International Journal of Neural Systems, Vol. 23, No. 5 (2013) 1350020 (18 pages)

DOI: 10.1142/S0129065713500202

World Scientific
www.worldscientific.com

ASSESSMENT OF FEATURE SELECTION AND CLASSIFICATION APPROACHES TO ENHANCE INFORMATION FROM OVERNIGHT OXIMETRY IN THE CONTEXT OF APNEA DIAGNOSIS

DANIEL ÁLVAREZ*, ROBERTO HORNERO and J. VÍCTOR MARCOS
Biomedical Engineering Group (GIB), University of Valladolid
Paseo Belén 15, 47011, Valladolid, Spain
**dalvgon@ribera.tel.uva.es*

NIELS WESSEL
Cardiovascular Physics, Humboldt-Universität zu Berlin
Robert Koch Platz 4, 10115, Berlin, Germany
wessel@physik.hu-berlin.de

THOMAS PENZEL and MARTIN GLOS
Center of Sleep Research, Charité Universitätsmedizin Berlin
Chariteplatz 1, 10117, Berlin, Germany
thomas.penzel@charite.de

FÉLIX DEL CAMPO
Department of Pneumology
Hospital Universitario Pío del Río Hortega
Dulzaina 2, 47013, Valladolid, Spain
fsas@telefonica.net

Accepted 3 May 2013
Published Online 2 July 2013

This study is aimed at assessing the usefulness of different feature selection and classification methodologies in the context of sleep apnea hypopnea syndrome (SAHS) detection. Feature extraction, selection and classification stages were applied to analyze blood oxygen saturation (SaO_2) recordings in order to simplify polysomnography (PSG), the gold standard diagnostic methodology for SAHS. Statistical, spectral and nonlinear measures were computed to compose the initial feature set. Principal component analysis (PCA), forward stepwise feature selection (FSFS) and genetic algorithms (GAs) were applied to select feature subsets. Fisher's linear discriminant (FLD), logistic regression (LR) and support vector machines (SVMs) were applied in the classification stage. Optimum classification algorithms from each combination of these feature selection and classification approaches were prospectively validated on datasets from two independent sleep units. FSFS + LR achieved the highest diagnostic performance using a small feature subset (4 features), reaching 83.2% accuracy in the validation set and 88.7% accuracy in the test set. Similarly, GAs + SVM also achieved high generalization capability using a small number of input features (7 features), with 84.2% accuracy on the validation set and 84.5% accuracy in the test set. Our results suggest that reduced subsets of complementary features (25% to 50% of total features) and classifiers with high generalization ability could provide high-performance screening tools in the context of SAHS.

Keywords: Sleep apnea hypopnea syndrome; oximetry; blood oxygen saturation; feature selection; principal component analysis; stepwise selection; genetic algorithms; Fisher's discriminant; logistic regression; support vector machines.

1. Introduction

The sleep apnea hypopnea syndrome (SAHS) is a respiratory disorder characterized by frequent breathing cessations (apneas) or partial collapses (hypopneas) during sleep. These respiratory events lead to deep oxygen desaturations, blood pressure and heart rate acute changes, increased sympathetic activity and cortical arousals.[1] Daytime hypersomnolence, neurocognitive dysfunction, metabolic deregulation and/or cardiovascular and cerebrovascular diseases could affect people having undiagnosed SAHS.[1,2] Common epidemiological data reflects a high SAHS prevalence in western countries: 1% to 5% of adult men and 2% of women. However, recent studies suggest that 20% of adults have at least mild SAHS and 7% of adults have moderate-to-severe SAHS.[3] Unlike its high prevalence and negative influence in the quality of life, it is estimated that 90% of cases in men and 98% of cases in women may be undiagnosed for many years.[2]

The gold standard method for SAHS diagnosis is in-hospital, technician-attended overnight polysomnography (PSG).[4] However, this methodology is labor-intensive, expensive and time-consuming,[4] which has led to large waiting lists, delaying diagnosis and treatment.[5] Thus, there is a great demand on new techniques aimed at simplifying the standard procedure and/or reducing the number of PSGs needed.[6] The main alternatives to PSG focus on developing automated analysis using a reduced set of cardiorespiratory-derived signals. Blood oxygen saturation (SaO_2) from overnight oximetry provides relevant information to detect apneas, it can be easily recorded ambulatory and it is less expensive and highly reliable.[6] However, there is still a great demand on new studies to improve the usefulness of SaO_2 in SAHS diagnosis.[7]

Several studies applied multivariate analysis to assist in SAHS detection.[8–11] Multivariate adaptive regression splines[8] and stepwise linear regression[9] have been used to classify subjects from conventional oximetric indexes. Discriminant analysis, logistic regression and neural networks have also been applied in the context of SAHS.[10–12] However, few studies applied feature selection before classification, which could improve diagnostic performance.

In the present study, feature extraction, selection and classification procedures were carried out to analyze SaO_2 recordings. Signal processing techniques were applied to compose an initial feature set: statistical, spectral and nonlinear measures were computed to obtain as much information as possible from oximetry. At this point, we hypothesized that an exhaustive analysis of the search space by means of variable selection could provide further knowledge on SaO_2 dynamics. Dimensionality reduction and feature selection techniques could be very useful to derive a smaller but optimal subset for classification purposes. There are many potential benefits of variable selection after feature extraction[13,14]: simplifying data representation, reducing measurement, storage and computational requirements, avoiding redundant and noisy information, selecting complementary features and defying the curse of dimensionality to improve classification accuracy. Feature subset selection methodologies are essentially divided into wrapper, filter and embedded methods.[14,15] Wrapper methods use a classifier of interest to score subsets of variables according to their predictive power, whereas filter methods select subsets of variables as a pre-processing stage independent of the predictor. Finally, embedded methods integrate variable selection into the learning machine training process. Additionally, feature construction and dimensionality reduction techniques are a different and useful approach when the number of variables is not too large and time and computational cost is not a concern.[14,16] Filter, wrapper and embedded techniques select features in the original space, which makes new subsets easy to interpret. On the other hand, feature construction approaches select variables in a transformed space, providing a more efficient representation of patterns. However, new features could not have clear physical meaning.[17] In the present study, three different approaches were assessed for feature selection: conventional principal component analysis (PCA),[18] forward stepwise feature selection (FSFS)[19] and genetic algorithms (GAs).[20] Additionally, three classifiers were used to investigate classification performance: Fisher's linear discriminant (FLD),[13] logistic regression (LR)[18] and support vector machines (SVMs).[21] Previous studies already applied these feature selection algorithms in different contexts, such as image processing,[22] signal monitoring,[23,24] structural monitoring[25,26] or model optimization.[27–29] Similarly, FLD and LR are

conventional classifiers extensively assessed in many fields[11,13,30,31] and SVMs are optimal state-of-the-art classifiers widely applied in different contexts, such as fMRI data analysis,[31] document classification,[32] biomedical signal processing[33,34] or motor pump faults detection.[35]

The goal of this study is to assess the usefulness of these algorithms for feature selection and classification in the context of SAHS diagnosis. We hypothesized that a prospective evaluation of different feature subsets from oximetry could provide further knowledge on SaO_2 dynamics. Thus, we wanted to test if the proposed classification schemes will be suitable for applying at another sleep laboratory. To achieve this goal, oximetric recordings from two independent sleep units were analyzed.

2. DataSet

Subjects under study were recruited from two independent sleep units: the "Río Hortega Hospital" (RHH) from Valladolid (Spain) and the "Philipps University Hospital" (PUH) from Marburg (Germany). First, a population set composed of 249 consecutive subjects (191 males and 58 females) was studied, with a mean ± standard deviation (SD) age of 52.2 ± 13.5 years and an average body mass index (BMI) of $29.9 \pm 4.9\,kg/m^2$. All subjects were derived to the sleep unit of the RHH due to a suspicion of suffering from SAHS. This population set was divided into training set and validation set. Table 1 shows the demographic and clinical characteristics of the population groups. The training set was used to compose optimum feature subsets from oximetric features and build the classifiers, whereas the validation set was subsequently used to assess their performance. In order to test whether proposed classification schemes will fit recordings from another sleep laboratory, optimum classifiers were further assessed on an independent test set. The Marburg subset (71 recordings) of the SIESTA database from the PUH was used. In this dataset, healthy subjects with no sleep disturbances composed the control group, whereas patients with a positive diagnosis of SAHS from PSG composed the SAHS-positive group. Table 2 shows the demographic and clinical features of this population.

The standard apnea–hypopnea index (AHI) from PSG was used to diagnose SAHS. Apnea was defined as a drop in the airflow signal greater than or equal to 90% from baseline lasting at least 10 s, whereas hypopnea was defined as a drop greater than or equal to 50% during at least 10 s accompanied by a desaturation greater than or equal to 3% and/or an arousal. Subjects with an AHI $\geq$ 10 events per h (e/h) were diagnosed as suffering from SAHS. Regarding the population under study from the RHH, a positive diagnosis of SAHS was confirmed in 165 patients.

