Med Biol Eng Comput (2013) 51:1367–1380
DOI 10.1007/s11517-013-1109-7

ORIGINAL ARTICLE

Pattern recognition in airflow recordings to assist in the sleep apnoea–hypopnoea syndrome diagnosis

Gonzalo C. Gutiérrez-Tobal · Daniel Álvarez · J. Víctor Marcos · Félix del Campo · Roberto Hornero

Received: 11 March 2013 / Accepted: 1 September 2013 / Published online: 22 September 2013

Abstract This paper aims at detecting sleep apnoea–hypopnoea syndrome (SAHS) from single-channel airflow (AF) recordings. The study involves 148 subjects. Our proposal is based on estimating the apnoea–hypopnoea index (AHI) after global analysis of AF, including the investigation of respiratory rate variability (RRV). We exhaustively characterize both AF and RRV by extracting spectral, nonlinear, and statistical features. Then, the fast correlation-based filter is used to select those relevant and non-redundant. Multiple linear regression, multi-layer perceptron (MLP), and radial basis functions are fed with the features to estimate AHI. A conventional approach, based on scoring apnoeas and hypopnoeas, is also assessed for comparison purposes. An MLP model trained with AF and RRV selected features achieved the highest agreement with the true AHI (intra-class correlation coefficient = 0.849). It also showed the highest diagnostic ability, reaching 92.5 % sensitivity, 89.5 % specificity and 91.5 % accuracy. This suggests that AF and RRV can complement each other to estimate AHI and help in SAHS diagnosis.

G. C. Gutiérrez-Tobal (✉) · D. Álvarez · J. V. Marcos · R. Hornero
Biomedical Engineering Group, E.T.S.I. de Telecomunicación, University of Valladolid, Paseo Belén 15, 47011 Valladolid, Spain
e-mail: gonzalo.gutierrez@gib.tel.uva.es

F. del Campo
Servicio de Neumología, Hospital Universitario Río Hortega, c/Dulzaina 2, 47012 Valladolid, Spain

F. del Campo
Facultad de Medicina, University of Valladolid, Avenida Ramón y Cajal 7, 47005 Valladolid, Spain

Keywords Sleep apnoea–hypopnoea syndrome · Airflow · Respiratory rate variability · AHI estimation · Pattern recognition

1 Introduction

The sleep apnoea–hypopnoea syndrome (SAHS) is a disease characterized by recurrent episodes of total absence (apnoeas) or significant reduction (hypopnoeas) in airflow (AF) during sleep. SAHS is highly prevalent since up to 5 % of adults are affected [41]. It has been usually related to cardiovascular illnesses [25], motor vehicle collisions [35], and occupational accidents [24]. Recently, it has been also associated with cancer incidence [8].

The current diagnostic standard test is nocturnal polysomnography (PSG). It requires monitoring and recording multiple physiological signals from patients [32]. The origin of the signals can be electrical or mechanical, and each of them can involve one or several channels. The apnoea–hypopnoea index (AHI), which is derived from PSG, is used to establish SAHS. Physicians have to perform an offline inspection of signals such as electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG), oxygen saturation (SpO2), or AF to obtain AHI. Thus, PSG is technically complex and time-consuming [6, 14]. Moreover, it is also costly since requires expensive equipment as well as expert workforce overnight [14]. These restrictions limit the availability of specialized sleep units, leading to long waiting lists and increasing the time until diagnosis and treatment [11]. Thereby, simplifying SAHS diagnosis has become a major concern.

New alternative methods have been proposed to overcome the PSG drawbacks. A common approach is to

analyse reduced sets of signals from PSG in order to decrease complexity, cost, and diagnostic time [14]. We propose evaluating the utility of single-channel AF data to assist in SAHS diagnosis. The respiratory rate variability (RRV), derived from AF [10], is also investigated. The waveform of both signals is directly modified by the occurrence of apnoea and hypopnoea events [10, 19]. Hence, their study is a natural way of dealing with the problem. There exist many recent works focused on processing AF to determine SAHS. Most of them are aimed at scoring apnoeic events to estimate AHI [5, 11, 30, 36, 39]. By contrast, our proposal performs a direct estimation of AHI after a comprehensive analysis of AF and RRV. Thus, the first step is the extraction of statistical moments, nonlinear measures and spectral parameters from the recordings in order to characterize them [2, 15, 26]. This exhaustive characterization of AF and RRV may lead to obtain redundant or non-relevant features. Hence, we include a second step consisting of a feature selection procedure using the fast correlation-based filter (FCBF) [42]. FCBF relies on symmetrical uncertainty (SU) and has been already involved in biomedical applications for cancer recognition [18], neonatal seizure detection [1], or gene classification [13]. Its purpose is to filter data according to their relevancy and redundancy. A final step is included to estimate AHI. Thus, we feed three pattern recognition techniques with the extracted features: multiple linear regression (MLR), multi-layer perceptron neural network (MLP), and radial basis function neural network (RBF). They represent common linear (MLR) and nonlinear (MLP, RBF) methodologies to perform regression tasks [7]. We evaluate the agreement between these estimations and the true AHI of subjects as well as their diagnostic ability. Additionally, we also conduct a conventional approach (scoring apnoeas and hypopnoeas) for comparison purposes. Our hypothesis is that relevant and non-redundant features from single-channel AF could help in SAHS diagnosis by estimating AHI.

2 Materials and methods

Figure 1 presents a scheme of the general methodology carried out in this study. It includes the feature extraction, the feature selection, and the AHI estimation steps, as well as the two kinds of evaluations applied to the estimations from each pattern recognition method and the conventional approach.

2.1 Subjects and signals

This study involved recordings from 148 subjects (100 SAHS-positive and 48 SAHS-negative). The AF data were obtained from nocturnal PSG, which was conducted in the sleep unit of the Hospital Universitario Río Hortega (Valladolid, Spain). All subjects were suspected of suffering from SAHS before undergoing PSG due to common symptoms such as daytime sleepiness, loud snoring, nocturnal choking, awakenings, and referring apnoeic events. The physicians established the AHI threshold for a positive diagnosis in 10 events per hour (e/h). The score of apnoeic events was done following the rules of the American Academy of Sleep Medicine (AASM) [19]. Thus, apnoeas were defined as 10-s-or-more episodes of complete cessation of AF. Accordingly, hypopnoeas were defined as 10-s-or-more episodes of 30 % of AF reduction accompanied by a 4 % or more decrease in the saturation of haemoglobin. The Review Board on Human Studies accepted the protocol, and all the subjects gave their informed consent to participate in the study.

The proportion of male subjects was 79 %. No statistically significant differences between SAHS-positive and SAHS-negative samples were encountered in the body

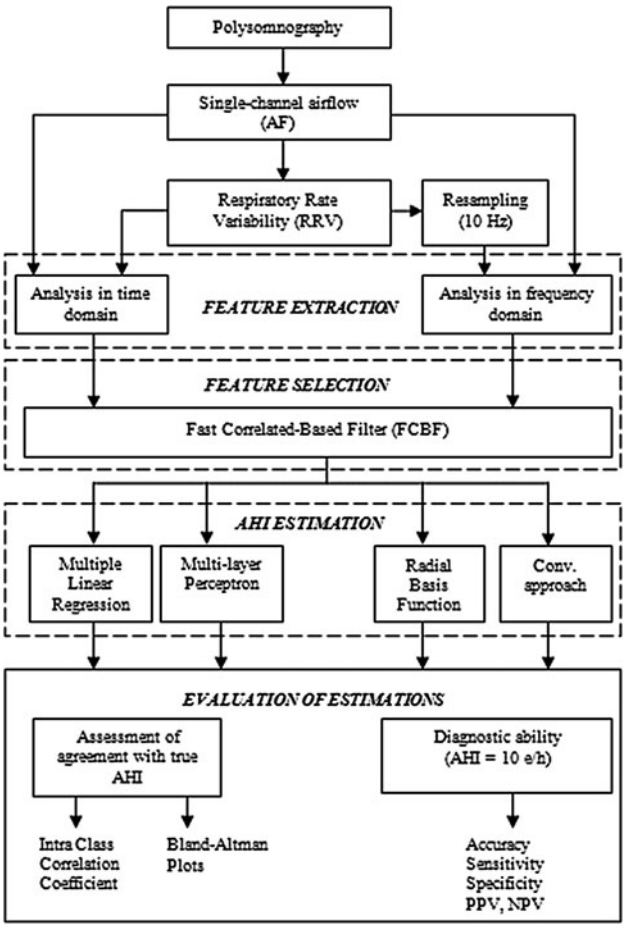

Fig. 1 General scheme of the methodology carried out in the study. *AHI* apnoea-–hypopnoea index, *PPV* positive predictive value, *NPV* negative predictive value

mass index (BMI) or age. The entire group was randomly divided into a training group (60 %) and a test group (40 %). Table 1 summarizes demographic and clinical data from the entire sample, the training group and the test group.

