


An automatic method for arterial pulse waveform recognition using KNN and SVM classifiers

Tânia Pereira¹  · Joana S. Paiva² · Carlos Correia¹ · João Cardoso¹

Received: 24 June 2014 / Accepted: 10 September 2015 / Published online: 24 September 2015
© International Federation for Medical and Biological Engineering 2015

Abstract The measurement and analysis of the arterial pulse waveform (APW) are the means for cardiovascular risk assessment. Optical sensors represent an attractive instrumental solution to APW assessment due to their truly non-contact nature that makes the measurement of the skin surface displacement possible, especially at the carotid artery site. In this work, an automatic method to extract and classify the acquired data of APW signals and noise segments was proposed. Two classifiers were implemented: k-nearest neighbours and support vector machine (SVM), and a comparative study was made, considering widely used performance metrics. This work represents a wide study in feature creation for APW. A pool of 37 features was extracted and split in different subsets: amplitude features, time domain statistics, wavelet features, cross-correlation features and frequency domain statistics. The support vector machine recursive feature elimination was implemented for feature selection in order to identify the most relevant feature. The best result (0.952 accuracy) in discrimination between signals and noise was obtained for the SVM classifier with an optimal feature subset.

Keywords Arterial pulse waveform · Optical system · Feature creation · Recursive feature elimination · K-nearest neighbour algorithm · Support vector machine

1 Introduction

The blood pulse pressure waveform has long been known as a fundamental medical signal. Several non-invasive techniques have been developed to detect cardiovascular pulsation, blood pressure and arterial pulse waveform [49]. The arterial stiffness, identified as the most important cardiovascular risk factor, is responsible for changes in mechanical impedance of the arterial system [7, 58], which affects the artery compliance and, therefore, the distension profile. Several hemodynamic parameters can be extracted from the APW analysis that allows the assessment of the arterial stiffness and cardiac function [11].

Current commercial devices dedicated to non-invasive measurement of the APW show great limitations inherent to the used methods (e.g. arterial applanation tonometry, ultrasound and plethysmography) that require direct contact with the patient skin at the artery site using small amounts of compression over the artery that lead to the distortion of the pulse waveform [20, 13]. Optical sensors are an attractive instrumental solution for APW assessment due to their truly non-contact nature that allows the measurement of the skin surface vibrations at the carotid artery site [40, 42, 44, 45]. A new laboratory research prototype, based on optical sensors, was developed and tested in several volunteers which proved to be a reliable method to the arterial distension waveform detection [44]; however, the accuracy of APW detection is dependent on the operator's skills and experience [46]. An automatic method to select the APW will standardize the analysis and objectify APW identification. The variability between operators is introduced by the subjective criteria in the selection the pulse segment to be analysed. This subjective influence is responsible for systematic differences amongst the operators (interoperator reproducibility) and deviations of a

✉ Tânia Pereira
taniapereira@lei.fis.uc.pt

¹ Physics Department, Instrumentation Center, University of Coimbra, Rua Larga, 3004-516 Coimbra, Portugal

² Physics Department, University of Coimbra, Coimbra, Portugal

particular operator's score on a particular patient (intraoperator reproducibility). The automatic method developed will be implemented in order to exclude the operator from the task of APW selection, and it will reduce the variability of measurement. The optical system is not compatible with clinical use if it requires a review of all data in order to select the APW by the operator. For this reason, an automatic method to reduce the time and the requirements for the operator was developed.

The main goal of this work was to develop an algorithm based on a classifier that could distinguish valid data (part containing APW) from non-relevant information acquired during a clinical examination by the optical system. A great potential in clinical applications for a system that assesses the APW by combining the non-contact optical probe and advanced machine-learning techniques for automatic signal detection is expected. With this combination, the system could be used in clinical routine, with simple use, without subjective influence on the signal selection and consequently less variability that allows the monitoring of the hemodynamic parameters and assessment cardiovascular condition.

The study of features and classification of pulse pressure waveform is a recent and interdisciplinary field of computer science and clinical interpretation [48]. This work combines an extensive study of features from pulse pressure waveform with the state-of-the-art classification techniques and feature selection method. A global methodology to determine the APW and an important advance in the study of APW features were described. Several other works tried to develop a methodology that allows identifying the normal APW [4, 19]. Previous studies mainly focused on wavelet analysis that demonstrated its capability to extract useful features for automated waveform classification [14, 39, 60]. Amplitude features also received close attention in other studies [37, 38], while less importance was given to frequency features [35]. However, other studies that analyse biosignals were focused on limited subset which comprises a reduced number of features [2, 38]. This work represents a wide study of APW features, covering the recognized important features and many others that never have been studied for this type of medical signal. Several methodologies were implemented, and the performance was evaluated in order to choose the most important. This work contributes to boosting the feature study of pulse pressure waveform.

Machine-learning techniques have been successfully employed in automatic classification of several physiological signals [10, 25, 31, 36, 16], including those related to cardiovascular diseases [1]. Two classifiers were implemented: k-nearest neighbours and support vector machine, in order to provide a comparison of performance results. This paper aims at establishing a systematic approach to

the automatic APW detection, with the focus placed on feature study and pattern classification. This study represents one important application for the optical system developed. The identification of abnormalities in the waveform opens the way for a range of applications. The pulse pressure waveform suffers changes according to modifications in the cardiovascular condition and several other pathologies that could be studied and identified with the extensive features determined in this study [12]. The APW is one of the most important vital signals used in clinical environment; however, the information contained in this signal is still poorly explored. The APW results in several phenomena of interaction of cardiac function, cardiovascular condition, respiration cycle, etc., and contains clinical information that could be explored in order to diagnose and for monitoring the disease states [5]. This work contributes to study and progress of this development line to assess the clinical condition using the pulse pressure waveform in a non-invasive way.

2 Methods

2.1 Measurement system

The signal with clinical interest is detected in peripheral arteries like the carotid, as the pressure wave travels from the heart across the arterial tree in a compliant way, forcing blood vessels to distend elastically according to the pressure wave profile and causing an external distension effect. This distension can be used to produce an optical signal which is correlated with the passing pressure wave. Several studies have shown that the distension and the pressure waveforms have an analogous wave contour and, therefore, can reciprocally be used for pulse wave analysis [9, 28, 47].

The functional structure of the optical system is represented in Fig. 1. The probe includes an illumination source, provided by monochromatic light emitting diodes (LED), and a planar photodiode (PD) acting as photodetector. The developed system is capable of emitting a constant intensity beam of light while, simultaneously, acquiring the reflected light. When the pressure wave travels underneath a deformable surface, like the skin, the distension of the artery, close to the skin surface, will alter the direction of the light reflection and thus the intensity of the detected signal.