Table 1. Demographic and clinical features of the population from the RHH sleep unit.

Features	All	SAHS-negative	SAHS-positive
Recordings (n)	249	84	165
Age (years)	52.2 ± 13.5	47.2 ± 11.5	54.7 ± 13.7
Males (n)	191	52	139
BMI (kg/m^2)	29.9 ± 4.9	28.0 ± 4.5	31.3 ± 4.7
Time (h)	7.2 ± 0.6	7.2 ± 0.4	7.2 ± 0.6
AHI (e/h)		3.9 ± 2.4	37.1 ± 25.8
Features	**Training set**	**SAHS-negative**	**SAHS-positive**
Recordings (n)	148	48	100
Age (years)	52.9 ± 14.1	48.3 ± 11.8	55.2 ± 14.6
Males (n)	116	32	84
BMI (kg/m^2)	29.8 ± 5.6	27.3 ± 6.3	30.8 ± 5.0
Time (h)	7.2 ± 0.4	7.2 ± 0.4	7.2 ± 0.4
AHI (e/h)		4.1 ± 2.4	40.9 ± 27.6
Features	**Validation set**	**SAHS-negative**	**SAHS-positive**
Recordings (n)	101	36	65
Age (years)	51.1 ± 12.7	45.8 ± 11.2	54.1 ± 12.5
Males (n)	75	20	55
BMI (kg/m^2)	29.0 ± 1.6	27.9 ± 0.8	30.8 ± 0.4
Time (h)	7.3 ± 0.7	7.2 ± 0.3	7.3 ± 0.9
AHI (e/h)		3.5 ± 2.3	31.4 ± 21.8

Table 2. Demographic and clinical features of the population from the PUH sleep unit.

Features	Test set	Normal subjects	SAHS-positive
Recordings (n)	71	50	21
Age (years)	40.37 ± 12.36	36.72 ± 11.59	49.05 ± 9.66
Males (n)	46	25	21
BMI (kg/m^2)	25.82 ± 5.86	22.93 ± 3.37	32.67 ± 4.68
Time (h)	7.7 ± 0.8	7.7 ± 0.7	7.9 ± 0.9
AHI (e/h)		0.60 ± 1.94	55.27 ± 33.44

The training set from the RHH was composed of 148 patients (48 SAHS-negative and 100 SAHS-positive), whereas the validation set was composed of 101 patients (36 SAHS-negative and 65 SAHS-positive). Every subject contributed one PSG study each (7.2 ± 0.6 h of recording, mean $\pm$ SD). On the other hand, nocturnal PSG was carried out during two consecutive nights at the PUH sleep unit. In the test set from the PUH, 50 PSG studies from 26 healthy subjects composed the control group (24 subjects contributed two recordings each and two subjects contributed one recording each), whereas 21 PSG studies from 11 SAHS-positive patients composed the SAHS-positive group (10 patients contributed two recordings each and 1 patient contributed with a single recording).

All SaO_2 recordings from PSG were saved to separate files and processed offline to compose the initial oximetric feature set. SaO_2 was recorded at a sampling rate of 1 Hz. SaO_2 signals presented zero samples at the beginning of the acquisition process and drops to zero due to patient movements along the recording time. An automatic signal pre-processing stage was carried out to remove these artifacts.

3. Methodology

Our methodology was divided into three stages: feature extraction, feature selection and classification. A total of 16 features composed the initial feature set from oximetry, which was the input to the subsequent feature selection stage. Three feature selection algorithms were evaluated: PCA, FSFS and GAs. Three classifiers were applied to assess classification performance in the third stage: FLD, LR and SVMs. Therefore, nine different classification schemes were proposed: PCA + FLD, PCA + LR, PCA + SVM, FSFS + FLD, FSFS + LR, FSFS + SVM, GAs + FLD, GAs + LR and GAs + SVM. Training and a double testing process were carried out. The training set was used to perform feature selection and compose classifiers, where a number of optimum feature subsets were automatically selected. Every optimum classifier from each proposed classification schema was subsequently assessed on two test sets: a validation group from the same sleep unit as the training set and a test set from an independent sleep unit. Figure 1 shows a block diagram to illustrate this methodology.

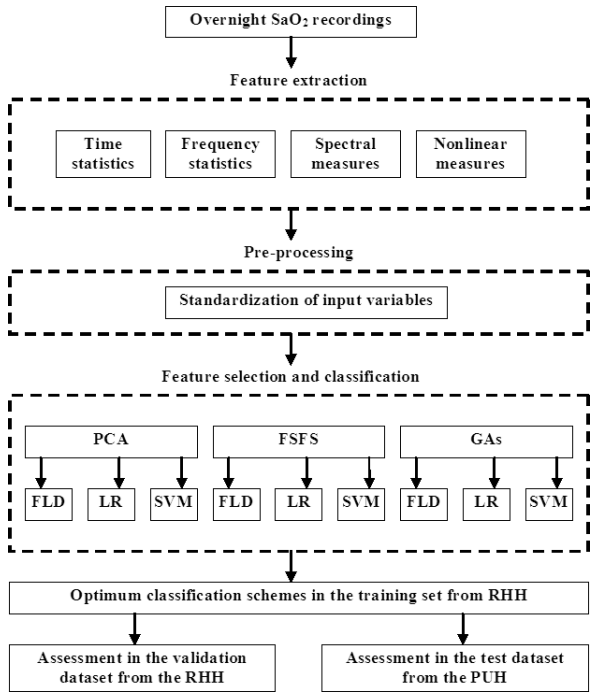

Fig. 1. System block diagram of the proposed methodology for feature extraction, selection and classification.

3.1. *Feature extraction stage*

Oximetric recordings were parameterized by means of 16 features from 4 feature subsets: time domain statistics, frequency domain statistics, conventional spectral measures and nonlinear features. All features were computed for each whole overnight recording.

3.1.1. *Time domain statistics*

The amplitude (%) of each SaO_2 signal was used to compute the normalized histogram. First to fourth-order statistical moments were computed[36]:

(i) Arithmetic mean ($M1t$), which is a measure of the central tendency of the data distribution:

$$M1t \equiv E[x] = \mu = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{1}$$

(ii) Variance ($M2t$), which quantifies the amount of dispersion in data, assigning higher values to higher variation:

$$M2t \equiv E[(x-\mu)^2] = \sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu)^2. \tag{2}$$

(iii) Skewness ($M3t$), which is a measure of symmetry in the data distribution. Large negative values suggest skewness (asymmetry) to the left while relatively large positive values suggest skewness to the right:

$$M3t = \frac{1}{\sigma^3} E[(x-\mu)^3], \tag{3}$$

where σ is the SD.

(iv) Kurtosis ($M4t$), which quantifies the peakedness, i.e. the frequency of data in the middle of the distribution. Positive peakedness suggests large concentration of probability in the center around μ accompanied by relative long tails, while negative values indicate relatively short tails:

$$M4t = \frac{1}{\sigma^4} E[(x-\mu)^4]. \tag{4}$$

3.1.2. *Frequency domain statistics*

The power spectral density (PSD) of each oximetric recording was estimated by applying the Welch's method. A 512-sample Hanning window with 50% overlap and 1024-points discrete Fourier transform were used. The following statistics were computed:

(i) First to fourth-order moments ($M1f$–$M4f$) in the frequency domain.[36] The amplitude (W/Hz) of the PSD function at each single spectral component was used to obtain the normalized histogram.

(ii) Median frequency (MF), which is defined as the spectral component which comprises 50% of the total signal power[37]:

$$0.5 \sum_{f_j=0\,\mathrm{Hz}}^{0.5 f_S} \mathrm{PSD}(f_j) = \sum_{f_j=0\,\mathrm{Hz}}^{\mathrm{MF}} \mathrm{PSD}(f_j). \tag{5}$$

(iii) Spectral entropy (SE), which is a disorder quantifier related to the flatness of the spectrum[37]:

$$\mathrm{SE} = -\sum_j p_j \ln(p_j), \tag{6}$$

where p_j is the normalized value of the PSD at each frequency component:

$$p_j = \frac{\mathrm{PSD}(f_j)}{\sum_{f_j=0\,\mathrm{Hz}}^{0.5 f_s} \mathrm{PSD}(f_j)}. \tag{7}$$

3.1.3. *Conventional spectral features*

The frequency band from 0.014 to 0.033 Hz proposed by Zamarrón *et al.* was parameterized. A significant power increase linked with suffering from SAHS was found in this frequency band.[38] The following measures were computed:

(i) Total spectral power (P_T), which is computed as the total area under the PSD.

(ii) Peak amplitude (PA) in the apnea frequency band, which is the local maximum of the spectral content in the apnea frequency range 0.014–0.033 Hz.

(iii) Relative power (P_R), which is the ratio of the area enclosed under the PSD in the apnea frequency band to the total signal power.

