Classification of oximetry signals using Bayesian neural networks to assist in the detection of the obstructive sleep apnoea syndrome

JV Marcos[1], R Hornero[1], D Álvarez[1], IT Nabney[2], F del Campo[3], C Zamarrón[4]

[1] Biomedical Engineering Group, E.T.S.I. de Telecomunicación, University of Valladolid, Camino del Cementerio s/n, 47011, Valladolid, Spain

[2] N~~eural Computing~~on-linearity and Complexity Research Group, School of Engineering and Applied Sciences, Aston University, Aston Triangle, B4 7ET, Birmingham, United Kingdom

[3] Hospital Universitario del Río Hortega, Servicio de Neumología, c/ Dulzaina 2, 47012, Valladolid, Spain

[4] Hospital Clínico Universitario, Servicio de Neumología, Travesía de la Choupana s/n, 15706, Santiago de Compostela, Spain

E-mail: jvmarcos@gmail.com, robhor@tel.uva.es, dalvgon@ribera.tel.uva.es, i.t.nabney@aston.ac.uk, fsas@telefonica.net, carlos.zamarron.sanz@sergas.es

Abstract

Nowadays, polysomnography (PSG) is the gold-standard to diagnose obstructive sleep apnoea syndrome (OSAS). However, it is complex, time-consuming and expensive. Nocturnal pulse oximetry, which provides oxygen saturation (SaO_2) recordings, allows us to overcome these difficulties~~ and could be an alternative to PSG~~. In the present study, multilayer perceptron (MLP) neural networks were applied to help in OSAS diagnosis using information from SaO_2 signals. We performed time and spectral analysis of these recordings to extract 14 features related to OSAS. ~~According to the principle used for network optimisation, w~~We compared the performance of two different MLP classifiers: maximum likelihood (ML) and Bayesian (BY) MLP networks. A total of 187 subjects suspected of suffering from OSAS took part in the study. Their SaO_2 signals were divided into a training set with 74 recordings and a test set with 113 recordings~~ to develop and validate the classifiers~~. BY-MLP networks achieved the best performance on the test set with 85.58% accuracy (87.76% sensitivity and 82.39% specificity). These results were substantially better than those provided by ML-MLP networks, which were affected by overfitting and achieved an accuracy of 76.81% (86.42% sensitivity and 62.83% specificity). Our results suggest that the Bayesian framework is preferred to implement our MLP classifiers. The proposed BY-MLP networks could be used for early OSAS detection~~, contributing to~~ and thus reduce the number of required PSGs.

Keywords: obstructive sleep apnoea syndrome (OSAS), nocturnal pulse oximetry, multilayer perceptron (MLP), maximum likelihood, Bayesian inference

1 Introduction

Feedforward neural networks represent a powerful tool for pattern recognition tasks. Several advantages ~~can be found on~~are associated with these algorithms. Firstly, neural networks can learn from the environment in which they operate without the requirement of any previous assumption (Haykin 1996, Zhang 2000). In addition, neural networks are capable of universal approximation, i.e. they can approximate any continuous mapping, provided that sufficiently many hidden units are available (Hornik 1991). Finally, neural networks are nonlinear models. Thus, they can be applied to model real-world complex relationships (Zhang 2000).

The multilayer perceptron (MLP) is the most commonly studied and used feedforward neural network. It is usually used for classification purposes since MLP networks can estimate posterior probabilities (Bishop 1995, Haykin 1999): see equation (6). ~~These neural networks have provided promising results in a wide variety of pattern recognition problems such as handwritten digit recognition, fault diagnosis, bankruptcity prediction and medical diagnosis (Zhang 2000).~~ Traditionally, MLP networks were based on the maximum likelihood (ML) approach ~~were applied to solve these problems~~. According to this principle, network weights ~~(~~**w** (which are the parameters of the model in a statistical sense)~~)~~ are determined by minimising an objective function that represents the error between the desired output and the network output. As a result, a single set of weights is obtained as optimum. In this context, more complex models are typically better able to fit the training data. However, this does not ~~involve~~ necessarily imply good generalisation capability (Bishop 1995). The Bayesian (BY) framework provides an alternative approach ~~to implement MLP algorithms. T~~which accounts for the uncertainty ~~on~~ in the network weights ~~is taken into account~~ by considering a prior probability distribution function $p(\mathbf{w})$ over weight space. The prior assumption is that networks with small weights are preferred: these give rise to smoother mappings and hence the network is likely to overfit. Once the training data D is observed, the posterior probability $p(\mathbf{w}|D)$ can be obtained (Bishop 1995). This distribution ~~is concentrated on weight values that are more consistent with data in the training set~~can be used to integrate over all possible parameters, weighted by their probability. Moreover, it can be used to evaluate the relevance of each input variable for the network predictions (Neal 1996, Nabney 2002).