The acquisition of signals during PSG was done by means of a polygraph (Alice 5, Respironics, Philips Healthcare, The Netherlands). AF was obtained through a thermistor (Pro-Tech, Respironics, Philips Healthcare, The Netherlands) at the sample rate of 10 Hz. The length of the AF recordings was 7.24 ± 0.38 h (mean ± standard deviation). An anti-aliasing filter was applied to satisfy the Nyquist–Shannon theorem. The RRV signal was obtained from AF by measuring the time between consecutive breaths [10]. Thereby, we examined the first derivative of AF to find time intervals in which the original signal grew. We located the AF maximums at each interval. To derive RRV, consecutive locations were used as references to measure the time from one breath to the next [21].

Table 1 Demographic and clinical data for all subjects under study (mean ± standard deviation)

	All	SAHS positive	SAHS negative
Subjects	148	100	48
All subjects			
Age (years)	50.9 ± 11.7	51.9 ± 11.4	48.7 ± 12.1
Males (%)	79.0	85.0	66.7
BMI (kg/m^2)	29.2 ± 4.7	29.7 ± 4.5	28.1 ± 5.0
Recording time (h)	7.24 ± 0.38	7.23 ± 0.36	7.27 ± 0.43
AHI (h^{-1})		37.15 ± 25.81	4.13 ± 2.39
Subjects	All 89	SAHS positive 60	SAHS negative 29
Training set			
Age (years)	51.9 ± 11.8	52.8 ± 11.9	50.2 ± 11.7
Males (%)	80.9	88.3	65.5
BMI (kg/m^2)	29.8 ± 5.0	30.5 ± 5.2	28.4 ± 5.7
Recording time (h)	7.22 ± 0.43	7.21 ± 0.38	7.24 ± 0.52
AHI (h^{-1})		37.4 ± 27.2	3.8 ± 2.4
Subjects	All 59	SAHS positive 40	SAHS negative 19
Test set			
Age (years)	49.2 ± 11.3	50.5 ± 10.7	46.5 ± 12.5
Males (%)	76.3	80.0	68.4
BMI (kg/m^2)	28.3 ± 4.1	28.6 ± 3.5	27.7 ± 5.2
Recording time (h)	7.27 ± 0.29	7.26 ± 0.32	7.30 ± 0.23
AHI (events/h)		26.2 ± 17.2	4.3 ± 2.3

BMI body mass index, *AHI* apnoea–hypopnoea index

2.2 Definition of spectral bands of interest

The recurrent behaviour of apnoeas and hypopnoeas can be characterized by analysing AF and RRV in the frequency domain. Moreover, according to previous studies [15], differences in the spectrum of SAHS-positive and SAHS-negative samples are expected. Thus, the power spectral density (PSD) of the recordings was computed in order to establish these differences. PSD was estimated using the nonparametric Welch method, which is suitable for non-stationary signal analysis [38]. A Hamming window of 2048 (204.8 s) samples (50 % overlap and 4,096-point DFTs) was used. Cubic spline interpolation was previously applied to RRV series in order to resample the recordings to a constant sample rate (10 Hz). The interpolation is not needed to perform the analysis in time domain, and therefore, the resampled version of the RRV recordings was not used in that case.

Spectral bands of interest were defined for AF and RRV. The Mann–Whitney test was applied to each SAHS-positive and SAHS-negative full PSD from the training group. Thus, a p value was computed for each frequency. We located those frequencies at which the lowest p value for AF and RRV was reached (p value $\ll 0.01$). We set the corresponding band limits around these frequencies. In order to minimize type I errors, we chose those frequencies with a corresponding p value smaller than one order of magnitude. Thereby, we maximized the likelihood of defining bands in which truly exist significant differences. According to this procedure, the following spectral bands of interest were determined: [0.022–0.058] Hz for AF and [0.085–0.134] Hz for RRV. Figure 2a, b shows the averaged PSD of SAHS-positive and SAHS-negative samples for AF and RRV, respectively, in the training set.

2.3 Feature extraction

Up to 19 features were used to exhaustively characterize AF and RRV. Statistical moments, nonlinear measures, and spectral parameters were extracted from each full AF and RRV recordings. Thus, subjects were described by patterns composed of the corresponding values for each feature.

2.3.1 Statistical moments

We expected differences between the distribution of the time series amplitude values from SAHS-positive and SAHS-negative samples [15]. Hence, four statistical moments were extracted from AF and RRV. Mean (M_{t1}), standard deviation (M_{t2}), skewness (M_{t3}), and kurtosis (M_{t4}) were computed to quantify central tendency, dispersion, asymmetry, and peakedness of data, respectively.

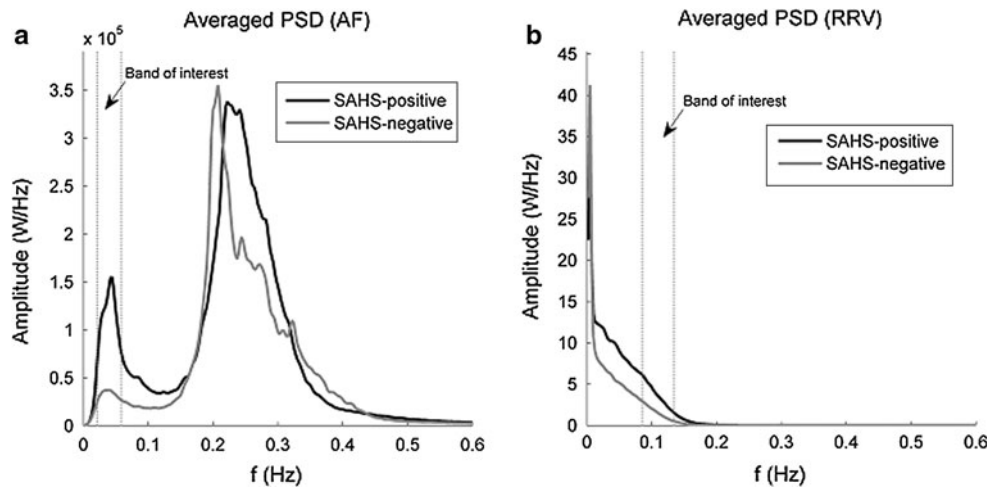

Fig. 2 Low-frequency representation of the averaged PSD for **a** AF and **b** RRV. SAHS-positive group in *solid black line*. SAHS-negative group in *solid grey line*. Corresponding bands of interest into *dashed lines*

2.3.2 *Nonlinear features*

Nonlinear features were used to measure the variability, complexity, and irregularity of the time series. We used central tendency measure (CTM), Lempel–Ziv complexity (LZC), and approximate entropy (ApEn) for this purpose. These methods have been already used to characterize SAHS in previous studies [2, 15, 26].

