensors in embedded platforms.

Andrés Iñiguez received a master's degree in Medicine and Surgery in 1980, Cardiology Specialist degree in 1986, and Ph.D. degree in 1987, all from Universidad Complutense, Madrid, Spain.
Dr. Iñiguez is the Head of the Cardiology Department at Álvaro Cunqueiro Hospital, Vigo, Spain. From 1993 to 2004 he was Chief of Interventional Cardiology at Fundación Jiménez Díaz, Madrid. From 1986 to 2004, he held positions as Associate Professor in Medicine at Universidad Autónoma and Universidad Complutense, Madrid. In 2016-2017 he was the President of the Spanish Society of Cardiology.

Shen Yin (M'12–SM'15) received the B.E. degree in automation from Harbin Institute of Technology, China, in 2004, the M.Sc. degree in control and information systems and the Ph.D. degree in electrical engineering and information technology from University of Duisburg-Essen, Germany in 2007 and 2012, respectively.
He is currently a Professor with the Research Center of Intelligent Control and Systems, Harbin Institute of Technology. His research interests include model-based and data-driven fault diagnosis and prognosis in process control, fault-tolerant control, and big data focused on industrial electronics applications.