3.1.4. *Nonlinear features*

Linear methods cannot capture all the information from biological signals due to their nonlinearities and nonstationary behavior.[39–42] Therefore, nonlinear measures of irregularity, variability and complexity were applied to obtain additional and complementary information from SaO_2 dynamics.[30,43,44]

(i) Sample entropy ($SampEn$), which is a nonlinear measure of irregularity in time series, with larger values corresponding to more irregular data[45]:

$$\mathrm{SampEn}(m, r, N) = -\ln\left[\frac{A^m(r)}{B^m(r)}\right], \tag{8}$$

where A^m and B^m are the average number of (m)-length and $(m+1)$-length segments $X_m(i)$ $(1 \le i \le N-m+1)$ with $d[X_m(i), X_m(j)] \le r (1 \le j \le N-m,\ j \ne i)$, respectively, and

$$d[X_m(i), X_m(j)] = \max_{k=0,\ldots,m-1} (|x(i+k) - x(j+k)|). \tag{9}$$

(ii) Central tendency measure (CTM), which is a nonlinear measure of variability from second-order difference plots, assigning larger values to lower variability[46,47]:

$$\mathrm{CTM} = \frac{1}{N-2} \sum_{i=1}^{N-2} \delta(d_i), \tag{10}$$

where

$$\delta(d_i) = \begin{cases} 1 & \text{if } [(x(i+2) - x(i+1))^2 \\ & \quad + (x(i+1) - x(i))^2]^{1/2} < \rho \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

(iii) Lempel–Ziv complexity (LZC), which is a nonlinear measure of complexity linked with the rate of new subsequences and their repetition along the original sequence.[48,49] The complexity counter $c(n)$ is increased every time a new subsequence is encountered:

$$LZC = \frac{c(n)}{b(n)}, \tag{12}$$

where $b(n)$ is a normalization parameter.[48]

3.2. *Pre-processing stage*

Units used to measure input variables or changes in scale of measurement can influence the performance of classifiers.[13,50] Therefore, standardizing each feature by subtracting its mean and dividing by its SD is a common practice in the context of pattern recognition.[50,51] A linear re-scaling of each individual variable was carried out to obtain a zero mean and unit variance distribution for each input feature:

$$x_k(i) = \frac{x_k^{\text{raw}}(i) - \bar{x}_k}{\sigma_{x_k}}, \quad k = 1, \ldots, p, \tag{13}$$

where $x_k(i)$ is the standardized value for sample i of feature k, $x_k^{\text{raw}}(i)$ is the original raw value for sample i of feature k, $\bar{x}_k$ is the mean value of feature k and σ_{x_k} is its SD.

3.3. *Feature selection stage*

3.3.1. *Principal component analysis*

PCA is probably the best-known orthogonal transform for variable construction, which has been widely used as reference methodology for dimensionality reduction in pattern recognition.[16,17] As a variable construction technique, PCA is aimed at finding an appropriate transform that maps the pattern vector $\mathbf{x}(i)$ from the original p-dimensional feature space to a new d-dimensional feature space, where $d \leq p$.[17] When the number of features in the original space is large, the high correlation between variables under study becomes a problem in multivariate analysis. In order to avoid this issue, all variables or principal components from PCA in the new d-dimensional space are uncorrelated and mutually orthogonal.[13,18] New variables from PCA are linear transformations of the original features in a d-dimensional space, providing pattern representation with minimum mean-squared error for a given dimension d.[17] In the transformed space, new patterns are the projection of the original observations onto the eigenvectors of the original covariance matrix.[13,17] Each eigenvector accounts for a portion of the total variation of original data and the variance linked with each eigenvector is represented by its associated eigenvalue.[13,18] The portion of the total variation accounted for by the eigenvalue λ_d is given by its explained variance (EV):

$$\text{EV} = \frac{\lambda_d}{\sum_{k=1}^{p} \lambda_k}. \tag{14}$$

Regarding dimensionality reduction, PCA is commonly applied as a filter method to select variables in the transformed space as a pre-processing stage independent of the classifier. PCA allows discarding the components with lower EV to deal with a transformed space with lower dimension without significant loss of information.[18] The optimum number of components to accomplish dimensionality reduction can be estimated using some cut-off proportion. In this study, new variables from PCA were ranked according to their EV and the *average criterion* or *eigenvalue-one-criterion* was used as threshold to filter principal components. According to this rule, the components whose variance ($\lambda_j, j = 1, \ldots, p$) exceeds the average variance $\bar{\lambda}$ were selected:

$$\lambda_j > \bar{\lambda} = \sum_{j=1}^{p} \lambda_j \Big/ p. \tag{15}$$

In the present study, we applied PCA to the original dataset of 16 features from oximetry. PCA + FLD, PCA + LR and PCA + SVM classification schemes were subsequently built using the principal components automatically selected.

3.3.2. *Forward stepwise feature selection*

Sequential forward selection and backward elimination algorithms allow exploring the original p-dimensional feature space looking for a small subset that could reasonably describe the original data and avoiding the need to compute all possible 2^p combinations, which becomes impracticable when p

is large.[19,52] Both forward selection and backward elimination techniques yield nested subsets of features, where variables are progressively added into larger and larger subsets or progressively removing the least promising ones starting from the complete set of variables, respectively.[14] Advantages of both methodologies of feature selection are computational efficiency and robustness against overfitting. On the other hand, their main limitation is that once a variable has been included or removed from the subset, there is not a feedback process to modify the inclusion or exclusion of previous variables, which could improve the information provided by the model.[17] Forward stepwise selection and backward stepwise elimination improve sequential approaches by considering both feature addition and feature deletion at each step.[53]

Forward and backward stepwise strategies are usually classified as wrapper feature selection methods.[14,15] However, they can also be used as an embedded method if the criterion to decide whether or not to include or exclude a feature is not based directly on the accuracy of a classifier but on another objective function.[17] In the present study, we used a forward stepwise classifier-building strategy to find the simplest feature subset that still significantly explains original data.[19] Bidirectional FSFS decides to add or to remove a variable from the current feature subset through an iterative process. FSFS selects the strongest variables in the dataset and removes variables that provide redundant information in terms of statistical significant differences: at each iteration, the stepwise method performs a test for backward elimination followed by a forward selection procedure.[19] Different tests of statistical significance are used to compare models differing in one degree of freedom (1 input variable) depending on the output of the classifier. FSFS + FLD, FSFS + LR and FSFS + SVM schemes were analyzed in this study. The likelihood ratio test is used when output values can be interpreted as probabilities, such as in LR.[19] The output of a SVM can also be mapped to pseudo-probabilities using a logistic function.[54] In stepwise linear problems, an F-test is used since the errors are assumed to be normally distributed.[19] Therefore, the Rao's R approximate F-test was used for FLD.

In FSFS, a new variable is selected if the p-value associated to the statistical test was lower than a significance level α_E, which usually varies between 0.05 and 0.25.[19] Similarly, a variable was removed if the p-value was higher than a significance level α_R, commonly between 0.20 and 0.90[19]:

$$p_{\text{feature}}^{(\text{step})} = \min(p_j^{(\text{step})}) < \alpha_E \rightarrow \text{add feature}, \quad (16)$$

$$p_{\text{feature}}^{(\text{step})} = \max(p_j^{(\text{step})}) > \alpha_R \rightarrow \text{remove feature}. \quad (17)$$

The FSFS algorithm stops when all variables from the original feature set are selected or when all variables in the model have p-values lower than α_R and the remaining variables have p-values greater than α_E. In the present study, we used the less restrictive $\alpha_E = 0.25$ and a moderate $\alpha_R = 0.40$ significance thresholds to let the algorithm significantly explore the original feature space.[19]

3.3.3. *Genetic algorithms*

GAs are usually used as optimization schema to efficiently inspect the search space of variables or parameters that govern a model.[28,29] They encode a potential solution as a chromosome-like data structure and apply recombination operators on these structures.[24] A population from a GA optimization procedure comprises a group of chromosomes or candidate solutions that are modified iteratively: A particular group of chromosomes (parents) are selected from an initial population to generate the offspring by means of predefined genetic operations (crossover and mutation). The offspring replaces chromosomes in the current population based on certain replacement strategies.[28] The optimization process is carried out in cycles called generations.

In this study, GAs were applied as a wrapper feature selection procedure to obtain the optimum input feature subset of a classifier in terms of classification performance. In this case, an individual or chromosome from the population is just a combination of a predetermined number of features from SaO_2 recordings.[24] While conventional approaches just evaluate and improve a single feature subset, a GA intensively analyzes the whole feature space by modifying and improving a group of subsets at the same time.