- Central tendency measure quantifies the degree of variability in time series [8]. It is based on first-order difference plots that can be generated representing $x[n+2]-x[n+1]$ versus $x[n+1]-x[n]$, where $x[n]$ are the time series values. CTM is computed by counting the points falling within a preselected radius ρ and dividing that count by the total number of points [9]. Values closer to 1 indicate lower variability, whereas values closer to 0 indicate higher variability.
- Lempel–Ziv complexity is a measurement of complexity in finite sequences [23]. Thus, the conversion of time series into a finite sequence of symbols is needed. Binary conversion has been commonly applied by using the median as a threshold [29]. Once the sequence is obtained, it is scanned from left to right in order to find new subsequences of consecutive characters [43]. The final number of these subsequences is normalized to make the method independent of the length of sequences. Larger values of LZC correspond to higher complexity [43].
- ApEn measures the irregularity of time series. It assigns higher values to higher irregularity [34]. ApEn was originally developed to be applied over short and noisy data sets and requires the specification of two design parameters: a length m and a tolerance window r [33]. These are used to establish the logarithmic likelihood resulting from those close patterns (within r) for m contiguous observations, which remain close (within the same r) for $m+1$ contiguous observations.

Optimum radius ρ (CTM), length m, and tolerance r (ApEn) were determined by a p value-based methodology [17]. In the case of ApEn, we evaluated $m = 1, 2$ and r ranging 0.10–0.25 times the standard deviation of the times series (with a 0.05 step). These values produce good statistical reproductibility for ApEn [34]. A wide range of values for ρ were also assessed (0.1–30, with a 0.1 step). We selected those configurations, which showed the lowest p value between SAHS-positive and SAHS-negative samples in the training group:

- AF: $\rho = 0.8$ (CTM), $m = 2$, $r = 0.2$ times standard deviation (ApEn).
- RRV: $\rho = 4.8$ (CTM), $m = 2$, $r = 0.2$ times standard deviation (ApEn).

2.3.3 *Spectral features*

A total of 12 parameters were extracted from the full PSD (6) and the band of interest (6) for every AF and RRV recording.

- First-to-fourth statistical moments, which were also extracted in the frequency domain ($M_{f1} - M_{f4}$).
- Peak amplitude (PA), taken as the maximum value of PSDs in a given frequency interval.
- The Wootters distance (WD) [40], which is a disequilibrium measure. WD assigns higher values when the PSD is concentrated into a narrow frequency band (as in sum of sinusoids). If it is uniformly distributed along frequencies (white noise), WD equals zero [27].

2.4 Automatic feature selection: FCBF

After the feature extraction stage, the FCBF algorithm automatically selected relevant and non-redundant features [42]. FCBF is a filter method, which is not dependent on posterior analysis. It relies on symmetrical uncertainty

(SU), which is a normalized measure of information gain (IG) between two variables [42]. The method is divided into two steps. First, a relevance analysis of features was done. SU between the features (X_i) and AHI (Y) was computed as follows:

$$\mathrm{SU}_i(X_i, Y) = 2\frac{\mathrm{IG}_i(X_i, Y)}{H_i(X_i) + H(Y)} \quad i = 1, 2, \ldots, N, \tag{1}$$

where H refers to Shannon's entropy [42], and N is the number of features extracted. SU is restricted to the range [0, 1]: 1 indicates that knowing one feature it is possible to completely predict the other, whereas 0 indicates that the two features are independent [42]. Once SU_i were computed, the features were ranked from more relevant (higher SU_i) to less relevant (lower SU_i). The mean of all SU_i values was used as a cut-off to perform a preselection. The second step was a redundancy analysis. SU between each pair of preselected features ($\mathrm{SU}_{i,j}$) was sequentially computed beginning from the most relevant ones. When $\mathrm{SU}_{i,j} \geq \mathrm{SU}_i$, the feature j was discarded due to redundancy and was not taken into account in successive comparisons. The final selected features were those not discarded after ending the procedure.

2.5 Pattern recognition methods

As described above, the extracted features were used to form patterns (vectors). Thus, a subject n was characterized by a pattern x_n. Each subject and its corresponding x_n are associated with an AHI value (t_n). We modelled the statistical relationship between patterns and AHI by means of pattern recognition techniques. The utility of three methods to provide a reliable estimation (y) of the AHI was evaluated.

2.5.1 Multiple linear regression (MLR)

Multiple linear regression is a traditional method to predict an output variable, y, through data from a multivariate pattern, $x_1, x_2, \ldots, x_N$. It assumes a linear relationship between the former and the latter [20]:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_N x_N = \mathbf{w}^T \mathbf{x}, \tag{2}$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_N)^T$ are the regression coefficients for each input variable and the intercept (w_0). The computation of $\mathbf{w}$ is done by means of the sum of squares error (E_D) minimization [7]:

$$E_\mathrm{D} = \frac{1}{2}\sum_{n=1}^{N} [y(\mathbf{x}_n, \mathbf{w}) - t_n]^2. \tag{3}$$

2.5.2 Multi-layer perceptron (MLP) network

The MLP network is a model inspired by the human brain. The architecture of MLP is arranged in several interconnected layers (input, hidden layers, and output), which are composed of simple units known as perceptrons [7]. Each perceptron is characterized by an activation function $g(\bullet)$, and their connections to perceptrons from other layers are associated with adaptive weights (w_{ij}).

The output layer provides the response, y. Since our purpose is to estimate a continuous variable, a single output unit with a linear activation function was used [28]. Additionally, we implemented a single hidden layer composed of perceptrons with nonlinear activation functions. This configuration is known to be able of providing universal approximation [7]. Thus, y can be expressed as follows:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{N_H} \left[w_j g\left(\sum_{i=1}^{d} w_{ij} x_i + b_j \right) + b_0 \right], \tag{4}$$

where $\mathbf{w}$ is a vector with all the adaptive parameters (weights and bias), w_j is the weight connecting hidden units h_j with the output unit, b_0 is the bias associated with the output unit, w_{ij} is the weight connecting the input unit i with hidden unit h_j, and b_j is its associated bias. N_H, the number of perceptrons in the hidden layer, is a design parameter. Weights were optimized with patterns from the training group, by sum of squares error function minimization. Scaled conjugate gradient was used for this purpose [7].

Weight decay regularization was used to achieve good generalization. Thus, a penalty term (Ω) was added to the error function E_D, to favour small weights [7]:

$$\begin{aligned} E_T &= E_\mathrm{D} + \Omega \\ &= E_\mathrm{D} + \upsilon \sum_i w_i^2 = \frac{1}{2}\sum_{n=1}^{N} [y(\mathbf{x_n}, \mathbf{w}) - t_n]^2 + \upsilon \sum_i w_i^2, \end{aligned} \tag{5}$$

where Ω is the sum of squares of the network weights, and υ is known as the regularization parameter, which has to be configured.

2.5.3 Radial basis function (RBF) network

Radial basis function is a different neural network approach. This network is composed of a hidden and an output layer. The output y is computed from the responses provided by the basis functions $\psi(\bullet)$ in the hidden layer nodes. These functions only depend on the radial distance (typically the Euclidian distance) between the input vector $\mathbf{x}$ and a set of suitable centres $\mathbf{c}_j$ [7]. A single output neuron with a linear activation function was used to implement the output layer, since the problem was a single variable regression task. Thus, y is given by the following expression [7]:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{N_B} w_j \psi_j(||\mathbf{x} - \mathbf{c}_j||) + b, \quad (6)$$

where N_B is the number of basis functions (or centres), $\mathbf{c}_j$ is the centre of function ψ_j, w_j is the weight connecting ψ_j and the output neuron, and b is the bias parameter for this neuron. A Gaussian function is commonly used for $\psi(\bullet)$ [7]:

$$\psi_j(\mathbf{x}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{c}_j||^2}{2\sigma_j^2}\right). \quad (7)$$

where σ_j is the standard deviation (width) of each function. Thus, the numbers of centres (N_B) and their locations $\mathbf{c}_j$ as well as the widths of radial basis functions σ_j and the weights w_j are parameters to be optimized. N_B and σ_j were experimentally determined during a design stage. *K*-means algorithm was used to optimize the location of centres [7], and w_j was computed from the solution of linear equations following the sum of squares error minimization [7]. All of them were configured by patterns from the training group.