The aim of this study is to analyse the performance of Bayesian MLP networks to assist in the diagnosis of the obstructive sleep apnoea syndrome (OSAS). ~~Moreover, we compared these algorithms with maximum likelihood MLP networks. Both network models were applied in OSAS detection using information from nocturnal oxygen saturation signals.~~ Patients affected by OSAS suffer repetitive occlusion of the upper airway during sleep, leading to a complete (apnoea) or partial (hypopnoea) cessation of the airflow (Qureshi and Ballard 2003). The recurrence of these episodes has severe implications ~~on~~ for the health of the patient~~. Indeed~~: for example, OSAS is considered a risk factor for the development of cardiovascular diseases such as hypertension, cardiac failure, arrhythmias and atherosclerosis (Lattimore *et al* 2003). Additionally, excessive daytime sleepiness due to sleep fragmentation has been pointed out as a major cause of traffic and industrial accidents (George 2001). As a result, early diagnosis of OSAS is required in order to apply an effective treatment. ~~Nowadays~~Currently, nocturnal polysomnography (PSG) is the gold standard ~~in~~ for OSAS diagnosis (Qureshi and Ballard 2003). This test is performed in a special sleep unit and must be supervised by a trained technician. Usually, the following recordings are monitored during PSG: electroencephalogram (EEG), electrocardiogram (ECG), electromy~~i~~ogram (EMG), electro-oculogram (EOG), oximetry, nasal airflow and respiratory effort. Subsequently, a medical expert must analyse ~~these~~ this data to provide a final diagnosis. Despite its diagnostic reliability, PSG is complex, time-consuming and expensive (Bennet and Kinnear 1999). Therefore, simplified diagnostic techniques would be of great practical interest.

Comment [n1]: Can these be quantified here?

Arterial oxygen saturation (SaO_2) recordings from nocturnal pulse oximetry represent an alternative to PSG. These could be acquired in the home of the patient, resulting in reduced complexity and cost. Pulse oximetry is widely known in pulmonary medicine (Netzer *et al* 2001). It provides useful information about respiratory dynamics during sleep. Subjects suffering from OSAS are usually characterised by SaO_2 signals with high instability. The

repetition of apnoeas is reflected by frequent drops and the corresponding restorations of the saturation value due to the lack of oxygen during each apnoeic event. In contrast, control subjects tend to present a constant SaO_2 value around 97% (Netzer *et al* 2001). Thus, oximetry data ~~could be useful~~is useful for~~to~~ detecting OSAS.

Different methodologies have been previously proposed to perform OSAS diagnosis from SaO_2 recordings. Visual inspection represents the easiest analysis technique (Rodríguez *et al* 1996). However, automated analysis would ~~allow to~~ reduce the time required for a final diagnosis. Conventional oximetry indices were suggested for this purpose. These indices are usually provided by the oximetry equipment. They include the oxygen desaturation index over 2% (ODI2), 3% (ODI3) and 4% (ODI4), and the cumulative time spent below a given level of saturation. Typically, a saturation level of 90% (CT90) is applied (Lévy *et al* 1996, Roche *et al* 2002, Vázquez *et al* 2000, Netzer *et al* 2001, Magalang *et al* 2003). Recently, signal processing techniques have been also used for automated analysis of oximetry recordings through spectral and nonlinear methods (Zamarrón *et al* 2003, Álvarez *et al* 2006, Del Campo *et al* 2006, Álvarez *et al* 2007, Hornero *et al* 2007). ~~Moreover~~In addition, pattern classification techniques such as neural networks have been applied for the development of diagnostic algorithms (El-Solh *et al* 2003, Marcos *et al* 2008, Polat *et al* 2008).

In the present study, we modelled the OSAS diagnosis problem as a pattern recognition task. Subjects must be assigned to one of two possible groups: OSAS positive or negative. Several features extracted from SaO_2 recordings were fed into MLP network classifiers to identify subjects with OSAS. We evaluated and compared the capability of MLP networks developed within two different frameworks: the maximum likelihood and the Bayesian approaches.