A feature subset in the GA search space is codified with a finite binary sequence, where the kth bit denotes the absence (0) or the presence (1) of the kth feature. Each sequence has p bits, where p is the dimension of the original space, i.e. the

number of features in the whole set.[20] The classification accuracy is used as the objective value, in order to assess each chromosome performance and to achieve parent selection. A fitness function is used to map each objective value to a proportional predefined fitness interval. In this study, a proportional fitness scaling function was used. Additionally, roulette and tournament schemes were used as parent selection strategies. One-point crossover was applied to produce offspring: a crossover point is randomly selected and the portions of both parents beyond this point are exchanged to form the offspring.[28] Uniform mutation was applied to introduce variations into the offspring. In the present study, probability of crossover (P_c) values between 0.5 and 0.9 and probability mutation rate (P_m) values between 0.01 and 0.09 were used.[20] The elite or percentage of the best individuals in the old population preserved after each generation were varied between 0% and 25%. A number of realizations were carried out varying the parent selection strategy, P_c, P_m and elite. Each implementation of the GA was run with an initial population size of 16 individuals during 100 generations.[24] For each realization, the feature subset with the highest accuracy at the last generation was saved. Finally, the optimum feature subset in terms of diagnostic performance was selected. In this study, GAs + FLD, GAs + LR and GAs + SVM classification schemes were assessed.

3.4. *Feature classification stage*

3.4.1. *Fisher's linear discriminant*

In a binary (two class) context, FLD performs a linear projection of p-dimensional input data to a one-dimensional space:

$$y = w^T x, \tag{18}$$

where w is the projection weight matrix whose components maximize the class separation in the transformed space.[13] The Fisher criterion can be written as follows:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} d, \tag{19}$$

where S_B is the *between-class* covariance matrix and S_W is the total *within-class* covariance matrix. Differentiating $J(w)$ with respect to w, the separation of classes in the projected space is maximized when[13]:

$$w \propto S_W^{-1}(m_2 - m_1), \tag{20}$$

where m_i is the mean vector of the class i. The projected data can be used to construct a discriminant by choosing a threshold y_0 so that we classify a new point as belonging to C_1 if $y(x) \geq y_0$ and classify it as belonging to C_2 otherwise.

3.4.2. *Logistic regression*

LR relates a categorical dependent variable Y with a set of input features X_i. For dichotomous problems, input patterns are classified into one of two mutually exclusive categories (SAHS-positive or SAHS-negative in the context of SAHS diagnosis) and the probability density for the response variable can be modeled by a Bernoulli distribution[18]:

$$f(y \,|\, p(d)) = [p(d)]^y [1 - p(d)]^{(1-y)}, \tag{21}$$

where

$$p(d) = p(\beta_0 + \sum_{i=1}^{p} \beta_i x_i), \tag{22}$$

models the linear relationship between input features X_i. The maximum likelihood criterion is used to optimize coefficients of the independent input features.[18] LR classifiers assign an input vector to the class with the maximum *a posteriori* probability value. The LR model is expressed as follows[18]:

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \sum_{i=1}^{p} \beta_i x_i. \tag{23}$$

3.4.3. *Support vector machines*

SVMs are binary classifiers that search for the optimum separating hyperplane between classes.[13] The hyperplane is built in a transformed high-dimensional space in order to maximize separation, resulting in the following mapping function:

$$y(x, w) = w^T z + w_0, \tag{24}$$

where $x \in \Re^p$ is the input pattern, $z = \varphi(x) \,|\, z \in \Re^d$, $d > p$ performs the transformation of input data to a high-dimensional space, y is the output of the classifier and w is the weight vector. w is obtained by minimizing the following functional[21]:

$$E_c(w, \xi) = \frac{1}{2}\|w\|^2 + C \sum_{n=1}^{N} \xi^n, \tag{25}$$

subject to the constrains

$$t^n(w^T z^n + w_0) \geq 1 - \xi^n \quad \text{and} \quad \xi^n \geq 0 \qquad n = 1, \ldots, N, \quad (26)$$

where N is the number of observations in the training set, t^n is the target or desired output (+1 for the positive class and −1 for the negative class), ξ^n measures a deviation of a data point x^n from the ideal condition of separability (nonseparable classes) in the transformed space and C is a regularization parameter that controls the trade-off between the maximum margin of separation between classes and minimizing the classification error.[55] This optimization problem is commonly reformulated in terms of Lagrange multipliers η^n, so that the weight vector is expressed as follows:

$$w = \sum_{n=1}^{N} \eta^n t^n \varphi(x^n). \quad (27)$$

Only the support vectors, those for which their Lagrange multipliers are nonzero, contribute to the definition of the decision boundary. The output of the SVM classifier is expressed in terms of these support vectors as follows[21]:

$$y = \sum_{n \in S} \eta^n t^n K(x^n, x) + w_0, \quad (28)$$

where S is a subset of the indices $\{1, \ldots, N\}$ corresponding to the support vectors and $K(\cdot, \cdot)$ represents the inner product kernel function in the transformed space. In the present study, a linear kernel is used. The linear combination of inputs is the simplest but most useful kernel for SVM classification in many contexts, such as fMRI data analysis[31] or document classification.[32] Leave-one-out cross-validation (loo-cv) was carried out in the training set to obtain the optimum value of the regularization parameter C for each SVM classifier. The following values were assessed: $10^{-4}, 10^{-3}, 10^{-2}, \ldots, 10^{3}, 10^{4}$. For each value of C, we computed the accuracy of the classifier applying loo-cv. The value of C that achieved the highest accuracy was selected and the classifier was re-trained using the whole training set.

3.5. *Statistical analysis*

Matlab R2012a (7.14.0.739) and IBM SPSS Statistics 20 were used to implement feature extraction methods and to develop the feature selection and classification stages. Sensitivity (proportion of SAHS-positive patients correctly classified), specificity (proportion of SAHS-negative subjects rightly classified) and accuracy (the total percentage of subjects correctly classified) were computed to quantify classification performance. For every classifier, a ROC analysis was carried out to obtain its optimum decision threshold in the training set. This threshold was applied on further assessments in the validation and test sets.

4. Results

4.1. *Training*

Feature extraction was carried out for each SaO_2 recording from the populations under study. Figure 2(a) shows the nocturnal SaO_2 profile of a common SAHS-negative subject and a common SAHS-positive patient from the training set. Figure 2(b) shows the normalized averaged histogram envelope of recordings in the time domain for the whole SAHS-negative (dashed black) and SAHS-positive (dotted gray) groups in the training set. We can observe that the histogram envelope corresponding to the SAHS-negative group showed higher mean, skewness (symmetry) and kurtosis (peakedness) and lower variance in the time domain than that corresponding to SAHS-positive patients. This agrees with the fact that recordings from subjects without sleep apnea tend to remain constant around 96%,[6] i.e. higher mean and peakedness, whereas SAHS patients show deep desaturations during the night, i.e. higher variability and lower symmetry due to the left tail of the histogram envelope as a result of lower saturation values. Figure 2(c) shows the normalized averaged PSD for the whole SAHS-negative (dashed black) and SAHS-positive (dotted gray) groups in the training set. In the frequency domain, spectral power of oximetric recordings from SAHS-negative subjects concentrates on very low frequencies, showing lower mean and variance and higher skewness and kurtosis than SAHS-positive patients due to the continuous component (baseline) in the time domain around 96%. We can observe from Fig. 2(c) that spectral power of recordings from SAHS-positive patients spreads in a wider frequency band due to the repetitive apnea events during the night, leading to higher *MF* and *SE*. As a result, P_T, *PA* and P_R from SAHS-positive patients were also higher