The optical probe is enclosed in a plastic box, in an ergonomic configuration, ensuring a non-contact signal acquisition at the artery site, by keeping a 3-mm distance between the probe and the skin surface. The contact between the box probe and tissues was done in the surround area of the arterial vessel that allows the non-contact in the measurement site. Signals were digitized with a 16 bits National

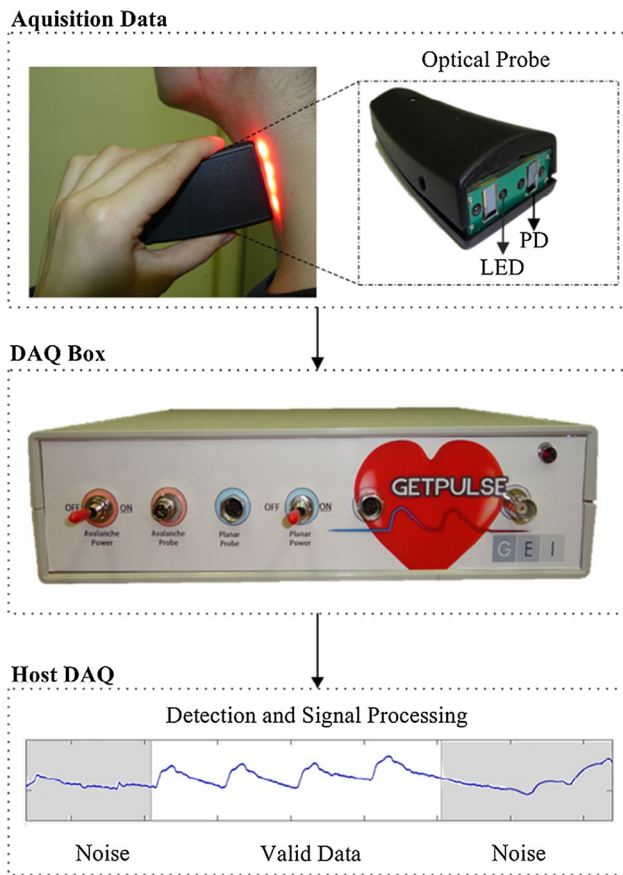


Fig. 1 Overview of the optical system for hemodynamic multi-parameter assessment

Instruments, USB6210[®] (DAQ represented in Fig. 1). The sampling frequency of the data acquisition is of 20 kHz; the cut-off frequency for the DAQ is 250 kHz. The signals were stored for offline analysis by MATLAB[®] [41, 43].

Previously, a comparison test was carried out between an ultrasound image system GE Vivid e[®] (30 Hz of sampling rate), as source of reference data, and the optical system with higher-resolution acquisition signals (20 kHz), adequate to feed feature extraction algorithms [40]. A large study was performed in 131 young subjects, and the results showed that the use of this new technique is a trustworthy method to determine pulse wave velocity (PWV) and for pulse waveform analysis [44, 45]. The PWV is defined as the speed at which the pulse pressure propagates along the arterial tree. It is known that PWV increases with age, blood pressure and arterial stiffness [22, 54].

2.2 Data collection

In the first stage, the pulse waveforms were collected using the optical system. The data set used in this work was

obtained during the acquisitions in 213 patients and contains parts with PWA and noise, segmented in 1752 pieces (1096 are labelled as APW, and 656 are noise). The study protocol was approved by the Ethical Committee of the Centro Hospitalar e Universitário de Coimbra (CHUC), Portugal, and all the subjects were volunteers and gave a written informed consent.

The 1752 acquisitions represent segments of acquisition with noise or arterial pulse pressure waveform detected by the optical system. The signals were previously labelled by physicians as noise or APW. The physician is an experienced operator of this type of devices and trained to identify and classify the data. This work studied the relation of each feature and patterns with valid information and those with no information (noise).

2.3 Feature creation

The signals were parameterized by means of 37 pulse features divided in the following subsets: amplitude features (reflection point; full width at half maximum), time domain statistics (mean; median; standard deviation; variance; interquartile range; skewness; kurtosis; root mean square; entropy), cross-correlation features (maximum of cross-correlation with template waveform), wavelet features (relative power at six levels of wavelets for two mother wavelets, Haar and Db4) and frequency domain statistics (first- to fourth-order moments in the frequency domain; median frequency; spectral entropy; total spectral power and peak amplitude in frequency band) [3, 15, 29, 37]. Most of these features are summarized in Table 1. The signal x_i represent the segments of data, for $i = 1, \dots, 1752$ in Table 1.

For the amplitude feature creation, a low-pass filter (with a cut-off frequency of 30 Hz) was used for noise filtering, improving the signal differentiation that is essential to detect the waveform inflection points [44]. The reflection point represents the amplitude of the inflection point in APW that corresponds to the arrival of the reflected component added to the incident wave. The full width at half maximum is determined by the difference between the two values at which amplitude is equal to half of its maximum value [17].

In the time domain statistics: mean and median quantitatively represent the magnitude of each signal, whereas standard deviation, variance and interquartile represent its variability. Skewness and kurtosis reflect the shape of the amplitude distribution [29, 37], while the root mean square measures the magnitude of a varying quantify [15]. The skewness quantifies how symmetrical the distribution is, and Kurtosis quantifies whether the shape of the data distribution matches the Gaussian distribution. These features allow exploring the differences in the distribution of sample and deviations from a normal distribution.

Table 1 Feature subsets and expressions

Group	Feature	Expression
Amplitude features	Reflection point (RP)	x' (RP) = 0 but is not the absolute maximum (x' is the first derivative)
Time domain statistics	Mean (M)	$M = \frac{1}{n} \sum_{i=1}^n x_i$ (n is the number of elements in the sample)
	Standard deviation (σ)	$\sigma = \left(\left(\frac{1}{n-1} \right) \sum_{i=1}^n (x_i - M)^2 \right)^{\frac{1}{2}}$ (where M is the mean of x)
	Variance (Var)	Var = $E[(x - M)^2]$ (E represents the expected value)
	Skewness (s)	$s = \frac{E(x-M)^3}{\sigma^3}$ (where σ is the standard deviation of x)
	Kurtosis (k)	$k = \frac{E(x-M)^4}{\sigma^4}$
	Root mean square (RMS)	$RMS = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}$
	Entropy (H)	$H = - \sum_{i=1}^n p(x_i) \log_{10} p(x_i)$ (where $p(x_i)$ is a probability of x)
Cross-correlation features	Cross-correlation (CC)	CC = $\max \sum_{m=-\infty}^{\infty} x[n+m]y^*[m]$ (y is the signal template; y^* denotes the complex conjugated of y)
Wavelet features	Relative power for each decomposition detail (P_j)	$P_j = \int D_j^2 dt$ (D_j is the reconstructed signal detail at the j th level)
Frequency domain statistics	Moments in the frequency ($M1f$)	$M1f = \sum_{f=1}^L f_i \text{PSD}(f_i)$ (where L is the length of FFT and f representing the frequency component)
	Median frequency (MF)	MF = $0.5 \sum \text{PSD}(f)$
	Spectral entropy (SE)	SE = $-\sum p_j \ln(p_j)$ (p_j is the normalized value of PSD at each frequency)
	Total spectral power (P_T)	$P_T = \sum \text{PSD}(f)$
	Peak amplitude (PA)	PA = $\max(\text{PSD}(f))$