2.6 Conventional approach

A conventional way of dealing with the problem of automatic SAHS diagnosis is to detect and score respiratory events in AF signal. Then, an estimation of AHI (AHI_c) can be derived by dividing the number of these events by the sleep time. We implemented a scoring algorithm to compare it with the proposed pattern recognition techniques. A peak detection algorithm was used to locate inspiratory onsets and endings in AF [21]. These values determined the amplitude of every inspiration. Following the rules of the AASM, we scored those respiratory events that matched 30 % or more drop from the baseline and lasted a minimum of 10 s [19]. The baseline was determined by the mean amplitude of the s previous inspirations [16]. Hence, s was a design parameter. The same methodology than in the case of the parameters of nonlinear features was used to optimize s. We computed AHI_c in the training group by varying s from 1 to 10 (with a 1 step). For each s, the Mann–Whitney test was used to obtain the p value between the AHI_c from the SAHS-positive and the SAHS-negative samples. The greatest statistical difference, i.e. the lowest p value, was obtained for $s = 3$, which was established as the optimum value.

2.7 Statistical analysis

Data did not pass the Lilliefors normality test. Hence, the nonparametric Mann–Whitney significance test was used to assess the differences in SAHS-positive and SAHS-negative samples. We used the intra-class correlation coefficient (*ICC*) and Bland–Altman plots as assessment of agreement between estimated and true AHI. The diagnostic ability of the estimations was assessed by means of sensitivity (proportion of SAHS-positive patients correctly classified), specificity (proportion of SAHS-negative subjects correctly classified), accuracy (percentage of subjects correctly classified over the entire sample), positive predictive value (proportion of positive test result which are true positives), and negative predictive value (proportion of negative test result which are true negatives).

3 Results

Three sets of complete patterns were defined: patterns composed of the 19 AF features (P^c_{AF}); patterns composed of the 19 RRV features (P^c_{RRV}); and patterns composed of the 38 AF and RRV features (P^c_{AF-RRV}). Then, we used the training group to select relevant and non-redundant features through FCBF algorithm. Thus, three new sets of reduced patterns, formed with filtered features, were obtained (P^r_{AF}, P^r_{RRV}, and P^r_{AF-RRV}). The training group was also used in the process of obtaining specific pattern recognition models. This process was divided into two stages: design and training. In the first one, the *ICC* was computed using a leave-one-out cross-validation (loo-cv) procedure to find optimum design parameters for MLP and RBF. In the second one, MLR, MLP, and RBF models were trained by the use of the entire training group.

The test group was used to evaluate our methodology. *ICC* and Bland–Altman plots were used to assess the agreement between the AHI estimations (MLR, MLP, RBF, and the conventional approach) and the actual values of AHI. Furthermore, the diagnostic ability of these estimations was also evaluated. Thus, we used the AHI threshold established by the physicians (AHI = 10 e/h) to derive Se, Sp, Acc, PPV, and NPV in each case.

3.1 Feature selection stage

The FCBF algorithm was applied to P^c_{AF}, P^c_{RRV}, and P^c_{AF-RRV}. The complete patterns were significantly filtered. Thus, the reduced patterns P^r_{AF}, P^r_{RRV}, and P^r_{AF-RRV} were, respectively, composed of: 7 out of 19 AF features (from higher to lower SU: WD_b, M_{f1b}, ApEn, CTM, M_{f3b}, WD, M_{f1}), 5 out of 19 RRV features (from higher to lower SU: CTM, M_{f1b}, M_{t3}, M_{f3}, M_{f1}), and 10 out of 38 AF and RRV features (from higher to lower SU: CTM^{RRV}, WD_b^{AF}, M_{f1b}^{RRV}, M_{t3}^{RRV}, M_{f1b}^{AF}, M_{f3}^{RRV}, M_{f1}^{RRV}, $ApEn^{AF}$, CTM^{AF}, LZC^{RRV}). All the selected features from the spectral bands of interest were more relevant than the features from the full PSDs. Linear and nonlinear features, as well as

frequency and time domain features, were selected in all cases. The presence of AF and RRV features was balanced in P^{r}_{AF-RRV}. Nonetheless, the features from RRV tended to be more relevant than those from AF. CTM from RRV was the most relevant feature in terms of SU.

3.2 Design and training stages

3.2.1 Design of MLP and RBF

A proper design of MLP and RBF networks is required to achieve high generalization ability. It refers to selecting the appropriate model complexity in order to prevent over-fitting and under-fitting effects [7]. The effective complexity of the MLP and RBF models is governed by the design parameters [7]. Thus, we experimentally determined the number of hidden nodes (N_H and N_B), the regularization parameter (υ), and a smoothing parameter (τ), which governs the widths of kernel functions (σ_j) in RBF. Only the training group was used for this purpose.

Figures 3 and 4 show the results of the experiments conducted to determine these parameters. The MLP and RBF were fed with complete (P^{c}_{AF}, P^{c}_{RRV}, P^{c}_{AF-RRV}) and reduced (P^{r}_{AF}, P^{r}_{RRV}, P^{r}_{AF-RRV}) patterns. In each case, the *ICC* was computed for N_H/υ (MLP) or N_B/τ (RBF) pairs, and it was used as selection criterion. *ICC* was estimated through loo-cv, which was repeated ten times due to random initialization of weights and centres of MLP and RBF networks. Then, we averaged the ten *ICC*s to obtain the final value.

Figure 3a–f displays the performance of the MLP networks following this procedure. Figures in the same column correspond to complete (left) or reduced (right) input patterns, respectively. Figures in the same row indicate the origin of the features included in the patterns: AF, RRV, or both signals. υ was assessed according to each set. We chose those υ for which their *ICC* was higher throughout the number of nodes. N_H was varied from 1 to 50, and the optimum value was selected for the sake of the network complexity, i.e. we chose those values from which no substantial *ICC* improvement was observed. Thus, the optimum values were N_H/υ = 18/6 (P^{c}_{AF}), 20/11 (P^{c}_{RRV}), 22/8 (P^{c}_{AF-RRV}), 17/3 (P^{r}_{AF}), 13/7 (P^{r}_{RRV}), and 18/2 (P^{r}_{AF-RRV}). Since N_H/υ govern the effective complexity of the networks [7], less complex models were selected as optimum when using reduced patterns.

Figure 4 follows the same scheme for the RBF networks. We varied N_B from 1 to 50 and evaluated τ in 1, 2, 3, 4 and 5. Since the evolution of the *ICC* presented clear absolute maximums, we selected those pairs N_B/τ corresponding with these points. Hence, N_B/τ were the following: 21/2 (P^{c}_{AF}), 7/4 (P^{c}_{RRV}), 7/4 (P^{c}_{AF-RRV}), 18/3 (P^{r}_{AF}), 4/1 (P^{r}_{RRV}), and 5/4 (P^{r}_{AF-RRV}). The optimum models were also less complex in the case of reduced patterns, i.e. fewer nodes N_B were used.

3.2.2 Training of MLR, MLP and RBF models

Specific MLR, MLP and RBF models were obtained from the entire training group. A single MLR model was computed for each set of complete (P^{c}_{AF}, P^{c}_{RRV}, and P^{c}_{AF-RRV}) and reduced (P^{r}_{AF}, P^{r}_{RRV}, and P^{r}_{AF-RRV}) patterns. In the case of MLP and RBF, we computed 100 models for each set, due to random initializations in these networks. The optimum design parameters values, which were obtained in the previous stage, were used in the process.