2 Subjects and signals

A total of 187 SaO_2 recordings from subjects suspected of suffering from OSAS were available for the study. Usually, sleep analysis was carried out from midnight to 8:00 AM in the Sleep Unit of the Hospital Clínico Universitario of Santiago de Compostela (Spain). The Review Board on Human Studies at this institution approved the protocol. Conventional PSG and nocturnal pulse oximetry were simultaneously performed on each of the subjects. Oximetry signals were recorded by means of a Criticare 504 oximeter (CSI, Waukeska, U.S.A.) at a sampling frequency of 0.2 Hz. The equipment used to perform PSG was a polygraph (Ultrasom Network, Nicolet, Madison, W.I., U.S.A.). Signals obtained in the polysomnographic study were EEG, EOG, chin EMG, airflow (three-port thermistor), ECG and measurement of chest wall movement. These recordings were analysed by an expert according to the system by Rechtschaffen and Kales to obtain a diagnosis for each subject (Rechtschaffen and Kales 1968). Apnoea was defined as a cessation of airflow for 10 seconds or longer. Hypopnoea was defined as a reduction, without complete cessation, in airflow of at least 50%, accompanied by a decrease of more than 4% in the saturation of haemoglobin. The average apnoea-hypopnoea index (AHI) was calculated for hourly periods of sleep from apnoeic/hypopnoeic episodes captured in PSG. Finally, a threshold of AHI $\geq$ 10 events/h was established to determine the presence of OSAS in a subject.

A positive diagnosis of OSAS was confirmed in 111 subjects, i.e. 59.36% of the population under study. There were no significant differences between ~~OSAS~~ OSAS-positive and negative groups in age, body mass index (BMI) and recording time. On the other hand, the percentage of males was higher in the ~~OSAS~~ OSAS-positive group (84.68%) than in the ~~OSAS~~ OSAS-negative (69.74%). The initial population was randomly divided into training and test sets to develop both types of neural network classifiers. The proportion of ~~OSAS~~ OSAS-positive and negative subjects was preserved in each of these sets. The training set (74 subjects) was used for network optimisation. The test set (113 subjects) was applied to estimate the generalisation capability of the classifiers. Table 1 summarises the demographic and clinical data for the whole population as well as for training and test sets.

Comment [n2]: What was the methodology used? Expert judgment?

Comment [n3]: When 'OSAS positive' is used as an adjective, it should be hyphenated. I have tried to fix all of these, but may have missed a few.

Table 1. Demographic and clinical statistics of all subjects, training set and test set. Data are presented as mean ± standard deviation. *n*: number of subjects; BMI: body mass index; AHI: apnoea/hypopnoea index computed as events for hourly periods.

All subjects			
	All (*n* = 187)	OSAS Positive (*n* = 111)	OSAS Negative (*n* = 76)
Age (years)	57.97 ± 12.84	58.30 ± 12.88	57.57 ± 12.87
Males (%)	78.61	84.68	69.74
BMI (kg/m^2)	29.54 ± 5.51	30.45 ± 4.92	28.42 ± 6.02
Recording Time (h)	8.19 ± 0.62	8.17 ± 0.75	8.22 ± 0.33
AHI (events/h)		40.07 ± 19.64	2.04 ± 2.36
Training set			
	All (*n* = 74)	OSAS Positive (*n* = 44)	OSAS Negative (*n* = 30)
Age (years)	58.25 ± 12.14	56.73 ± 13.61	59.59 ± 10.19
Males (%)	75.68	79.55	70.00
BMI (kg/m^2)	29.62 ± 5.71	30.19 ± 5.09	28.93 ± 6.40
Recording Time (h)	8.22 ± 0.41	8.20 ± 0.49	8.25 ± 0.27
AHI (events/h)		38.11 ± 18.18	2.60 ± 2.51
Test set			
	All (*n* = 113)	OSAS Positive (*n* = 67)	OSAS Negative (*n* = 46)
Age (years)	57.91 ± 13.39	59.37 ± 12.38	56.03 ± 14.54
Males (%)	80.53	88.06	69.57
BMI (kg/m^2)	29.49 ± 5.41	30.63 ± 4.84	28.07 ± 5.80
Recording Time (h)	8.17 ± 0.72	8.14 ± 0.88	8.20 ± 0.37
AHI (events/h)		41.36 ± 20.58	1.67 ± 2.21

3 Methods

Pattern recognition techniques were applied to model the OSAS diagnosis problem using oxymetry data. Our methodology involved several stages: 1) feature extraction, 2) feature preprocessing and 3) pattern classification.