A cross-correlation between the signal segments and an APW template was performed. Its maximum is a measurement of the coupling between these two time series. Differences in the structure of APW indicate an alteration of cardiovascular system, and four types of pulse pressure profile are described with pathology processes [14, 58]. The APW templates, used for the cross-correlation tests, belong to four waveform groups classified as A, B, C and D, which represent the most clinical-relevant APW types [14]. The arterial pulse pressure waveforms detected in the volunteers with specific features of each type of waveform described in the literature were used as a template. Four types of carotid pressure waveforms have been identified: type A, where systolic peak appears after the maximum wave; type B, which corresponds to cases of pronounced artery stiffness where the systolic peak appears after the inflection point (although the difference between them is very close to zero); type C waveforms, where the systolic peak appears before the inflection point (commonly seen in healthy individuals that have low arterial stiffness and high elasticity); and type D, whose waves are similar to type A waveforms, but the inflection point cannot be identified [14].

The wavelet features capture relative energies in different spectral bands. Two mother wavelets (Haar and Db4) were tested, and the relative power of the reconstructed signal detail (one to four) was determined [14,

26, 39]. The choice of the mother wavelet depends on the application, and, generally, a wavelet similar in shape to the analysed signal is considered appropriate. The mother wavelet function Daubechies 4 is most commonly used for the APW analysis [14, 39, 60]. Although the Haar function does not feature an excellent time–frequency localization, it is more suitable for analysis that requires peak detections [19, 23], which in this case could be an advantage for detection of typical inflection point in the APW.

First- to fourth-order frequency moments represent the amplitude of the power spectral density (PSD) at each single spectral component. The median frequency is defined as the spectral component which comprises 50 % of the signal power, with higher values corresponding to signals with significant spectral components at higher frequencies. The spectral entropy is a disorder quantifier related to the flatness of the spectrum. The total spectral power is computed as the total area under the PSD. The peak amplitude is the local maximum of the spectral content in the frequency range [9, 19].

Due to the differences in the characteristics (amplitude and variation) of the feature components, a normalization procedure is required. This task has a strong impact on KNN and SVM classification algorithms and consists in subtracting the mean over all training values and dividing by the corresponding standard deviation [9].

2.4 Feature selection

The choice of the most relevant features reduces the computational cost and avoids the inclusion of redundant information, which could compromise classification performance, on the training model. In an initial stage of a pattern recognition project, it is desirable to discard features with no significant contribution. There are methods that generate new features from the existing ones, which retain the most meaningful attributes, as is a process called feature extraction. An example is the principal component analysis (PCA) algorithm. On the contrary, feature selection algorithms preserve the most significant features contained in the original feature vector [30]. PCA must be applied for feature extraction with caution because there are some disadvantages related to its use. This method could discard components with negligible contribution to the overall variance, which may nevertheless provide a crucial contribution to pattern discrimination and, inadvertently, impair the classification accuracy [34].

Feature selection algorithms are divided into two types: filter methods and wrapper methods. The first one is based on indirect measurements, for example, with distances, reflecting segregation between classes. Wrapper methods, on the contrary, select a subset of features based on the classification accuracy, which means they are dependent on the used classifier. The support vector machine recursive feature elimination (SVM RFE) algorithm is a wrapper method based on a recursive process of feature elimination. Starting with all available features in the original data, it determines each feature's contribution in terms of the classification performance. The algorithm eliminates the feature with the least impact on the classification accuracy until a stopping criterion, indicating that a good solution has been found, is reached [55]. At the end of process, the algorithm returns a feature ranking, sorted by order of significance in terms of the classification accuracy [33]. In this work, a linear kernel function in SVM RFE was used. This type of kernel is the most simple of the kernel functions, ensuring a smaller computational cost and complexity in comparison with the other kernel types more commonly used. Only the regularization constant (RC) has to be tuned, and it has been demonstrated that very low values can improve the performance of SVM RFE [6]. The recursive feature elimination techniques have been successfully applied to physiological signals, with the aim to find feature subsets with high diagnostic relevance [3, 55]. The SVM RFE algorithm was adopted in the current project, and the SPIDER toolbox for MATLAB[®] was used to perform the feature selection [24].

2.5 Classifiers

In biomedical data classification systems are frequently applied classifier that is not suitable for the given data set

[32]. The most important literature in this area emphasizes the application of KNN and SVM to solve problems related to cardiac signals [57]; for that reason, those classifiers were compared in order to confirm the higher performance of SVM [48, 53]. SVM is defined as a classifier that can be linear or nonlinear. When applied to data, it can distinguish two different types of classes by finding a separating hyperplane with the maximal margin between two classes [32]. SVMs are defined by two general attributes: C , a hyperparameter that controls the trade-off between having large normalized margin and having less constraint violation; and kernel, a function that maps training data into a higher-dimensional space [30]. The kernel function is used to train the SVM, and, usually, the most common kernel types are the linear and the Gaussian radial basis function (RBF) [30]. In this study, the last one was adopted because the Gaussian kernel is considered the best option on the biological context [1, 18]. By using SVMs with this type of mapping function, a third parameter must be optimized: sigma, the width of the Gaussian function. Therefore, it is always necessary to define the best combination of the two hyper-parameters, C and sigma, that define the kernel RBF model.

Contrary to SVM, KNN does not make any assumptions about the underlying data pattern distributions. It is an algorithm used for object classification based on the closest training examples in the problem space [34]. The object is then assigned to the most common class amongst its k nearest neighbours. KNN is considered the simplest algorithm of all machine-learning techniques, where the function used is only approximated locally [48]. Generally, the performance of SVM classifier is better than KNN, but a comparing study between them was made in order to select the best model. The Statistical Pattern Recognition Toolbox for MATLAB[®] was used to design both the KNN and SVM classifiers [52].