3.3 Test stage

3.3.1 Intra-class correlation coefficient and Bland–Altman plots

Table 2 shows the *ICC* values reached by the MLR, MLP and RBF models for each set of patterns in the test group. The values for MLP and RBF are presented as mean ± standard deviation of the 100 models previously obtained. One model for each method was selected according to their *ICC*: MLR^{c}_{AF-RRV} (P^{c}_{AF-RRV} from MLR), MLP^{r}_{AF-RRV} (P^{r}_{AF-RRV} from MLP), and RBF^{r}_{AF} (P^{r}_{AF} for RBF). Thus, MLP^{r}_{AF-RRV} outperformed AHI_c in terms of agreement and both of them outperformed MLR^{c}_{AF-RRV} and RBF^{r}_{AF}. This tendency was also observed when applying graphical analysis. Figure 5 displays the "Bland–Altman"—(a, c, e, g)—and "estimated versus true AHI" plots—(b, d, f, h). Both graphs show smaller deviation from the target AHI in the case of MLP^{r}_{AF-RRV} and AHI_c. These models also reached less dispersion in the scatter of the points, which is reflected in the corresponding 95 % confidence interval: [−15.6, 19.9] e/h in the case of MLP^{r}_{AF-RRV} and [−16.6, 19.3] e/h for AHI_c.

3.3.2 Diagnostic performance of the models

To complete the analysis, we evaluated the diagnostic ability of the four AHI estimations obtained from the test group. Table 3 shows sensitivity (Se), specificity (Sp), accuracy (Acc), positive predictive value (PPV), and negative predictive value (NPV) for each method. The highest performance was achieved by MLP^{r}_{AF-RRV}, which reached 92.5 % Se, 89.5 % Sp, 91.5 % Acc, 94.9 % PPV, and 85.0 % NPP. MLR^{c}_{AF-RRV} and RBF^{r}_{AF} also outperformed AHI_c at each statistic.

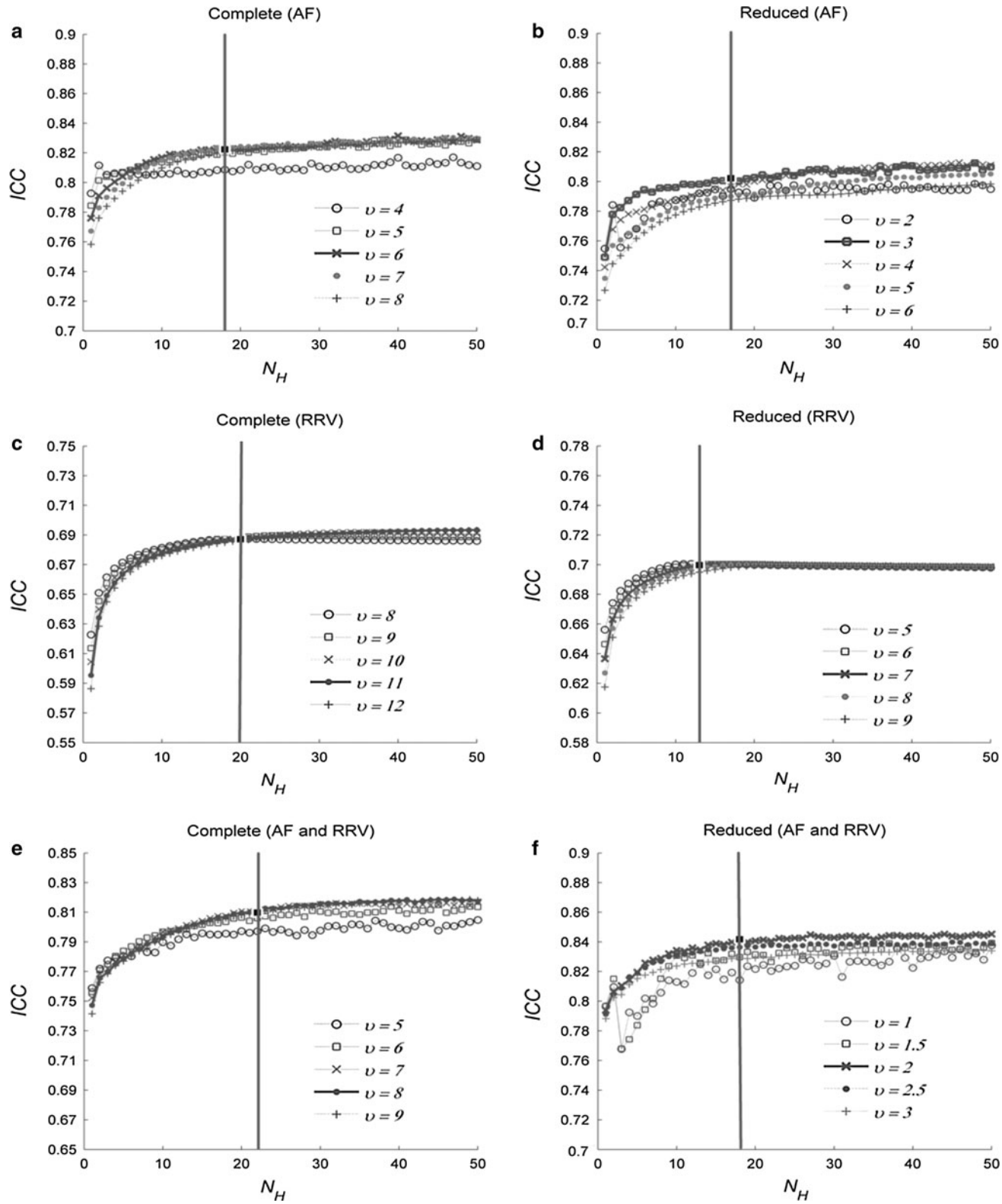

Fig. 3 MLP design stage: *ICC* for different N_H and υ values. Optimum values of: υ marked in *solid line*; N_H marked in *vertical line*