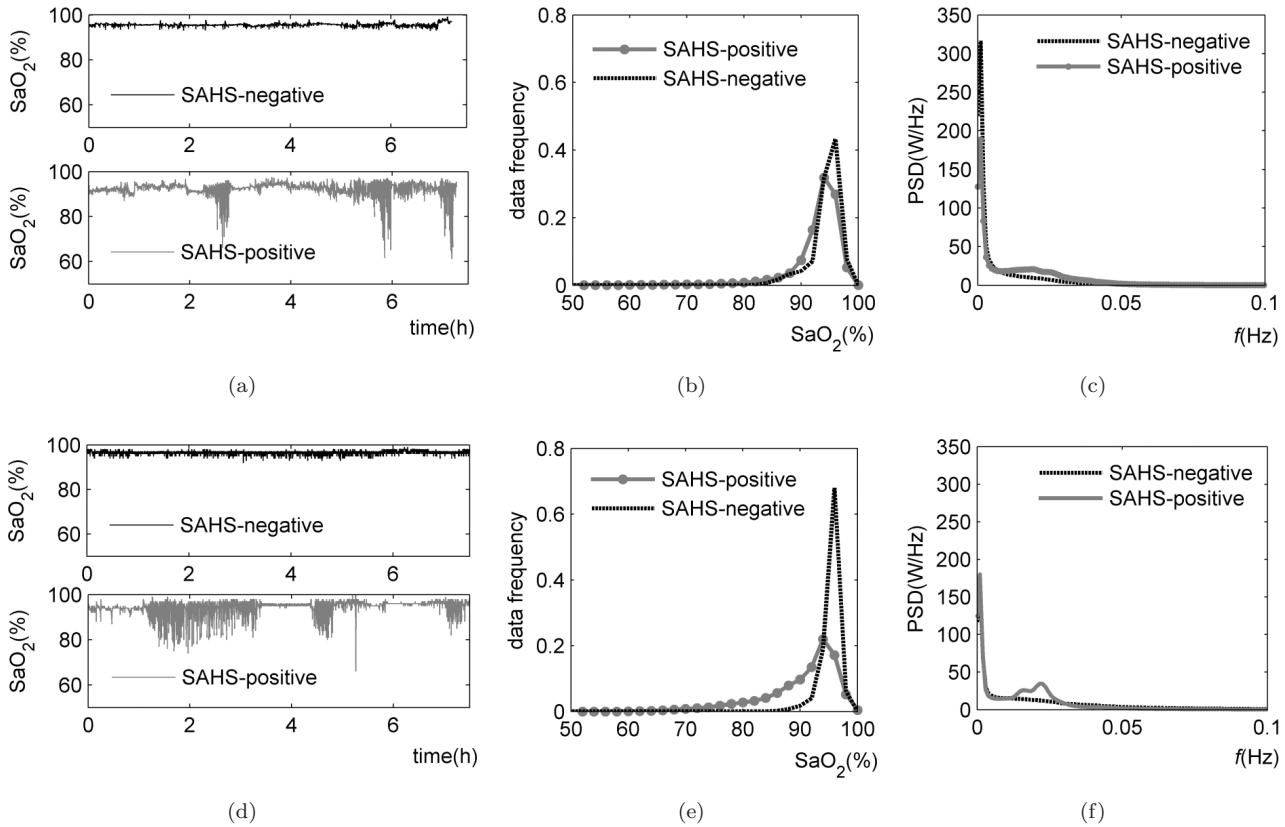

Fig. 2. Overnight SaO_2 profiles for a common SAHS-negative subject and a common SAHS-positive patient (a) from the RHH hospital database and (b) from the PUH database. Average histogram envelopes in the time domain for the whole SAHS-negative and SAHS-positive group (b) in the training set from the RHH and (e) in the test set from the PUH. Average PSD functions for the whole SAHS-negative and SAHS-positive group (c) in the training set from the RHH and (f) in the test set from the PUH.

than conventional spectral measures from the SAHS-negative group. Finally, common oximetric recordings in the time domain plotted in Fig. 2(a) show marked changes in the SaO_2 profile due to recurrent desaturations during the night in SAHS-positive patients, leading to higher irregularity (*SampEn*), variability (lower CTM) and complexity (LZC) than non-SAHS subjects. This trend was also present in the test set, although some differences between patient groups from both sleep units under study (RHH versus PUH) can be seen both in time and frequency domains. Figure 2(d) shows the SaO_2 profile of a normal subject and a SAHS patient from the PUH database, whereas Figs. 2(e) and 2(f) show the normalized averaged histograms and PSDs for the whole normal (dashed black) and SAHS-positive (dotted gray) groups in this test set. Differences between databases agree with heterogeneity of population commonly derived to sleep units. Additionally, the histogram envelope in the time domain of the normal group from the PUH shows a marked peak, higher than that corresponding to the RHH. This is due to the fact that the dataset from the PUH is composed of non-SAHS subjects with lower average AHI than SAHS-negative patients from the RHH. Similarly, the PSD of the SAHS-positive group from the PUH show higher power increase in the apnea frequency band than SAHS-positive patients from the RHH due to the fact that, on average, they have higher SAHS severity.

PCA, FSFS and GAs were applied for feature selection in the training set and a number of FLD, LR and SVM classifiers were composed. Table 3 shows principal components from PCA in the training set ranked in decreasing order of their EV. The three first consecutive principal components were selected according to the *average criterion*. Table 4 summarizes the performance of

Table 3. Explained variance for each principal component from PCA in the training set.

Principal components	*EV*
First principal component	42.60994
Component 2	26.69255
Component 3	11.82037
Component 4	6.00241
Component 5	3.64536
Component 6	3.24952
Component 7	1.69924
Component 8	1.48557
Component 9	1.23793
Component 10	0.88975
Component 11	.33337
Component 12	0.20245
Component 13	0.06906
Component 14	0.05557
Component 15	0.00690
Component 16	0.00001

feature selection and classification schemes under study. Regarding PCA dimensionality reduction, PCA + LR achieved the highest diagnostic accuracy in the training set (90.5%), while PCA + FLD and PCA + SVM achieved similar but lower performance than LR (83.8% and 84.5%, respectively). Similarly, FSFS + LR also achieved the highest accuracy (91.9%) after bidirectional feature selection in the training set. A reduced LR model composed of 4 features was built. FSFS + FLD (8 features) and FSFS + SVM (5 features) achieved slightly lower performance (90.5% and 87.8%, respectively). Exhaustive feature selection by means of evolutionary algorithms built more complex classifiers composed of a larger number of features, ranging from 7 to 15 variables. GAs + LR also obtained the highest diagnostic accuracy in the training set (96.6% using 14 and 15 features). GAs + FLD (7 and 9 features) and GAs + SVM (7 and 8 features) yielded to lower performances in the training set (93.2% and 86.5%, respectively).

4.2. *Validation and testing*

Each feature selection and classification schema was prospectively assessed. Optimum classifiers were evaluated on two independent test sets from different sleep units. Table 5 summarizes the performance assessment of the proposed methodology. The accuracy of optimum classifiers from PCA significantly decreased, with accuracies ranging from 71.3% to 81.2% in the validation set and 40.9% to 54.9% in the test set. Similarly, the FSFS + FLD classifier composed of 8 features achieved 78.2% accuracy in the validation set and 57.8% accuracy in the test set. On the other hand, optimum classifiers from FSFS + LR and FSFS + SVM schemes showed lower performance decrease. The LR model composed of 4 features achieved 83.2% accuracy in the validation set and 88.7% accuracy in the test set, whereas the SVM classifier with 5 input features achieved 82.2% accuracy in the validation set and 80.3% accuracy in the test set from the PUH sleep unit. Optimum classification schemes from GAs showed different

Table 4. Optimum feature subsets for each feature selection and classification methodology and their performance in the training set.

Algorithm	n	Features	Se	Sp	Ac
PCA + FLD	3	3 principal components	80.0	91.7	83.8
PCA + LR	3	3 principal components	92.0	87.5	90.5
PCA + SVM	3	3 principal components	81.0	91.7	84.5
FSFS + FLD	8	$M1t$, $M3t$, $M4t$, SE, P_R, SampEn, CTM, LZC	90.0	91.7	90.5
FSFS + LR	4	$M2t$, $M4t$, P_R, LZC	92.0	91.7	91.9
FSFS + SVM	5	$M4t$, PA, P_R, SampEn, LZC	87.0	89.6	87.8
GAs + FLD	7	$M1t$, $M3t$, $M4t$, $M1f$, SE, SampEn, LZC	94.0	91.7	93.2
	9	$M2t$, $M4t$, $M1f$, $M2f$, $M4f$, P_T, PA, P_R, LZC	94.0	91.7	93.2
GAs + LR	14	$M1t$, $M3t$, $M4t$, $M1f$, $M3f$, $M4f$, MF, SE, P_T, PA, P_R, SampEn, CTM, LZC	97.0	95.8	96.6
	15	$M1t$, $M2t$, $M3t$, $M4t$, $M1f$, $M2f$, $M3f$, $M4f$, MF, SE, P_T, PA, P_R, CTM, LZC	97.0	95.8	96.6
GAs + SVM	7	$M2t$, $M3t$, $M4t$, $M2f$, $M4f$, SE, CTM	84.0	91.7	86.5
	8	$M2t$, $M3t$, $M4t$, $M2f$, $M3f$, $M4f$, SE, CTM	84.0	91.7	86.5

Table 5. Diagnostic performance assessment of optimum feature subsets from each feature selection and classification methodologies in the validation set and in the test set from an independent sleep unit.