2.6 Performance evaluation

A set of tests is usually performed in order to assess the performance of each classifier. Performance analysis is conducted under accuracy (A), specificity (SP), sensitivity (SE) and F-Measure (F-M). In general, the performance of a binary classifier as the ones analysed here can be evaluated taking into account the following quantities: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In the context of the analysed data set, a 'positive' denotes a part of the signal classified as having a waveform profile and 'negative' denotes noise. Consequently, a TP is a portion of the signal which was correctly classified as having a waveform profile, and a FP is a portion of the signal with noise which was wrongly classified as having a waveform profile.

A classifier is considered as having a good performance if it simultaneously has a high sensitivity and a high

specificity [34]. F-Measure reflects indirectly the sensitivity and the specificity of the method. The performance evaluation task allows the determination of method that ensures the highest accuracy and F-Measure values. To select a good classifier from a set of classifiers (model selection), it is necessary to adopt an accuracy estimation method. Its implementation is important to predict the classifier's future prediction accuracy and to calculate the above performance measures. Cross-validation is one of the most common methods [27]. In this method, the whole data set is randomly split into n different subsets (folds). Usually, fivefold or tenfold cross-validation displays lower variance. Cross-validation ensures nearly unbiased estimate of the prediction error rate and avoids overlapping test sets [59].

3 Results

This section contains the results obtained after a thorough test of the two classifiers, for several values of the parameters. These values are represented in Tables 2 and 3.

Different combinations of input parameters were tested for each classifier to find out the one which ensured the best classification performance in the cross-validation. The KNN classifier was tested for 1–50 nearest neighbours and fivefold to tenfold in cross-validation method [1, 50, 51]. For the SVM classifier tests were used the values of a constraint factor between 0.01 and 200, the sigma from 0.01 to 100 and cross-validation with fivefold to tenfold. In the SVM RFE, three values for small RC were tested: 1×10^{-7} , 5×10^{-7} and 1×10^{-6} , and combined with different feature numbers. The best performance of classifiers was found for the lower values of RC. Similar results were also described in another study that used very low values for the RC in order to improve the performance of SVM RFE [6, 21].

The values of Table 2 show that the overall accuracy is higher than 0.9, for both classifiers. For the KNN, the higher accuracy (0.933) was achieved for four nearest neighbours using fivefold cross-validation scheme. In the SVM classifier with the cross-validation method, the best value of accuracy was 0.952, and the best combination of parameters

Table 2 Overall KNN and SVM accuracy for different parameters in the SVM RFE (N represents number of features and the RC the regularization constant)

RC	1×10^{-7}					5×10^{-7}					1×10^{-6}				
	5	10	19	25	30	5	10	19	25	30	5	10	19	25	30
KNN	0.921	0.919	0.925	0.919	0.926	0.932	0.933	0.931	0.930	0.924	0.923	0.923	0.928	0.924	0.923
SVM	0.947	0.948	0.949	0.952	0.948	0.942	0.949	0.947	0.947	0.949	0.946	0.950	0.949	0.949	0.948

The higher values are marked in bold

Table 3 Validation results for KNN and SVM classifiers for different parameters of SVM RFE

RC	N features	KNN			SVM		
		SP	SE	F-M	SP	SE	F-M
1×10^{-7}	5	0.9018	0.9290	0.9347	0.9259	0.9623	0.9560
	10	0.9018	0.9337	0.9339	0.9203	0.9672	0.9569
	19	0.8836	0.9977	0.9387	0.9208	0.9672	0.9571
	25	0.8992	0.9337	0.9336	0.9312	0.9665	0.9600
	30	0.8619	0.9743	0.9395	0.9253	0.9640	0.9568
5×10^{-7}	5	0.8942	0.9624	0.9432	0.9115	0.9624	0.9515
	10	0.8921	0.9658	0.9446	0.9221	0.9668	0.9572
	19	0.9254	0.9506	0.9434	0.9179	0.9676	0.9563
	25	0.9054	0.9629	0.9426	0.9221	0.9644	0.9559
	30	0.8649	0.9583	0.9369	0.9214	0.9672	0.9573
1×10^{-6}	5	0.8500	0.9748	0.9372	0.9183	0.9647	0.9549
	10	0.8953	0.9355	0.9370	0.9297	0.9648	0.9585
	19	0.8898	0.9645	0.9409	0.9214	0.9672	0.9573
	25	0.8325	0.9747	0.9381	0.9253	0.9648	0.9572
	30	0.8943	0.9471	0.9371	0.9253	0.9640	0.9568

The higher values are marked in bold

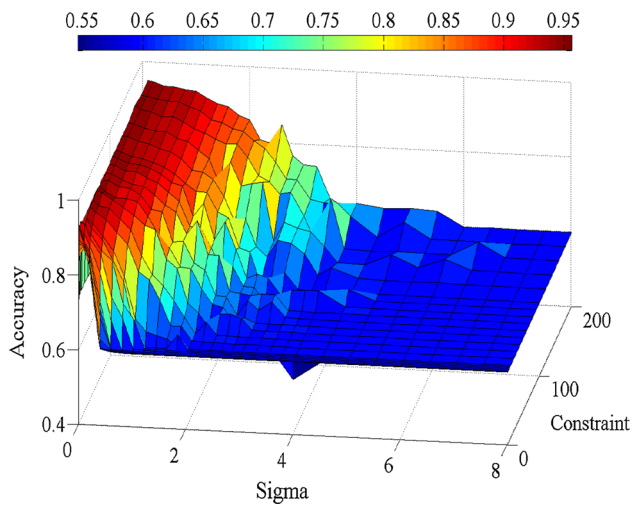


Fig. 2 Grid search showing the accuracy of SVM classifier for a wide range of σ and C for data set, and the colour code represents the accuracy result (colour figure online)

for the SVM RFE results in a small RC (1×10^{-7}) and 25 selected features out of 37. The number of final selected features, 25, was found for the best accuracy obtained by testing several combinations of parameters.

The evolution of the accuracy results corresponds to the combination of sigma and constraint values in the SVM classifier, and it is represented in Fig. 2. The highest value occurs using the values mentioned above, for a $C = 180$ and $\sigma = 0.1$ with fivefold cross-validation scheme. For sigma values higher than 2, there is a great decay of the accuracy results, and from $\sigma = 4$, it stabilizes at a low performance. Very small values of sigma keep a high score of performance for a wide range of constraint values.

From Table 2, it was possible to identify the best accuracy value, for each classifier, regarding the number of features tested. In Table 3, the best results are highlighted and it can be observed that all the algorithms achieved a good performance, since the overall results for SP, SE and F-M are extremely high, however, with lower results for KNN classifier when compared to SVM.

A good performance of a classifier can be considered when its values of SP and SE have some balance between them, i.e. the combination of higher values for SP and SE that could be expressed by the higher value of F-M. The best values of specificity and sensitivity were selected for each algorithm. In all cases, the SE value is higher than 0.9 which corresponds to a small false-negative rate. The high values of SP ensure that a small number of waveforms are classified as noise, and this affords an important characteristic of the classifier for this type of problem, ensuring that good waveform segments are not disposed.