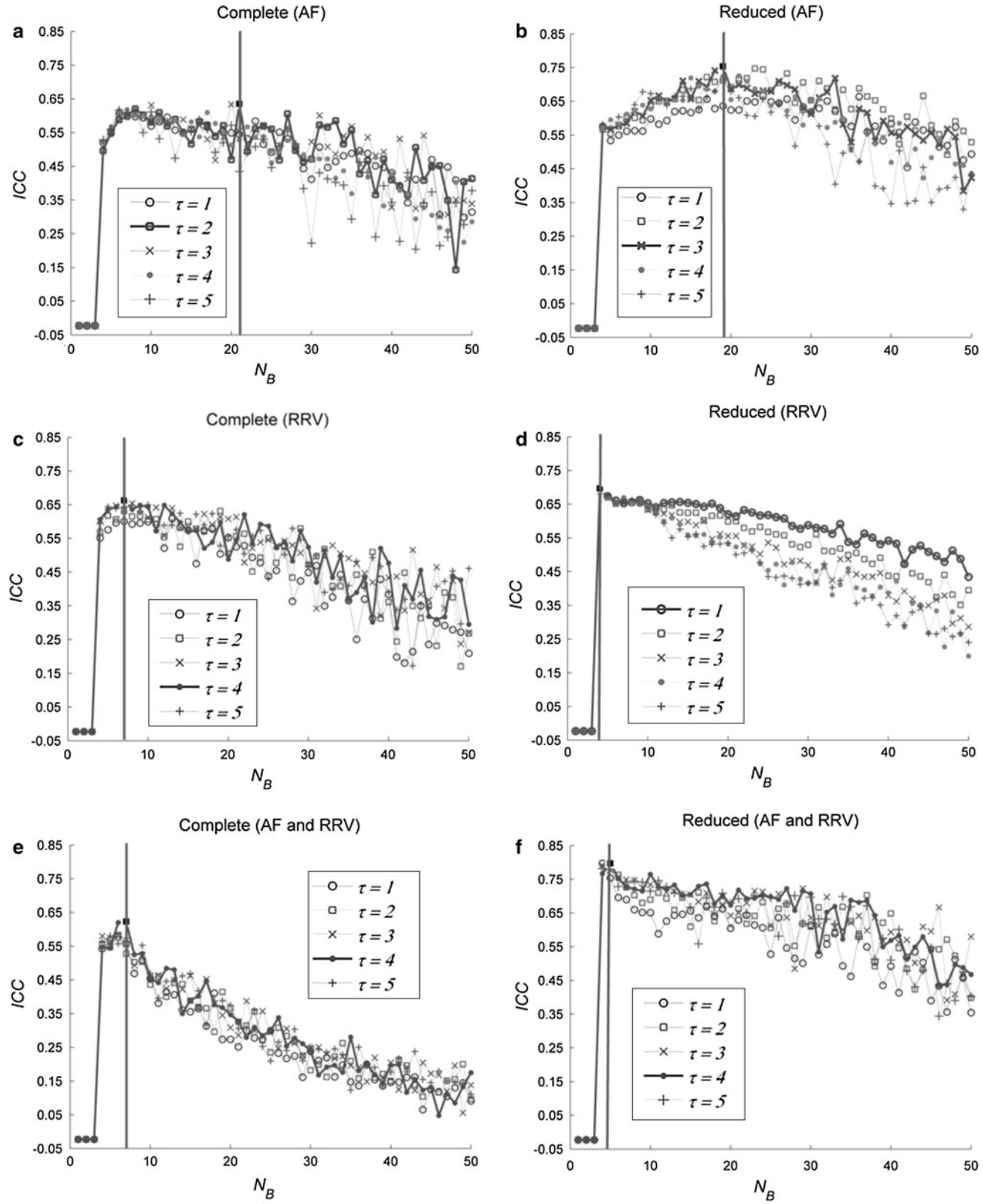

Fig. 4 RBF design stage: *ICC* for different N_B and τ values. Optimum values of: τ marked in *solid line*; N_B marked in *vertical line*

Table 2 *ICC* obtained from MLR, MLP, RBF, and the conventional approach (AHI_c)

	ICC test		
	AF	RRV	AF-RRV
AHI_c	**0.840**	–	–
MLR			
P^c	0.796	0.710	**0.809**
P^r	0.650	0.689	0.777
MLP			
P^c	0.782 ± 0.002	0.644 ± 4.3 e-4	0.808 ± 1.7-5
P^r	0.743 ± 0.002	0.685 ± 1.1 e-4	**0.849 ± 0.002**
RBF			
P^c	0.594 ± 0.094	0.617 ± 0.022	0.632 ± 0.170
P^r	**0.748 ± 0.037**	0.703 ± 0.006	0.732 ± 0.016

Best performance for each method in bold

P^c complete patterns, P^r reduced patterns

4 Discussion and conclusions

In this study, we addressed the estimation of AHI by pattern recognition in single-channel AF. Our approach focused on the exhaustive analysis of AF and RRV signals. Thus, spectral, nonlinear, and statistical features were obtained from all recordings. FCBF algorithm filtered these features, discarding those non-relevant or redundant. After filtering, both linear and nonlinear features from AF and RRV were selected. Moreover, all the features selected from the spectral bands of interest were more relevant in terms of SU than those selected from the full PSDs. The FCBF method was also useful in the design of MLP and RBF. Thereby, optimum less complex networks were selected in both cases when using reduced patterns (P^r_{AF}, P^r_{RRV}, and P^r_{AF-RRV}) instead of complete patterns (P^c_{AF}, P^c_{RRV}, and P^c_{AF-RRV}). These results support the use of the AF and RRV signals, as well as the methodology conducted to characterize them.

During the test stage, the agreement between the AHI estimations and the true AHI was evaluated. We selected specific models according to their *ICC*. Both *ICC* and graphical analysis supported MLP^r_{AF-RRV} and AHI_c as the best in terms of agreement. The conventional approach, however, systematically overestimated AHI in the SAHS-negative sample (15 out of 19 subjects) and underestimated AHI in the SAHS-positive sample (27 out of 40 subjects) (Fig. 5 b). These two effects may have caused that, despite having lower *ICC* values, MLR^c_{AF-RRV} and RBF^r_{AF} reached higher global diagnostic ability than AHI_c.

The diagnostic ability of the methods was also assessed. The highest performance was achieved by the AHI estimation derived from the MLP^r_{AF-RRV} model. This model reached high sensitivity (92.5 %), specificity (89.5 %), and accuracy (91.5 %). Only 2 out of 19 SAHS-negative subjects (false positives) and 3 out of 40 SAHS-positive subjects (false negatives) were misclassified. Additionally, three out of them have borderline true AHI values (5.7, 10, and 15.8 e/h). Thus, 94.9 % of subjects that our model estimated SAHS-positive were actually suffering from SAHS. Moreover, 85.0 % of subjects that our model predicted SAHS-negative were not SAHS patients. These findings confirmed the usefulness of combining relevant and non-redundant features from AF and RRV.

Fig. 5 Bland–Altman plots (**a**, **c**, **e**, **g**) and "estimated versus true AHI" (**b**, **d**, **f**, **h**), for the specific models and the conventional approach (AHI_c). Results derived from the test group. *TP* true positives, *FP* false positives, *TN* true negatives, *FN* false negatives ▶

Recent studies aimed at identifying SAHS (AHI threshold = 10 e/h) from single-channel AF. Most of them detected and scored respiratory events to estimate AHI. Shochat et al. [36] investigated the usefulness of SleepStrip™ for this purpose. They acquired AF through a thermistor and involved 288 subjects. Sensitivity was 86.0 %, but specificity reached low values (57.0 %). Nakano et al. [30] scored events supported by a spectral analysis of AF. The best performance was achieved using 116 AF recordings acquired with a thermocouple: 92 % Se and 90 % Sp. Their results are similar to ours from MLP^r_{AF-RRV}. Nonetheless, no further comparison was possible since no data were reported to obtain Acc, PPV or NPV. Nasal prong pressure sensor (NPP) has been widely used to acquire AF in portable diagnostic devices. Thus, De Almeida et al. assessed SleepCheck™ [11]. The authors reported 85.7 % Se and 87.5 % Sp by using a small sample size (30 subjects). Additionally, Wong et al. [39] evaluated FlowWizard™. They achieved high diagnostic performance: 92 % Se, 86 % Sp, 96 % PPV, and 75 % NPV. However, only 27 SAHS-positive subjects and 7 SAHS-negative subjects were used. Finally, ApenaLink™ was recently evaluated by BaHammam et al. [5]. The study involved 95 AF recordings. Specificity and PPV reached high values (89.0 and 91.0 %, respectively), but sensitivity (70.0 %) and NPV (63.0 %) were low. In contrast to the conventional approach conducted in these studies, our methodology took into account not only the apnoeic events but also data from the whole single-channel AF. A similar approach was performed in a recent study of our research group [15]. The utility of AF and RRV signals was assessed by the use of a logistic regression model, i.e. into a binary classification task. After a loo-cv process, the diagnostic performance reached 88 % Se, 70.8 % Sp, 82.4 % Acc, 86.3 % PPV, and 73.9 % NPV.

There also exist SAHS studies not aimed at assessing the diagnostic ability of a given methodology, but focused on evaluating how well this methodology detects apnoeas