3.1 Feature extraction

The purpose of the feature extraction stage was to summarise the information in SaO_2 recordings using a reduced set of parameters or features. Events of apnoea are accompanied by hypoxaemia, which is reflected in oximetry signals with a marked decrease in the saturation value. These recordings tend to present a different behaviour for OSAS positive and negative subjects. Therefore, they can be useful for OSAS detection.

The extracted features must provide suitable measures in order to differentiate signals from both populations. We used signal processing techniques to extract features from oximetry data. Initially, conventional statistical analysis was applied. Moreover, we analysed SaO_2 signals using spectral and nonlinear methods. As stated in our previous research (Zamarrón *et al* 2003, Álvarez *et al* 2006, Del Campo *et al* 2006, Álvarez *et al* 2007, Hornero *et al* 2007), spectral and nonlinear features from oximetry signals provided significant statistical differences between OSAS positive and negative subjects. Finally, a total of 14 features were computed from SaO_2 recordings. These can be divided into two groups: time-domain and frequency-domain features.

3.1.1 Time-domain analysis of oximetry data

Time representation of oximetry signals reflects respiratory dynamics during sleep. Apnoea events are characterised by a decrease in the SaO_2 value due to airway obstruction and reduced airflow. Therefore, signals from positive subjects are usually associated to instability. They reflect continuous drops and subsequent restorations of the saturation value because of the repetition of apnoeas. In contrast, signals from OSAS negative subjects tend to present a near

constant saturation value around 97% (Netzer *et al* 2001). To illustrate this, the oximetry recordings corresponding to a control subject (AHI = 0 events/h), a patient suffering from OSAS (AHI = 44 events/h) and an uncertain OSAS positive patient (AHI = 14 events/h) from our database are depicted in figure 1. In order to quantify these dynamic differences, we analysed SaO_2 signals in the time domain using conventional statistics and nonlinear methods.

We estimated the first four standard moments for the distribution of the variable representing the SaO_2 value. The probability density function of this variable was modelled with a discrete uniform distribution. Each standard moment was estimated by averaging the values computed from signal epochs of 200 samples. The following features were extracted:

- Feature 1. First statistical moment in the time domain (SMT1). SMT1 represents the expected value ~~for~~ of the distribution of SaO_2 samples. It is ~~supposed to be~~usually lower for positive patients due to frequent drops in the saturation value.

- Feature 2. Second statistical moment in the time domain (SMT2). SMT2 represents the variance ~~for~~ of the distribution of SaO_2 samples. Positive patients are expected to provide higher values of SMT2 due to instability of their recordings.

- Feature 3. Third statistical moment in the time domain (SMT3). SMT3 measures the asymmetry ~~for~~ of the distribution of SaO_2 samples. Typically, it is negative for subjects from both populations. However, its magnitude is usually greater in positive subjects due to the higher concentration of samples with low values of saturation.

- Feature 4. Fourth statistical moment in the time domain (SMT4). SMT4 evaluates the sharpness ~~for~~ of the distribution of SaO_2 samples. It is expected to be higher in control subjects since their oximetry signals tend to be constant.

In addition, we analysed oximetry recordings with three different nonlinear methods: approximate entropy (ApEn), central tendency measure (CTM) and Lempel-Ziv complexity

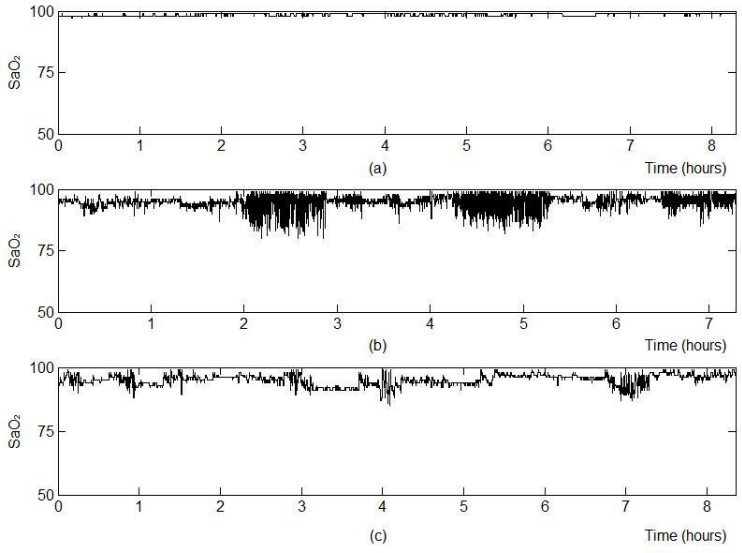

Figure 1. Time representation of oximetry recordings corresponding to (a) an OSAS negative subject (AHI = 0 events/h), (b) an OSAS positive subject (AHI = 44 events/h) and (c) an uncertain OSAS positive subject (AHI = 14 events/h).