			Validation set (RHH)			Test set (PUH)		
Algorithm	n	Features	Se	Sp	Ac	Se	Sp	Ac
PCA + FLD	3	First 3 principal components	66.2	80.6	71.3	52.4	44.0	46.5
PCA + LR	3	First 3 principal components	92.3	61.1	81.2	100.0	36.0	54.9
PCA + SVM	3	First 3 principal components	67.7	80.6	72.3	28.6	46.0	40.9
FSFS + FLD	8	$M1t$, $M3t$, $M4t$, SE, P_R, SampEn, CTM, LZC	76.9	80.6	78.2	9.5	78.0	57.8
FSFS + LR	4	$M2t$, $M4t$, P_R, LZC	83.1	83.3	83.2	95.2	86.0	88.7
FSFS + SVM	5	$M4t$, PA, P_R, SampEn, LZC	83.1	80.6	82.2	76.2	82.0	80.3
GAs + FLD	7	$M1t$, $M3t$, $M4t$, $M1f$, SE, SampEn, LZC	80.0	83.3	81.2	95.2	46.0	60.6
	9	$M2t$, $M4t$, $M1f$, $M2f$, $M4f$, P_T, PA, P_R, LZC	10.8	91.7	39.6	0.0	94.0	66.2
GAs + LR	14	$M1t$, $M3t$, $M4t$, $M1f$, $M3f$, $M4f$, MF, SE, P_T, PA, P_R, SampEn, CTM, LZC	89.2	77.8	85.2	100.0	2.0	31.0
	15	$M1t$, $M2t$, $M3t$, $M4t$, $M1f$, $M2f$, $M3f$, $M4f$, MF, SE, P_T, PA, P_R, CTM, LZC	100.0	11.1	68.3	100.0	0.0	29.6
GAs + SVM	7	$M2t$, $M3t$, $M4t$, $M2f$, $M4f$, SE, CTM	84.6	83.3	84.2	95.2	80.0	84.5
	8	$M2t$, $M3t$, $M4t$, $M2f$, $M3f$, $M4f$, SE, CTM	84.6	83.3	84.2	95.2	76.0	81.7

performance depending on the classifier. GAs + LR achieved moderate to high accuracies in the validation set, ranging from 68.3% (15 features) to 85.2% (14 features), but extremely low performance in the test set, with accuracies ranging from 29.6% (15 features) to 31.0% (14 features). GAs + FLD achieved unbalanced accuracies in the validation set, ranging from 39.6% (9 features) to 81.2% (7 features), and moderate performance in the test set, with accuracies ranging from 66.2% (9 features) to 60.6% (7 features). On the other hand, GAs + SVM provided higher performance and more stable classifiers, leading to 84.2% accuracy (7 and 8 features) in the validation set, and accuracies ranging from 81.7% (8 features) to 84.5% (7 features) in the test set.

5. Discussion

This study assessed the usefulness of 9 feature selection and classification schemes to enhance information from SaO_2 oximetric recordings in the context of SAHS diagnosis. An initial feature set composed of 16 features was developed to characterize SaO_2 dynamics. A filter-based selection approach from variable construction (PCA), an embedded feature selection approach (FSFS) and a wrapper methodology for exhaustive analysis of the feature space (GAs) were applied. FLD, LR and SVM classifiers were involved on each feature selection methodology. Optimum classification schemes from the training set were subsequently assessed in datasets from different sleep units.

Our results showed that all algorithms from different feature selection and classification procedures reached high performance in the training set, with accuracies ranging from 83.8% to 96.6%. In contrast, optimum classification schemes showed different behavior when they were further tested. Regarding results from PCA, significantly lower or unbalanced sensitivity and specificity values were reached in the validation set from the RHH, leading to accuracies ranging from 71.3% to 81.2%. The diagnostic performance was even lower in the test set from the PUH, with a maximum accuracy of 54.9% using a LR classifier. PCA performs feature selection as a pre-processing stage regardless of the classification method. This is the reason why PCA achieved the lowest performances in the training set and subsequently failed in the validation and test sets independent of the classifier. Optimum classification schemes from GAs showed high dependence on the number of selected features. GAs + LR achieved the highest performances in the training set using high-dimensional feature subsets automatically selected. However, extremely unbalanced sensitivity and specificity values were obtained in further

assessments, especially in the test set from the PUH, with accuracies ranging from 29.6% (15 features) to 31.0% (14 features). On the other hand, GAs + SVM provided higher performance and more stable classifiers using half of the features: 84.5% (7 features) and 81.7% (8 features) in the test set. GAs are optimization algorithms aimed at extensively inspecting the search space in the training set to maximize the fitness function, usually the performance of a classifier. SVMs provide high generalization performance on pattern classification problems.[31,55] Indeed, the regularization parameter C controls the trade-off between the maximum margin of separation between classes and minimizing the classification error.[21,31] Our results suggest that, when low generalization capability predictors are used, GAs might build classifiers composed of a high number of features that overfit the training set and fail on subsequent assessments in different population groups. It is noteworthy that GAs + FLD selected feature subsets of similar size than those from GAs + SVM. However, optimum classifiers from GAs + FLD reached significantly lower accuracy in the test set. Performance decrease could be due to the fact that SVMs do not hypothesize any *a priori* statistical distribution of variables, whereas input features are assumed to have normal distributions and equal covariance matrices when using FLD.[31] Similarly, FSFS + FLD achieved unbalanced sensitivity and specificity values and low accuracy in the test set using eight features. On the contrary, FSFS + LR and FSFS + SVM provided high performance and balanced classifiers with reduced input feature subsets composed of four and five features, respectively. This agrees with the aim of forward stepwise selection: features are selected taking into account the amount of information added to the model, instead of maximizing classification accuracy on a specific dataset. Using efficient search strategies instead of "brute force" techniques did not decrease prediction performance. Indeed, our results support previous studies reporting that greedy search strategies, such as stepwise feature selection, are computationally advantageous and robust against overfitting.[14]

Regarding the number of features, the highest and more balanced performances in the validation and test sets were obtained using reduced feature subsets (25–50% of input features). From FSFS, FSFS + LR and FSFS + SVM schemes selected the smallest feature subsets: 4 ($M2t$, $M4t$, P_R, LZC) and 5 ($M4t$, PA, P_R, $SampEn$, LZC) features, respectively. Similarly, GAs + SVM provided 2 models with 7 ($M2t$, $M3t$, $M4t$, $M2f$, $M4f$, SE, CTM) and 8 ($M2t$, $M3t$, $M4t$, $M2f$, $M3f$, $M4f$, SE, CTM) features that yield to high accuracy both in the validation and test sets. Our results suggest that the larger the number of features, the larger overfitting is on the training set, leading to poor performance in subsequent assessments. Regarding PCA, only the first three principal components were selected using the *average criterion*. However, each principal component is a linear transformation of the original features, i.e. all 16 features contribute to every new variable in the transformed space. Thus, information from a large amount of features is used to achieve high performance in the training set, whereas accuracy significantly decreases in the validation and test sets.

In order to obtain high-performance classifiers is essential to build an initial feature set that concentrates as much nonredundant information as possible about the problem under study. Therefore, in the present research we built an original feature set from oximetry composed of metrics from complementary analyses: time versus frequency and linear versus nonlinear. After the feature selection stage, time, spectral and nonlinear features are included in the optimum feature subsets from FSFS + LR, FSFS + SVM and GAs + SVM, which achieved the highest accuracies in both test populations. Both subsets from the FSFS feature selection approach share 60–75% of features (three features): a linear statistic in the time domain ($M4t$), a linear measure in the frequency domain (P_R) and a nonlinear measure in the time domain (LZC). These features jointly account for the main characteristics of overnight SaO_2 profiles of non-SAHS subjects and the influence of apnea events in the recordings of SAHS-positive patients. $M4t$ measures the peakedness of the data distribution in the time domain, which is especially high in the case of SaO_2 recordings from non-SAHS subjects due to its near-constant behavior. On the other hand, there is a significant power increase in the frequency band between 0.014 and 0.033 Hz due to the quasi-periodic components of overnight respiratory events. P_R quantifies the effect of repetitive apneic episodes on SaO_2 recordings in the frequency

domain. Finally, desaturations of different severity modify the normal SaO_2 profile by adding new patterns or subsequences. LZC quantifies to what extent these desaturations increase the complexity of the SaO_2 signal in the time domain. Similarly, subsets from the GAs + SVM schema share 87.5% of features (7 out of 8). Comparing shared optimum features from both feature selection techniques (FSFS and GAs), we can observe that *M4t* is present in subsets from both approaches, P_R is replaced by *SE*, which is also influenced by the presence of additional frequency components in the power spectrum due to recurrent apneic events, and the nonlinear measure of complexity *LZC* is replaced by the nonlinear measure of variability *CTM*, which also quantifies time domain changes in the SaO_2 profile due to overnight desaturations. Therefore, our results suggest that a suitable feature selection stage applied to a suited and balanced initial feature set could detect complementary information and thus increase the diagnostic performance of oximetry in the context of SAHS diagnosis.

Previous researchers applied multivariate analysis in the context of SAHS. Using conventional oximetric indexes based on the number, duration and amplitude of the desaturations, 88.0% sensitivity and 70.0% specificity were reached from stepwise linear regression,[9] whereas 90% sensitivity and 70% specificity were obtained using multivariate adaptive regression splines.[8] Using spectral features from the high-frequency range, a sensitivity of 82% and a specificity of 84% were obtained with a LR classifier.[10] Higher performance (91.1% sensitivity and 82.6% specificity) was obtained by applying linear discriminant analysis to conventional spectral features in the apnea frequency band.[11] Neural networks have also been applied using clinical and anthropomorphic features (94.9% sensitivity and 64.7% specificity)[56] and oximetric features (89.4% sensitivity and 81.4% specificity) as input variables.[12] Different approaches of multivariate analysis using features from nonportable ECG have also been developed in the context of SAHS detection, reaching accuracies ranging from 74.4% to 100% using populations with no more than 80 subjects.[57–59] Other researchers suggested the use of wavelet features as inputs to a SVM classifier to assist in SAHS diagnosis from ECG.[60,61] A diagnostic accuracy of 92.86% was achieved on a small test set composed of 42 subjects.[60] The proposed methodology was also assessed on a slightly larger database composed of 70 recordings.[61] An accuracy of 100% was reached on a test set with 30 subjects. However, borderline subjects were excluded from the study.