The parameters that lead to the best accuracy were selected for a more detailed testing of the implemented

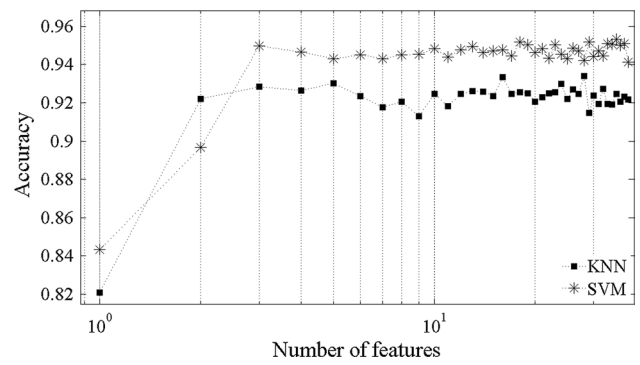


Fig. 3 Accuracy values corresponding to both SVM and KNN classifiers, for different numbers of features selected from the original set

algorithm. For these parameters, the graphic illustrated in Fig. 3 shows for the two different classifiers under study the evolution of accuracy with the number of features chosen from SVM RFE. The best values of accuracy were obtained for the SVM algorithms, which confirms that KNN has a weaker behaviour. Besides that, this illustration gives important information about the number of features that should be selected from SVM RFE.

It is clearly visible that accuracy has a strong increase until three features, but from that point the accuracy results not suffer substantial changes, which indicates that extracting a higher number of features will not affect the classifiers’ performance. However, few more features were introduced in the classifier in order to ensure that it always has a good performance (such as described in the plateau of curve); for this reason five features were selected. Through this analysis, it can be concluded that the first selected features, from SVM RFE, will have enough information to make a correct and safe distinction between APW and noisy signals. This point has important practical consequences for the classifier implementation, since being only five features necessary processing can be much faster.

The most important features identified by the SVM RFE procedure are: RMS; cross-correlation with signal template C and D; relative power for the first two levels of decomposition in the wavelet analysis using the mother wavelet function Haar. These five features are sufficient for the classifier to distinguish between APW and noise segments. Figure 4 shows the scatter plots between features from waveforms and noise acquisitions and almost distinguishable.

4 Discussion

A novel algorithm was developed for the detection of APW in the data acquired by the optical system.

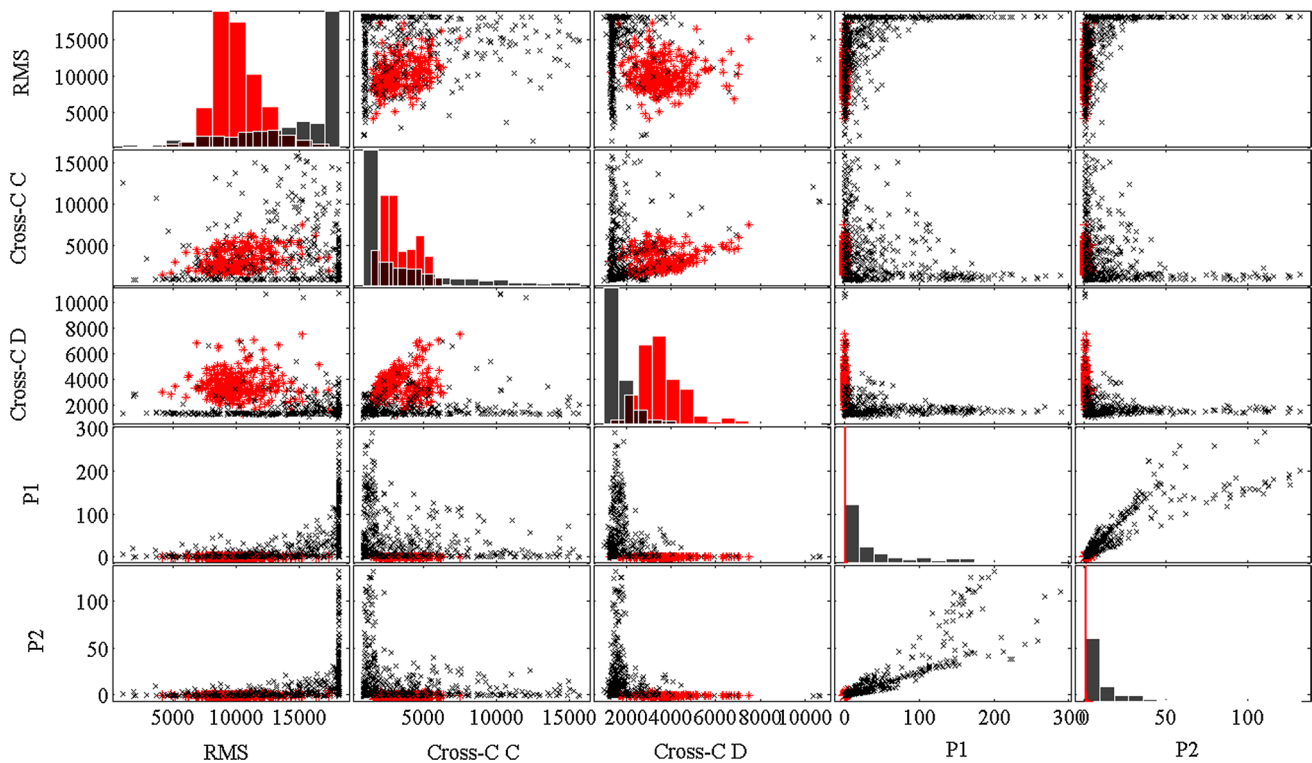


Fig. 4 Scatter plot matrix of the five variables distinguished by class; signals are marked in *red* and noise with *black*. The diagonal plots show the histogram of each feature. List of features: *RMS*; *Cross-C C* represents cross-correlation with signal template C; *Cross-C D* rep-

resents cross-correlation with signal template D; *P1* is the relative power for the first level and *P2* second level of decomposition with mother wavelet function Haar (colour figure online)

The extensive study of feature creation for APW has generated 37 features which allow the data characterization. This work represents a wide study in feature creation for APW. Several features used in this work have never been used with this kind of biomedical signal and therefore not have a physiological interpretation. This work opens, however, the opportunity to study these types of signals more broadly. Other features are well studied, such as the wavelet analysis. The arterial pulse pressure signals are the result of several biological processes (cardiac activity and arterial function) that correspond to different time and frequency scales in the analysed signal. The wavelet analysis represents one standard technique for PWA and provides the decomposition of a signal into different scales. Several hemodynamic parameters that allow the assessment of the arterial stiffness and cardiac function can be extracted from the APW analysis. Currently, the most accepted clinical parameters are determined by the mathematical relations in the time and amplitude features, such as the augmentation index (ratio between the amplitude of systolic peak and the amplitude of reflected waveform). This study was focused on the method that allows the selection of the APW that could be used for assessing relevant clinical parameters. The features used in this work were chosen in order

to develop a classifier that could be able to correctly select the waveforms. However, not all the features which were selected as the most significant ones have a clinical interpretation. Nevertheless, the results obtained in this study suggest that probably there are other features which clinical interpretation would be interesting and important to study for clinical assessment.