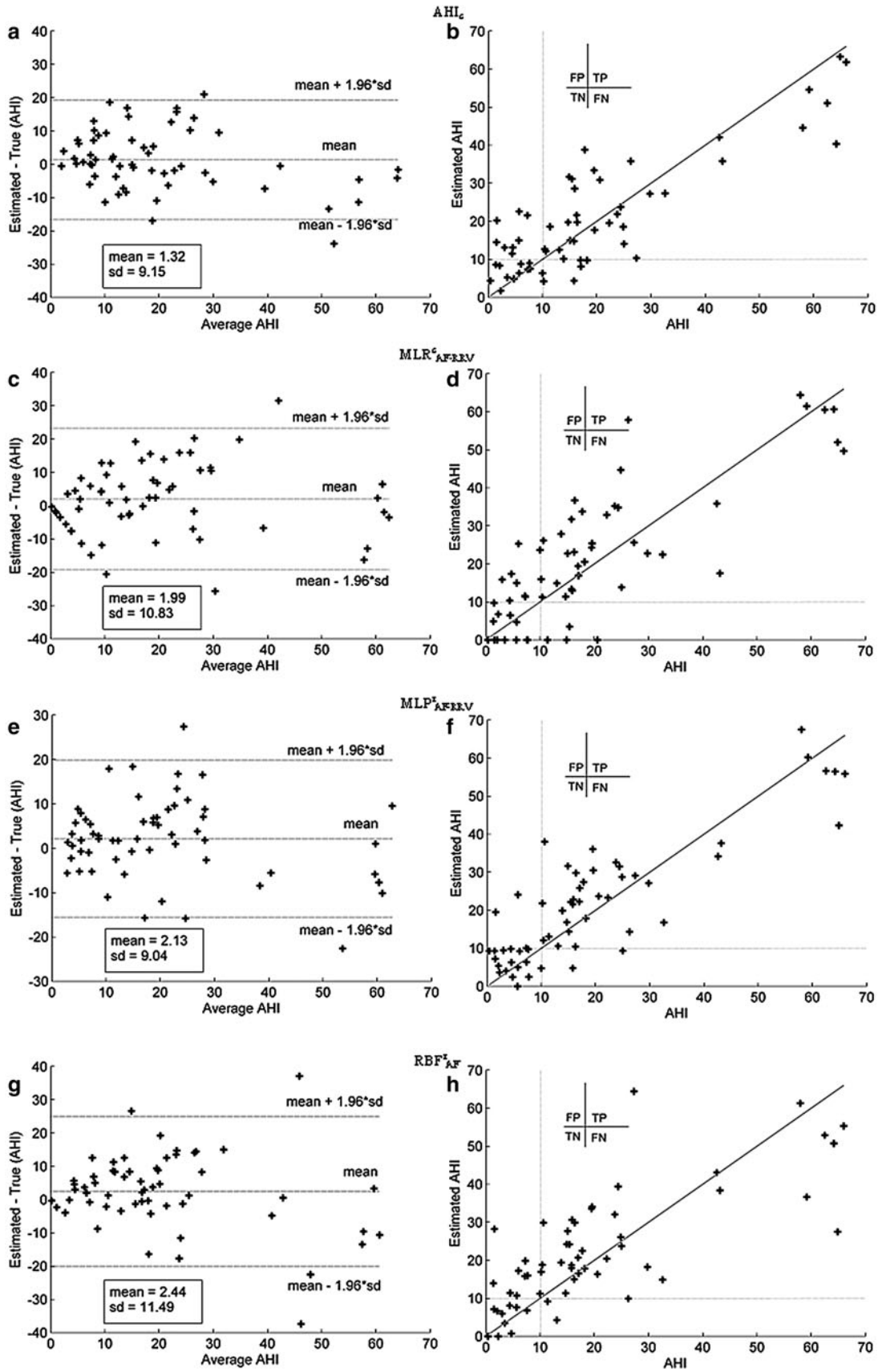
AHI_c
a
b
mean + 1.96*sd
mean
mean - 1.96*sd
mean = 1.32
sd = 9.15
Estimated - True (AHI)
Average AHI
FP TP
TN FN
Estimated AHI
AHI
MLR^c_AF-RRV
c
d
mean = 1.99
sd = 10.83
MLP^c_AF-RRV
e
f
mean = 2.13
sd = 9.04
RBF^c_AF
g
h
mean = 2.44
sd = 11.49

Table 3 Diagnostic performance of the specific models on the test group: comparison with the conventional approach (AHI_c)

	Se (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)
AHI_c	87.5	57.9	78.0	81.4	68.7
MLR^c_{AF-RRV}	90.0	63.2	81.4	83.7	75.0
MLP^r_{AF-RRV}	92.5	89.5	91.5	94.9	85.0
RBF^r_{AF}	92.5	57.9	81.4	82.2	78.6

Se sensitivity, *Sp* specificity, *Acc* accuracy, *PPV* positive predictive value, *NPP* negative predictive value

and/or hypopnoeas. Han et al. [16] used AF recordings from NPP, along with an automatic algorithm based on the mean magnitude of the second derivative, to detect apnoeas. They reported 92.4 % Se and 88.3 % Sp when comparing their methodology with the manual score of the events. Álvarez-Estévez and Moret-Bonillo [3] applied a fuzzy algorithm to AF, SpO_2, and respiratory movement recordings in order to detect respiratory events and classify them into apnoeas or hypopnoeas. Their results showed 87 % Se and 89 % Sp in the detection task, whereas they reported 92/85 % Se and 85/92 % Sp in the classification task (apnoeas/hypopnoeas). Otero et al. [31] propose several algorithms to detect different pathological events from polysomnographic recordings. Their results showed 97.4 and 94.0 % PPV when detecting apnoeas and hypopnoeas, respectively.

Pattern recognition techniques have been already shown to be useful in SAHS detection. Varady et al. [37] trained four feed-forward artificial neural networks to detect apnoeic segments in AF recordings. Data from AF and respiratory inductive plethysmography (RIP) were used. Up to 93 % of patterns were correctly classified into normal, apnoea, or hypopnoea categories. No assessment of diagnostic ability was performed. El-Shol et al. [12] trained a MLP network to predict AHI from demographic and clinical variables of subjects. Sensitivity and specificity reached 94.9 and 64.7 %, whereas PPV and NPV were 87.9 and 85.2 %, respectively. Additionally, in other study of our research group [26], 14 features extracted from 240 SpO_2 recordings were used along with MLR and MLP algorithms. The *ICCs* were 0.80 and 0.91, respectively. The MLP model showed the highest diagnostic performance: 89.6 % Se, 81.2 % Sp, 86.8 % Acc, 90.5 % PPV, and 79.6 % NPV.

Although our methods have revealed the usefulness of AF and RRV in SAHS detection, some limitations have to be addressed. A larger sample size would improve the generalization of our results. Accordingly, the validation of the proposed algorithms using different databases would be of great interest to enhance their statistical power [22]. Moreover, the use of subjects without previous suspects of suffering from SAHS would complement our findings. Nonetheless, this issue has no easy solution since subjects usually undergo overnight PSG after referring some symptoms. The cut-off AHI = 10 e/h is widely used to determine SAHS [5, 30, 36, 39]. Hence, our methodology was optimized according to this threshold. Future works, however, could assess our methodology for other common cut-offs such as 5 or 15 e/h. Another limitation is the use of a thermistor, instead of a thermistor and a NPP simultaneously. The AASM recommends using both sensors to acquire AF [19], due to weaknesses in the two of them [4]. Additionally, it is well known that NPP outperforms thermistor when recording respiratory events [4]. However, this work has shown that a global analysis of single-channel AF from thermistor can achieve high diagnostic performance and improve the results reported in recent studies only involving NPP [5, 11, 39]. The application of our methodology to AF recordings from NPP is a future goal. Another future goal is to assess relationships between the proposed features and the apnoeic events in order to clarify their physiological meaning. Additionally, our methodology does not offer flexibility to the physicians in order to change the AHI based on their expertise. However, the results reported in this study measure to what extent physicians can trust our AHI estimations. Finally, the main benefit of our approach would be obtained by applying our algorithms to single-channel AF recordings acquired at patient's domicile. Although there exist several portable devices to obtain AF [5, 11, 36, 39], these have limitations and need further investigation to ensure their reliability in unattended studies at home.

In summary, single-channel AF from thermistor can be used to assist in SAHS detection and simplify diagnosis. The methodology conducted over AF and RRV signals has shown its usefulness to estimate AHI. Particularly, the FCBF algorithm was successfully used to discard redundant and non-relevant information from recordings, which in turn decreased the complexity of the models obtained through neural networks. An MLP model, trained with relevant and non-redundant features from AF and RRV, achieved high results in terms of agreement with true AHI and diagnostic ability. It outperformed a conventional approach, based on scoring apnoeas and hypopnoeas, conducted over the same database. Additionally, the MLP approach also improved the diagnostic ability of the conventional one conducted in other studies. Our results suggest that AF and RRV complement each other in the AHI estimation and can help in SAHS diagnosis.