Recent studies by our group applied dimensionality reduction and stepwise feature selection procedures before classification.[30,62,63] PCA was applied to a small set of three spectral and three nonlinear features.[62] First-to-fifth principal components were selected and 93.0% accuracy (97.0% sensitivity and 79.3% specificity) was reached on a test set from the same sleep unit. FSFS + LR was previously applied to a larger feature set from oximetry, reaching 89.7% accuracy (92.0% sensitivity and 85.4% specificity) using cross-validation.[30] Similarly, FSFS + LR was also applied to a wide feature set (42 features) from single channel airflow and respiratory rate variability.[63] Using cross-validation, 82.4% accuracy was reached by the LR model composed of features automatically selected from both signals. Finally, a preliminary study on the usefulness of GAs for feature selection in the context of SAHS diagnosis from oximetry has been recently carried out.[64] A LR model composed of six features achieved the highest accuracy (87.5%) in the test set from the same sleep unit. Nevertheless, these studies tested their approaches on populations from the same hospital. In the present research, we analyzed SaO_2 datasets from two different sleep units to assess our methodologies. To our knowledge, this is the first study where several complementary feature selection and classification algorithms are prospectively tested in the context of SAHS diagnosis from oximetry.

We should take into account some limitations regarding the general application of our methodology. Recurrent desaturations during sleep are not exclusive of SAHS. The presence of other disorders, such as asthma, chronic obstructive pulmonary disease (COPD) or obesity-hypoventilation syndrome could influence the performance of methodologies based on oximetry alone.[4] Regarding this issue, the rules of the American Academy of Sleep Medicine (AASM) about the use of portable monitoring as an alternative to PSG were taken into account, which recommend that portable monitoring should not be used in patient groups with significant comorbid medical conditions, patients suspected of having other sleeps disorders and for general screening

of asymptomatic populations.[7] Our results suggest that LR and SVMs classifiers fed with reduced input feature subsets provide high performance and stable classifiers across independent populations form different sleep units. However, further analyses are needed to assess its robustness against common limitations of oximetry. Moreover, further work is required to test the performance of our methodology from ambulatory portable monitoring at patient's home. An additional limitation should be taken into account. In the present study, an AHI ≥ 10 e/h was used as threshold for a positive diagnosis of SAHS in both sleeps units under study. However, there is not a standardized AHI threshold for SAHS diagnosis[65] and different cut-off points (commonly 5, 10 and 15 e/h) have been widely applied. Therefore, further analysis is needed to assess the influence of changes in the diagnostic threshold in order to generalize our methodology. In addition, SAHS-positive patients are predominant in the training set, which could influence the model design and the performance of the classifiers. Finally, additional drawbacks regarding feature selection must be considered. As optimization algorithms, GAs achieved higher performance in the training set. However, significant unbalanced values of sensitivity and specificity were reached in the validation and the test sets when large feature subsets are selected. Genetic programming, which is a significant extension of GAs[66] could be applied to further assess the usefulness of evolutionary algorithms for feature selection in the context of SAHS diagnosis from oximetry. Moreover, additional feature selection techniques could be applied to further assess our methodology, such as independent component analysis, subspace clustering or simulated annealing.

6. Conclusions

In summary, three feature selection approaches (PCA, FSFS and GAs) and three classification algorithms (FLD, LR and SVMs) were assessed in the context of SAHS diagnosis using populations from two independent sleep units. Optimum classification schemes from PCA achieved highly unbalanced sensitivity–specificity pairs and poor accuracy both in the validation and test sets regardless of the classifier. Additionally, performance of optimum classifiers from GAs significantly decreased when large feature subsets are selected due to overfitting on the training set. On the other hand, FSFS + LR, FSFS + SMV and GAs + SVM classifiers, composed of a reduced number of features automatically selected, achieved a balanced sensitivity–specificity pair and high accuracy on populations from both sleep units. Thus, greedy search feature selection strategies and classifiers with high generalization ability against overfitting could be useful to avoid noisy and redundant information and to obtain complementary features in order to enhance SAHS detection from oximetry.

Acknowledgments

This research was supported in part by the Ministerio de Economía y Competitividad and FEDER under project TEC2011-22987, the Proyecto Cero 2011 on Ageing from Fundación General CSIC, Obra Social La Caixa and CSIC and project VA111A11-2 from Consejería de Educación (Junta de Castilla y León). D. Álvarez was in receipt of a PIRTU grant from the Consejería de Educación de la Junta de Castilla y León and the European Social Fund (ESF).

References

1. T. Young, J. Skatrud and P. E. Peppard, Risk factors for obstructive sleep apnea in adults, *J. Am. Med. Assoc.* **291** (2004) 2013–2016.
2. S. P. Patil, H. Schneider, A. R. Schwartz and P. L. Smith, Adult obstructive sleep apnea: Pathophysiology and diagnosis, *Chest* **132** (2007) 325–337.
3. F. Lopez-Jimenez, F. H. Sert, A. Gami and V. K. Somers, Obstructive sleep apnea: Implications for cardiac and vascular disease, *Chest* **133** (2008) 793–804.
4. W. W. Flemons, M. R. Littner, J. A. Rowlet, P. Gay, W. M. Anderson, D. W. Hudgel, R. D. McEvoy and D. I. Loube, Home diagnosis of sleep apnea: A systematic review of the literature, *Chest* **124** (2003) 1543–1579.
5. W. A. Whitelaw, R. F. Brant and W. W. Flemons, Clinical usefulness of home oximetry compared with polysomnography for assessment of sleep apnea, *Am. J. Respir. Crit. Care Med.* **171** (2005) 188–193.
6. N. Netzer, A. H. Eliasson, C. Netzer and D. A. Kristo, Overnight pulse oximetry for sleep-disordered breathing in adults, *Chest* **120** (2001) 625–633.
7. N. A. Collop, W. Mc, D. Anderson, B. Boehlecke, D. Claman, R. Goldberg, D. J. Gottlieb, D. Hudhel,

M. Sateia and R. Schwab, Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients, *J. Clin. Sleep Med.* **3** (2007) 737–747.
8. U. J. Magalang, J. Dmochowski, S. Veeramachaneni, A. Draw, M. J. Mador, A. El-Solh and B. J. B. Grant, Prediction of the apnea-hypopnea index from overnight pulse oximetry, *Chest* **124** (2003) 1694–1701.
9. L. G. Olson, A. Ambrogetti and S. G. Gyulay, Prediction of sleep-disordered breathing by unattended overnight oximetry, *J. Sleep Res.* **8** (1999) 51–55.
10. H. Chung-Ching H and Y. Chung-Chieh, Smoothed periodogram of oxyhemoglobin saturation by pulse oximetry in sleep apnea syndrome, *Chest* **131** (2007) 750–757.
11. J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo and C. Zamarón, Assessment of four statistical pattern recognition techniques to assist in obstructive sleep apnoea diagnosis from nocturnal oximetry, *Med. Eng. Phys.* **31** (2009) 971–978.
12. J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo, M. López and C. Zamarrón, Radial basis function classifiers to help in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry, *Med. Biol. Eng. Comput.* **46** (2008) 323–332.
13. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, New York, 2006).
14. I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Learn. Res.* **3** (2003) 1157–1182.
15. R. Kohavi and G. John, Wrappers for feature selection, *Artif. Intell.* **97** (1997) 273–324.
16. E. García-Cuesta, I. M. Galván and A. J. de Castro, Recursive discriminant regression analysis to find homogeneous groups, *Int. J. Neural Syst.* **21** (2011) 95–101.
17. K. Z. Mao, Fast orthogonal forward selection algorithm for feature subset selection, *IEEE Trans. Neural Netw.* **13** (2002) 1218–1224.
18. J. D. Jobson, *Applied Multivariate Data Analysis*, Vol. II: *Categorical and Multivariate Methods* (Springer-Verlag, New York, 1991).
19. D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression* (John Wiley & Sons, New York, 1989).
20. W. Siedlecki and J. Sklansky, A note on genetic algorithms for large scale feature selection, *Pattern Recognit. Lett.* **10** (1989) 335–347.
21. V. N. Vapnik, An overview of statistical learning theory, *IEEE Trans. Neural Netw.* **10** (1999) 988–999.
22. M. Al-Naser and U. Soderstrom, Reconstruction of occluded facial images using asymmetrical principal component analysis, *Integr. Comput. Aid. E* **19** (2012) 273–283.
23. P. Baraldi, R. Canesi, E. Zio, R. Seraoui and R. Chevalier, Genetic algorithm-based wrapper approach for grouping condition monitoring signals of nuclear power plant components, *Integr. Comput. Aid. E* **18** (2011) 221–234.
24. E. Yom-Tov G. F. and Inbar, Feature selection for the classification of movements from single movement-related potentials, *IEEE Trans. Neural Syst. Rehabil. Eng.* **10** (2002) 170–177.
25. G. C. Marano, G. Quaranta and G. Monti, Modified genetic algorithm for the dynamic identification of structural systems using incomplete measurements, *Comput. Aided Civ. Inf.* **26** (2011) 92–110.
26. R. Jafarkhani and S. F. Masri, Finite element model updating using evolutionary strategy for damage detection, *Comput. Aided Civ. Inf.* **26** (2011) 207–224.
27. Y. Lee and C. H. Wei, A computerized feature selection using genetic algorithms to forecast freeway accident duration times, *Comput. Aided Civ. Inf.* **25** (2010) 132–148.
28. K. S. Tang, K. F. Man, S. Kwong and Q. He, Genetic algorithms and their applications, *IEEE Signal Process. Mag.* **13** (1996) 22–37.
29. P. Patrinos, A. Alexandridis, K. Ninos and H. Sarimveis, Variable selection in nonlinear modeling based on RBF networks and evolutionary computation, *Int. J. Neural Syst.* **20** (2010) 365–379.
30. D. Álvarez, R. Hornero, J. V. Marcos and F. del Campo, Multivariate analysis of blood oxygen saturation recordings in obstructive sleep Apnea diagnosis, *IEEE Trans. Biomed. Eng.* **57** (2010) 2816–2824.
31. L. I. Kuncheva and J. J. Rodríguez, Classifier ensembles for fMRI data analysis: An experiment, *J. Magn. Reson. Imaging* **28** (2010) 583–593.
32. G. Forman, An extensive empirical study of feature selection metrics for text classification, *J. Mach. Learn. Res.* **3** (2003) 1289–1305.
33. V. P. Jumutc, P. Zayakin and A. Borisov, Ranking-based kernels in applied biomedical diagnostics using support vector machine, *Int. J. Neural Syst.* **21** (2011) 459–473.
34. U. R. Acharya, S. V. Sree and J. S. Suri, Automatic detection of epileptic EEG signals using higher order cumulant features, *Int. J. Neural Syst.* **21** (2011) 403–414.
35. E. D. Wandekokem, E. Mendel, F. Fabris, M. Valentim, R. J. Batista, F. M. Varejao and T. W. Rauber, Diagnosing multiple faults in oil rig motor pumps using support vector machine classifier ensembles, *Integr. Comput. Aid. E* **18** (2011) 61–74.
36. J. D. Jobson, *Applied Multivariate Data Analysis*, Vol. I: *Regression and Experimental Design* (Springer-Verlag, New York, 1991).
37. J. Poza, R. Hornero, D. Abásolo, A. Fernández and M. García, Extraction of spectral based measures