The SVM RFE was used to reduce the computational complexity and to improve the classifier's performance.

Two classifier architectures, KNN and SVM, were tested and compared to find which one guarantees the highest classification performance. The best method was an SVM classifier (constraint factor of 180 and sigma of 0.1) with an A (0.952), with a SP of 0.9312, and SE of 0.9665 and F-M of 0.9600. The implementation of the automatic method to select the APW in the optical system will decrease the influence of the operator and greatly improve data collection efficiency for the system based on the optical probe.

By varying the number of features, it was possible to determine the optimal feature number to get the best performance of the classifier. It was concluded that five features are enough for the classifier to distinguish between waveforms and noise with high accuracy. The main five features are: RMS; cross-correlation with signal templates

C and D; relative power for the first two levels of decomposition in the wavelet analysis using the mother wavelet function Haar. The higher relevance of the results from cross-correlation with template C and D reveals that data sample used as the test set was composed by the waveforms from healthy subjects (type C) and subjects with pathologic alterations in the cardiovascular condition originated by the arterial stiffness (type D). In fact, this set was acquired at the hospital and therefore was comprised of signals provided from both patients and healthy subjects. The relevant results obtained using wavelet features confirm which was concluded in previous studies [14, 26, 39]. The wavelet decomposition allows the study of the detailed waveform structure [14]. The RMS feature is correlated with the pulse area and waveform amplitude. This parameter reflects the cardiac and valve closure force [8].

The introduction of the correlated parameters' standard deviation and variance in the decision function could introduce redundant information, leading to impairment in classification performance. However, a feature selection method to raw data was applied, before the classification task, in order to select the feature set which retained the most relevant information. Therefore, if certain features are highly correlated between them, this algorithm selects only the most relevant ones. Taking into account that standard deviation and variance were similar features, the feature selection algorithm has discarded one of them.

The method developed for the automatic detection of APW showed higher accuracy values in comparison with other studies that also use machine-learning techniques to identify the APW, which only ensure accuracy results less than 92 % [4, 48, 56]. The feature set used in this work combined with the optimal classifier parameters represents an improvement in the methodologies applied so far for the automatic APW detection. However, other features might be created and tested, in order to improve the accuracy of APW estimation.

5 Conclusions

In this paper, a novel method to classify the data obtained by the optical system in pulse pressure waveform or noise segments was described. This study provides a review of emerging solutions to identify the APW and integrate the features from several domains to improve patient diagnosis by using collection and feature extraction in arterial waveforms by machine-learning algorithms. An implementation of one automatic method is extremely important to decrease the variability in the selection process and the time-consumed in the clinical examination. One experienced operator could be almost perfect, and for this reason, each sample of the database used in this work was

previously labelled by physicians as noise or APW. However, two of the most important requirements of a clinical system for daily use in the hospital are: easy to use and with effective application. The optical system is not compatible with clinical use if it requires that after the acquisition, the physicians need to review all data in order to select the APW. The automatic method developed will be implemented in order to exclude the operator from the task of APW selection and will reduce the variability of measurement and the time of clinical examination.

The algorithm developed was based on the SVM RFE to determine the most relevant feature and the SVM classifier to identify the pulse pressure signals and noise segments from the data obtained with non-contact optical system in the carotid site. An optimal feature subset was found by SVM RFE from one original extensive feature set composed by amplitude, cross-correlation, wavelet-based, time and frequency domain statistics-derived features. The set comprised of the most significant features was found, and it was composed by five features. The developed method showed high classification performances, achieving a maximal accuracy of 95.2 %.

This APW automatic detection method was developed to be applied at the end of each acquisition for processing the data detected by the optical system. Therefore, at the end of each clinical acquisition, each sample of the acquired signal will be classified as an APW or a noisy sample. However, a real-time solution could be implemented with the developed method and used in systems for real-time, beat-to-beat and continuous classification of pulse pressure waveform in dynamic clinical scenarios.

As future directions, it is the improvement of the classification algorithm to make it be able to discriminate not only between an APW and a noisy segment, but also between the different classes of APWs (from A to D). The automatic identification of APW type will be an important characteristic for the clinical implementation of the non-invasive optical system proposed. In addition, other types of classifiers could be tested for comparing the results using a different classification approach with the ones obtained with the SVM classifier, in order to verify whether other classifier would be more suitable to solve the proposed problem.

Acknowledgments Authors acknowledge the support from Fundação para a Ciência e a Tecnologia for funding (SFRH/BD/79334/2011). Project developed under the initiative of QREN, funding by UE/FEDER, through COMPETE.

References

1. Alty SR, Angarita-jaimés N, Millasseau SC, Chowienczyk PJ (2007) Predicting arterial stiffness from the digital volume pulse waveform. *Biomed Eng IEEE Trans* 54(12):2268–2275