Acknowledgments This research was supported in part by the "Consejería de Educación (Junta de Castilla y León)" under project VA111A11-2, the Project Cero 2011 on Ageing from Fundación General CSIC, and project TEC2011-22987 from Ministerio de Economía y Competitividad and FEDER. G. C. Gutiérrez-Tobal 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. Aarabi A, Wallois F, Grebe R (2006) Automated neonatal seizure detection: a multistage classification system through feature selection based on relevancy and redundancy analysis. Clin Neurophysiol 117:328–340
2. Álvarez D, Hornero R, Marcos JV, del Campo F (2010) Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. IEEE Trans Biomed Eng 57:2816–2824
3. Álvarez-Estévez D, Moret-Bonillo V (2009) Fuzzy reasoning used to detect apneic events in the sleep apnea-hypopnea syndrome. Expert Syst Appl 36:7778–7785
4. BaHammam A (2004) Comparison of nasal prong pressure and thermistors measurements for detecting respiratory events during sleep. Respiration 71:385–390
5. BaHammam A, Sharif M, Gacuan DE, George S (2011) Evaluation of the accuracy of manual and automatic scoring of a single airflow channel in patients with a high probability of obstructive sleep apnea. Med Sci Monit 17:MT13–MT19
6. Bennet JA, Kinnear WJM (1999) Sleep on the cheap: the role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome. Thorax 54:958–959
7. Bishop CM (1996) Neural networks for pattern recognition. Oxford University Press, Oxford, UK
8. Campos-Rodríguez F, Martínez-García MA, Martínez M, Duran-Cantolla J, de la Peña M, Masdeu MJ, González M, del Campo F, Gallego I, Martín JM, Barbe F, Monstserrat JM, Farre R (2013) Association between obstructive sleep apnea and cancer incidence in a large multicenter Spanish cohort. Am J Respir Crit Care Med 187:99–105
9. Cohen ME, Hudson DL, Deedwania PC (1996) Applying continuous chaotic modelling to cardiac signal analysis. IEEE Eng Med Biol Mag 15:97–102
10. Cysarz D, Zerm R, Bettermann H, Frühwirth M, Moser M, Kröz M (2008) Comparison of respiratory rates derived from heart rate variability, ECG amplitude, and nasal/oral airflow. Ann Biomed Eng 36:2085–2094
11. De Almeida FR, Ayas NT, Otsuka R, Ueda H, Hamilton P, Ryan FC, Lowe AA (2006) Nasal pressure recordings to detect obstructive sleep apnea. Sleep Breath 10:62–69
12. El-Solh AA, Mador MJ, Ten-Brock E, Shucard DW, Abul-Khoudoud M, Grant BJB (1999) Validity of neural network in sleep apnea. Sleep 22:105–111
13. Fernández-Navarro F, Hervás-Martínez C, Ruiz R, Riquelme JC (2012) Evolutionary generalized radial basis function neural networks for improving prediction accuracy in gene classification using feature selection. Appl Soft Comput 12:1787–1800
14. Flemons WW, Littner MR, Rowley JA, Gay P, Anderson WM, Hudgel DW, McEvoy RD, Loube DI (2003) Home diagnosis of sleep apnea: a systematic review of the literature. Chest 124:1543–1579
15. Gutiérrez-Tobal GC, Hornero R, Álvarez D, Marcos JV, del Campo F (2012) Linear and nonlinear analysis of airflow recordings to help in sleep apnoea-hypopnoea syndrome diagnosis. Physiol Meas 33:1261–1275
16. Han J, Shin HB, Jeong DU, Park KS (2008) Detection of apnoeic events from single channel nasal airflow using 2nd derivative method. Comput Methods Progr Biomed 98:199–207
17. Hornero R, Alonso A, Jimeno N, Jimeno A, López M (1999) Nonlinear analysis of time series generated by schizophrenic patients. IEEE Eng Med Biol Mag 3:84–90
18. Hu Q, Pan W, An S, Ma P, Wei J (2010) An efficient gene selection technique for cancer recognition based on neighborhood mutual information. Int J Mach Learn Cybern 2:63–74
19. Iber C, Ancoli-Israel S, Chesson A, Quan SF (2007) The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. American Academy of Sleep Medicine, Westchester, IL
20. Jobson JD (1991) Applied multivariate data analysis. Regression and experimental design, vol I. Springer, New York
21. Korten JB, Haddad GG (1989) Respiratory waveform pattern recognition using digital techniques. Comput Biol Med 19:207–217
22. Lado MJ, Vila XA, Rodríguez-Liñares L, Méndez AJ, Oliveri DN, Félix P (2011) Detecting sleep apnea by heart rate variability analysis: assessing the validity of databases and algorithms. J Med Syst 35:473–481
23. Lempel A, Ziv J (1976) On the complexity of finite sequences. IEEE Trans Inform Theory 24:530–536
24. Lindberg E, Carter N, Gislason T, Janson C (2001) Role of snoring and daytime sleepiness in occupational accidents. Am J Respir Crit Care Med 164:2031–2035
25. López-Jiménez F, Kuniyoshi FHS, Gami A, Somers VK (2008) Obstructive sleep apnea: implications for cardiac and vascular disease. Chest 133:793–804
26. Marcos JV, Hornero R, Álvarez D, Aboy M, del Campo F (2012) Automated prediction of the apnea-hypopnea index from nocturnal oximetry recordings. IEEE Trans Biomed Eng 59:141–149
27. Martín MT, Plastino A, Rosso OA (2003) Statistical complexity and disequilibrium. Phys Lett A 311:126–132
28. Nabney IT (2002) NETLAB: algorithms for pattern recognition. Springer, Berlin
29. Nagarajan R (2002) Quantifying physiological data with Lempel-Ziv complexity—certain issues. IEEE Trans Biomed Eng 49:1371–1373
30. Nakano H, Tanigawa T, Furukawa T, Nishina S (2007) Automatic detection of sleep-disordered breathing from single-channel airflow record. Eur Respir J 29:728–736
31. Otero A, Félix P, Álvarez MR (2011) Algorithms for the analysis of polysomnographic recordings with customizable criteria. Expert Syst Appl 38:10133–10146
32. Patil SP, Schneider H, Schwartz AR, Smith PL (2007) Adult obstructive sleep apnea: pathophysiology and diagnosis. Chest 132:325–337
33. Pincus SM (1991) Approximate entropy as a measure of system complexity. Proc Natl Acad Sci 88:2297–2301
34. Pincus SM (2001) Assessing serial irregularity and its implications for health. Ann N Y Acad Sci 954:245–267
35. Sassani A, Findley LJ, Kryger M, Goldlust E, George C, Davidson TM (2004) Reducing motor-vehicle collisions, cost, and fatalities by treating obstructive sleep apnea syndrome. Sleep 27:453–458
36. Shochat T, Hadas N, Kerkhofs M, Herchuelz A, Penzel T, Peter JH, Lavie P (2002) The SleepStripTM: an apnoea screener for the early detection of sleep apnoea syndrome. Eur Respir J 19:121–126
37. Várady P, Micsik T, Benedek S, Benyó Z (2002) A novel method for the detection of apnea and hypopnea events in respiration signals. IEEE Trans Biomed Eng 49:936–942
38. Welch PD (1967) The use of fast Fourier transform of the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust AU-15:70–73
39. Wong KKH, Jankelson D, Reid A, Unger G, Dungan G, Hedner JA, Grunstein RR (2008) Diagnostic test evaluation of a nasal flow monitor for obstructive sleep apnea detection in sleep apnea research. Behav Res Methods 40:360–366

40. Wootters WK (1981) Statistical distance and Hilbert space. Phys Rev D 23:357–362
41. Young T, Peppard PE, Gottlieb DJ (2002) Epidemiology of obstructive sleep apnea. Am J Respir Crit Care 165:1217–1239
42. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224
43. Zhang XS, Roy RJ, Jensen EW (2001) EEG complexity as a measure of depth anesthesia for patients. IEEE Trans Biomed Eng 48:1424–1433