from MEG background oscillations in Alzheimer's disease, *Med. Eng. Phys.* **29** (2007) 1073–1083.
38. C. Zamarrón, P. V. Romero, J. R. Rodríguez and F. Gude, Oximetry spectral analysis in the diagnosis of obstructive sleep apnea, *Clin. Sci.* **97** (1999) 467–473.
39. S. M. Pincus, Assessing serial irregularity and its implications for health, *Ann. NY Acad. Sci.* **954** (2001) 245–267.
40. U. R. Acharya, E. C-P. Chua, K. C. Chua, L. C. Min and T. Tamura, Analysis and automatic identification of sleep stages using higher order spectra, *Int. J. Neural Syst.* **20** (2010) 509–521.
41. U. R. Acharya, S. V. Sree, S. Chattophadyay, W. Yu and P. C. A. Ang, Application of recurrence quantification analysis for the automated identification of epileptic EEG signals, *Int. J. Neural Syst.* **21** (2011) 199–211.
42. U. R. Acharya, S. V. Sree, A. P. C. Alvin and J. S. Suri, Application of non-linear and wavelet based features for the automated identification of epileptic EEG signals, *Int. J. Neural Syst.* **22** (2012) 1250002–14.
43. D. Álvarez, R. Hornero, D. Abásolo, F. del Campo and C. Zamarrón, Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection, *Physiol. Meas.* **27** (2006) 399–412.
44. D. Álvarez, R. Hornero, M. García, F. del Campo and C. Zamarrón, Improving diagnostic ability of blood oxygen saturation from overnight pulse oximetry in obstructive sleep apnea detection by means of central tendency measure, *Artif. Intell. Med.* **41** (2007) 13–24.
45. J. S. Richman and J. R. Moorman, Physiological time series analysis using approximate entropy and sample entropy, *Am. J. Physiol. Heart Circ. Physiol.* **278** (2000) H2039–H2049.
46. M. E. Cohen, D. L. Hudson and P. C. Deedwania, Applying continuous chaotic modeling to cardiac signals analysis, *IEEE Eng. Med. Biol.* **15** (1996) 97–102.
47. M. E. Cohen and D. L. Hudson, New chaotic methods for biomedical signal analysis, in *Proceedings of the 2000 IEEE EMBS Int. Conf. Information Technology Applications in Biomedicine* (Arlington USA, 2000), pp. 123–128.
48. X.-S. Zhang, R. J. Roy and E. W. Jensen, EEG complexity as a measure of depth of anesthesia for patients, *IEEE Trans. Biomed. Eng.* **48** (2001) 1424–1433.
49. C. J. Stam, Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field, *Clinical Neurophysiol.* **116** (2005) 2266–2301.
50. G. Claeskens, C. Croux and J. V. Kerckhoven, An information criterion for variable selection in support vector machines, *J. Mach. Learn. Res.* **9** (2008) 541–558.
51. A. Gelman, Scaling regression inputs by dividing by two standard deviations, *Stat. Med.* **27** (2008) 2865–2873.
52. J. M. Sutter and J. H. Kalivas, Comparison of forward selection, backward elimination and generalized simulated annealing for variable selection, *Microchem. J.* **47** (1993) 60–66.
53. G. H. John, R. Kohavi and K. Pfleger, Irrelevant features and the subset selection problem, in *Machine Learning*: *Proc. Eleventh International Conf.* (1994) pp. 121–129.
54. M. R. Boutell, J. Luo, X. Shen and C. M. Brown, Learning multi-label scene classification, *Pattern Recognit.* **37** (2004) 1757–1771.
55. S. Haykin, *Neural Networks*: *A Comprehensive Foundation* (Prentice Hall Inc., New Jersey, 1999).
56. A. A. El-Solh, M. J. Mador, E. Ten-Brock, D. W. Shucard, M. Abul-Khoudoud and B. J. B. Grant, Validity of neural network in sleep apnea, *Sleep* **22** (1999) 105–111.
57. T. Penzel, J. W. Kantelhardt, L. Grote, J.-H. Peter and A. Bunde, Comparison of detrended fluctuation analysis and spectral analysis of heart rate variability in sleep and sleep apnea, *IEEE Trans. Biomed. Eng.* **50** (2003) 1143–1151.
58. P. De Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan and M. O'Malley, Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnea, *IEEE Trans. Biomed. Eng.* **50** (2003) 686–696.
59. M. O. Mendez, J. Corthout, S. Van Huffel, M. Matteucci, T. Penzel, S. Cerutti and A. M. Bianchi, Automatic screening of obstructive sleep apnea from the ECG based on empirical mode decomposition and wavelet analysis, *Physiol. Meas.* **31** (2010) 273–289.
60. A. H. Khandoker, M. Palaniswami and C. K. Karmakar, Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings, *IEEE Trans. Inf. Technol. Biomed.* **13** (2009) 37–48.
61. A. H. Khandoker, C. K. Karmakar and M. Palaniswami, Automated recognition of patients with obstructive sleep apnoea using wavelet-based features of electrocardiogram recordings, *Comput. Biol. Med.* **39** (2009) 88–96.
62. J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo and M. Aboy, Automated detection of obstructive sleep apnoea syndrome from oxygen saturation recordings using linear discriminant analysis, *Med. Biol. Eng. Comput.* **48** (2010) 895–902.
63. G. C. Gutiérrez-Tobal, R. Hornero, D. Álvarez, J. V. Marcos and F. del Campo, Linear and nonlinear analysis of airflow recordings to help in sleep apnoea–hypopnoea syndrome diagnosis, *Physiol. Meas.* **33** (2012) 1261–1275.
64. D. Álvarez, R. Hornero, J. V. Marcos and F. del Campo, Feature selection from nocturnal oximetry

using genetic algorithms to assist in obstructive sleep apnoea diagnosis, *Med. Eng. & Phys.* **34** (2012) 1049–1057.
65. N. A. Collop, S. L. Tracy, V. Kapur, R. Mehra, D. Kuhlmann, S. A. Fleishman and J. M. Ojile, Obstructive sleep apnea devices for out-of-center (OOC) testing: Technology evaluation, *J. Clin. Sleep Med.* **7** (2011) 531–548.
66. P. Day and A. K. Nandi, Evolution of super features through genetic programming, *Expert Syst.* **28** (2011) 167–184.