2. Alvarez D, Member S, Hornero R (2010) Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. *Biomed Eng IEEE Trans* 57(12):2816–2824
3. Álvarez D, Hornero R, Marcos JV, Del Campo F (2012) Feature selection from nocturnal oximetry using genetic algorithms to assist in obstructive sleep apnoea diagnosis. *Med Eng Phys* 34(8):1049–1057
4. Angarita-jaimés N, Alty SR, Millasseau SC, Chowieniczky PJ (2006) Classification of aortic stiffness from eigendecomposition of the digital volume pulse waveform. In: 2006 IEEE international conference on acoustics, speech and signal processing, 2006. ICASSP 2006 proceedings, pp 1168–1171
5. Avolio AP, Butlin M, Walsh A (2010) Arterial blood pressure measurement and pulse wave analysis—their role in enhancing cardiovascular assessment. *Physiol Meas* 31(1):1–47
6. Bedo J, Sanderson C, Kowalczyk A (2006) An efficient alternative to svm based recursive feature elimination with applications bioinformatics. In: Sattar A, Kang B-H (eds) *AI 2006: advances in artificial intelligence*. Springer, Berlin, Heidelberg, pp 170–180
7. Blacher J, Asmar R, Djane S, London GM, Safar ME (1999) Aortic pulse wave velocity as a marker of cardiovascular risk in hypertensive patients. *Hypertension* 33(5):1111–1117
8. Bombardini T, Gemignani V, Bianchini E, Venneri L, Petersen C, Pasanisi E, Pratali L, Pianelli M, Faita F, Giannoni M, Arpesella G, Picano E (2008) Arterial pressure changes monitoring with a new precordial noninvasive sensor. *Cardiovasc Ultrasound* 6:41
9. Boutouyrie P, Briet M, Collin C, Vermeersch S, Pannier B (2009) Assessment of pulse wave velocity. *Artery Res* 3(1):3–8
10. Cilla M, Martínez J, Pena E, Martínez MA (2012) Machine learning techniques as a helpful tool toward determination of plaque vulnerability. *Biomed Eng IEEE Trans* 59(4):1155–1161
11. Crilly M, Coch C, Bruce M, Clark H, Williams D (2007) Indices of cardiovascular function derived from peripheral pulse wave analysis using radial applanation tonometry: a measurement repeatability study. *Vasc Med* 12(3):189–197
12. Dart AM, Kingwell BA (2001) Pulse pressure—a review of mechanisms and clinical relevance. *J Am Coll Cardiol* 37(4):975–984
13. De Melis M, Morbiducci U, Scalise L, Tomasini EP, Delbeke D, Baets R, Van Bortel LM, Segers P (2008) A preliminary study for the evaluation of large artery stiffness: a non contact approach. *Artery Res* 2(3):100–101
14. De Melis M, Morbiducci U, Rietzschel ER, De Buyzere M, Qasem A, Van Bortel L, Claessens T, Montecocchi FM, Avolio A, Segers P (2009) Blood pressure waveform analysis by means of wavelet transform. *Med Biol Eng Comput* 47(2):165–173
15. Diez PF, Mut V, Laciari E, Torres A, Avila E (2009) Application of the empirical mode decomposition to the extraction of features from EEG signals for mental task classification. In: Conference on proceedings of the IEEE engineering in medicine and biology society, vol 2009, pp 2579–2582
16. Dong S, Boashash B, Azemi G, Lingwood BE, Colditz PB (2014) Automated detection of perinatal hypoxia using time–frequency-based heart rate variability features. *Med Biol Eng Comput* 52(2):183–191
17. Elgendi M (2012) On the analysis of fingertip photoplethysmogram signals. *Curr Cardiol Rev* 8(1):14–25
18. Frontzek T, Lal TN, Eckmiller R, Bonn D, Germany FR (2001) Predicting the nonlinear dynamics of biological neurons using support vector machines with different kernels. In: International joint conference on neural networks, 2001. proceedings. IJCNN'01, vol 2, pp 1492–1497
19. He W, Li S, Xiao H, Yu C, Lin H (2012) An arterial elasticity index algorithm based on wavelet transform and curve fitting. *J Inf Comput Sci* 9(12):3379–3389
20. Horváth IG, Németh A, Lenkey Z, Alessandri N, Tufano F, Kis P, Gaszner B, Cziráki A (2010) Invasive validation of a new oscillometric device (Arteriograph) for measuring augmentation index, central blood pressure and aortic pulse wave velocity. *J Hypertens* 28(10):2068–2075
21. Huang TM, Kecman V (2005) Gene extraction for cancer diagnosis by support vector machines an improvement and comparison with nearest. *Artif Intell Med* 35(1–2):185–194
22. Huck CJ, Bronas UG, Williamson EB, Draheim CC, Duprez DA, Dengel DR (2007) Noninvasive measurements of arterial stiffness: repeatability and interrelationships with endothelial function and arterial morphology measures. *Vasc Health Risk Manag* 3(3):343–349
23. Janney JB, Sruthi SP (2012) Dicrotic notch detection and analysis of arterial pulse by using discrete wavelet. *OSIET J Commun Electron* 4:93
24. Jason Weston FS, Elisseeff A, BakIr G The spider. <http://www.kyb.tuebingen.mpg.de/bs/people/spider>
25. Kim K-A, Choi JY, Yoo TK, Kim SK, Chung KS, Kim DW (2013) Mortality prediction of rats in acute hemorrhagic shock using machine learning techniques. *Med Biol Eng Comput* 51(9):1059–1067
26. Kips J, Vanmolkot F, Mahieu D, Vermeersch S, Fabry I, de Hoon J, Van Bortel L, Segers P (2010) The use of diameter distension waveforms as an alternative for tonometric pressure to assess carotid blood pressure. *Physiol Meas* 31(4):543–553
27. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *IJCAI'95 proceedings of the 14th international joint conference on artificial intelligence*. pp 1137–1143
28. Laurent S, Cockcroft J, Van Bortel L, Boutouyrie P, Giannattasio C, Hayoz D, Pannier B, Vlachopoulos C, Wilkinson I, Struijker-Boudier H (2006) Expert consensus document on arterial stiffness: methodological issues and clinical applications. *Eur Heart J* 27(21):2588–2605
29. Lee J, Mark RG (2010) An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. *Biomed Eng Online* 9(1):62
30. Liu Y, Zheng YF (2006) FS_SFS: A novel feature selection method for support vector machines. *Pattern Recognit* 39(7):1333–1345
31. Liu NT, Holcomb JB, Wade CE, Batchinsky AI, Cancio LC, Darrah MI, Salinas J (2014) Development and validation of the machine learning algorithm and hybrid system to predict the need for life-saving interventions in trauma patients. *Med Biol Eng Comput* 52(2):193–203
32. Lotte F, Congedo M, Lécuyer A, Lamarche F, Arnaldi B (2007) A review of classification algorithms for EEG-based brain-computer interfaces. *J Neural Eng* 4(2):R1–R13
33. Maldonado S, Weber R (2009) A wrapper method for feature selection using support vector machines. *Inf Sci* 179(13):2208–2217
34. Marques JP (2001) *Pattern recognition: concepts, methods, and applications*, 1st edn. Springer, Berlin, Heidelberg
35. Mason L (2002) *Signal processing methods for non-invasive respiration monitoring*. University of Oxford, Oxford
36. Monkaresi H, Calvo RA, Yan H (2014) A machine learning approach to improve contactless heart rate monitoring using a webcam. *Biomed Heal Inform IEEE J* 18(4):2168–2194
37. Nayak GS (2012) Classification of ECG signals using ANN with resilient back propagation algorithm. *Int J Comput Appl* 54(6):20–23
38. Nayak GS, Davide O (2010) Classification of bio optical signals using k-means clustering for detection of skin pathology. *Int J Comput Appl* 1(2):92–96

39. Pachauri A, Bhuyan M (2012) Wavelet transform based arterial blood pressure waveform delineator. *Int J Biol Biomed Eng* 6(1):16–25
40. Pereira T, Oliveira T, Cabeleira M, Matos P, Pereira HC, Almeida V, Borges E, Santos H, Pereira T, Cardoso J, Correia C (2011) Signal analysis in a new optical pulse waveform profiler for cardiovascular applications. In: SIPA 2011—proceedings of the IASTED international conference on signal and image processing and applications, no. Sipa. pp 19–25
41. Pereira T, Cabeleira M, Matos P, Borges E, Cardoso J, Correia C (2011) Optical methods for local pulse wave velocity assessment. In: BIOSIGNALS 2011—4th international conference on bio-inspired systems and signal processing. Rome, Italy, pp 74–81
42. Pereira T, Cabeleira M, Matos P, Borges E, Almeida V, Pereira HC, Cardoso J, Correia C (2012) Non-contact pulse wave velocity assessment using optical methods. In: Fred A, Filipe J, Gamboa H (eds) *Biomedical engineering systems and technologies*, vol 273. Springer, Berlin, Heidelberg, pp 246–257. doi:10.1007/978-3-642-29752-6_18
43. Pereira T, Oliveira T, Cabeleira M, Pereira H, Almeida V, Cardoso J, Correia C (2012) Comparison of low-cost and non-invasive optical sensors for cardiovascular monitoring. *IEEE Sens J* 13(5):1434–1441. doi:10.1109/JSEN.2012.2236549
44. Pereira T, Santos I, Oliveira T, Vaz P, Correia T, Pereira T, Santos H, Pereira H, Almeida V, Cardoso J, Correia C (2013) Characterization of optical system for hemodynamic multi-parameter assessment. *Cardiovasc Eng Technol* 4(1):87–97
45. Pereira T, Santos I, Oliveira T, Vaz P, Santos H, Pereira H, Almeida V, Cardoso J (2013) Local PWV and other hemodynamic parameters assessment: validation of a new optical technique in a healthy population. In: BIOSIGNALS 2013—6th international conference on bio-inspired systems and signal processing, vol 1. Barcelona, Spain, pp 61–69
46. Pereira T, Santos I, Santos H, Almeida V, Pereira H, Correia C, Cardoso J (2014) Reproducibility of pulse wave analysis and pulse wave velocity in healthy subjects. In: BIOSIGNALS 2014—7th international conference on bio-inspired systems and signal processing. Angers, France, pp 221–228
47. Pereira T, Santos I, Oliveira T, Vaz P, Pereira T, Santos H, Pereira H, Correia C, Cardoso J (2014) Pulse pressure waveform estimation using distension profiling with contactless optical probe. *Med Eng Phys* 36(11):1515–1520
48. Raikwal JS, Saxena K (2012) Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. *Int J Comput Appl* 50(14):35–39
49. Rajzer MW, Wojciechowska W, Klocek M, Palka I, Brzozowska-Kiszka M, Kawecka-Jaszcz K (2008) Comparison of aortic pulse wave velocity measured by three techniques: Complior, SphygmoCor and Arteriograph. *J Hypertens* 26(10):2001–2007
50. Scalzo F, Xu P, Asgari S, Bergsneider M, Hu X (2009) Regression analysis for peak designation in pulsatile pressure signals. *Med Biol Eng Comput* 47(9):967–977
51. Scalzo F, Asgari S, Kim S, Bergsneider M, Hu X (2010) Robust peak recognition in intracranial pressure signals. *Biomed Eng Online* 9(1):61
52. Schlesinger MI, Hlavac V Statistical pattern recognition toolbox. <http://cmp.felk.cvut.cz/cmp/software/stprtool/>
53. Thakker B, Lal Vyas A (2011) Support vector machine for abnormal pulse classification. *Int J Comput Appl* 22(7):13–19
54. Vermeersch SJ, Dynamics B, Society L (2010) Determinants of pulse wave velocity in healthy people and in the presence of cardiovascular risk factors: ‘establishing normal and reference values’. *Eur Heart J* 31(19):2338–2350
55. Wang X, Tian J (2012) A gene selection method for cancer classification. *Comput Math Methods Med* 2012:586246
56. Wang H, Zhang P (2008) A model for automatic identification of human pulse signals. *J Zhejiang Univ Sci A* 9(10):1382–1389
57. Wang K, Wang L, Wang D, Xu L (2004) SVM classification for discriminating cardiovascular disease patients from non-cardiovascular disease controls using pulse waveform variability analysis. In: Webb GI, Yu X (eds) *AI 2004: advances in artificial intelligence*. Springer, Berlin Heidelberg, pp 109–119
58. Weber T, Auer J, O’Rourke MF, Kvas E, Lassnig E, Berent R, Eber B (2004) Arterial stiffness, wave reflections, and the risk of coronary artery disease. *Circulation* 109(2):184–189
59. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19(13):1636–1643
60. Zajarevich N, Bia D, Pessana F, Codnia J, Armentano R (2010) Arterial pressure and diameter waveforms analysis by means of wavelet transform: application to artery de-endothelization. In: *Conference on proceedings of the IEEE engineering in medicine and biology society*, vol 2010. pp 4550–4553

Tânia Pereira received the M.S. degree in biomedical engineering from the University of Coimbra, Coimbra, Portugal, in 2009 and the Ph.D. degree in biomedical engineering in 2014. Her current research interests include development of noninvasive instrumentation, data acquisition and signal processing for the assessment of cardiovascular parameters for diagnosis and monitoring of diseases.

Joana S. Paiva was born in Santa Maria da Feira, Portugal, in 1991. She received the BSc and MSc Degrees in Biomedical Engineering from the University of Coimbra in 2012 and 2015, respectively. Her current research interests include Pattern Recognition, Machine Learning, Brain Imaging, Multimodal Data Analysis, Biostatistics, Biosignals Processing and Analysis (e.g. EEG, ECG, pulse wave) and Biomedical Image Analysis. She won two awards for being one of the top 3 % students of the Integrated Master in Biomedical Engineering in 2013 and 2014 at the University of Coimbra.

Carlos Correia received the Ph.D. degree in physics and his studies to signal processing, electronics and biomedical instrumentation in 1985. He has been a Full Professor with the Physics Department, University of Coimbra, Coimbra, Portugal, since 2000. He has over 100 published works in the literature.

João Cardoso received the Ph.D. degree on the research area of digital pulse processing techniques for nuclear spectrometry in 2006. He is a Research Scientist with the Instrumentation Center, University of Coimbra, Coimbra, Portugal. His current research interests include optimization of instrumentation for physics experiments in the fields of astrophysics and medium energy particle physics as well as the development of innovative instruments and sensors for biomedical engineering.