(LZC). According to our previous research, these methods can be used to evaluate properties of oximetry signals that are related to OSAS (Álvarez *et al* 2006, Del Campo *et al* 2006, Álvarez *et al* 2007, Hornero *et al* 2007). Signals were divided into epochs of 200 samples to compute these features. For each of them, the average value from all signal epochs represented the final estimate. The following nonlinear methods were applied on SaO_2 signals:

- Feature 5. Approximate entropy (ApEn). ApEn estimates irregularity of time series, with high values of ApEn corresponding to irregular signals (Pincus 2001). This method requires ~~to specify~~the specification of two design parameters: a run length *m* and a tolerance window *r*. These were set to 1 and 0.25 times the standard deviation of the original sequence, respectively (Hornero *et al* 2007). ApEn measures the logarithmic likelihood that runs of patterns that are close (within *r*) for *m* contiguous observations remain close (within the same tolerance window width *r*) on subsequent incremental comparisons (Pincus 2001). Usually, high values of ApEn are associated with SaO_2 signals from OSAS positive subjects. These tend to present a more irregular behaviour due to frequent changes in the saturation value (Hornero *et al* 2007).

- Feature 6. Central tendency measure (CTM). CTM quantifies the variability of the signal (Cohen *et al* 1996), assigning low values to signals with a high degree of chaos. It is computed from second-order difference plots representing $(s_{t+2} - s_{t+1})$ vs. $(s_{t+1} - s_t)$, where $\mathbf{s} = (s_1,\ldots, s_t,\ldots, s_T)$ is ~~the~~ a time series of length *T*. The number of points that fall inside a circular region of radius ρ centred on the origin is obtained. Then, it is divided by the total number of points to compute CTM. In the present study, a radius $\rho = 0.25$ was applied to estimate CTM from our SaO_2 signals (Álvarez *et al* 2006). Oximetry recordings from OSAS positive patients are characterised by high variability due to the recurrence of apnoeas, which is reflected in low CTM values for these subjects (Álvarez *et al* 2006).

- Feature 7. Lempel-Ziv complexity (LZC). LZC estimates the complexity of the signal (Lempel and Ziv 1976). Series with high complexity provide high values of LZC. It is related to the number of distinct substrings and the rate of their recurrence along a given sequence. The signal is transformed into a finite symbol sequence by comparing each sample with a fixed threshold. In this study, LZC was computed by converting SaO_2 signals into 0-1 sequences. The median value of the signal samples was used as threshold (Álvarez *et al* 2006). The resulting sequence is scanned from left to right, increasing the complexity counter by one unit every time a new subsequence of consecutive characters is encountered (Lempel and Ziv 1976). High values of LZC are expected for SaO_2 signals from OSAS positive subjects (Álvarez *et al* 2006).

3.1.2 Frequency-domain analysis of oximetry data

Our preceding studies concluded that spectral analysis of oximetry signals reflects significant differences between positive and negative subjects. The signal power associated with frequency components located in the band between 0.010 and 0.033 Hz is usually higher in subjects with OSAS than in normal controls (Zamarrón *et al* 2003). The duration of an apnoea usually ranges from 30 s to 2 min, including the awakening response after the event. The repetition of apnoeas during sleep originates phase-lagged changes in SaO_2 signals with the same periodicity. Thus, the minimum and maximum frequencies for the occurrence of apnoeas would approximately correspond to the limits of that band. This behaviour is illustrated in figure 2. It depicts the power spectral density (PSD) function computed from the SaO_2 recordings shown in figure 1. As it can be observed, frequent changes in the saturation value due to apnoeas lead to a higher bandwidth in SaO_2 signals from positive subjects. In addition, a peak in the band between 0.010 and 0.033 Hz reflects a periodic component in the oximetry recording from the patient with severe OSAS.

Different methods can be used to compute the PSD from non-stationary data such as our SaO_2 signals. We applied the non-parametric Welch's method to estimate the PSD using a

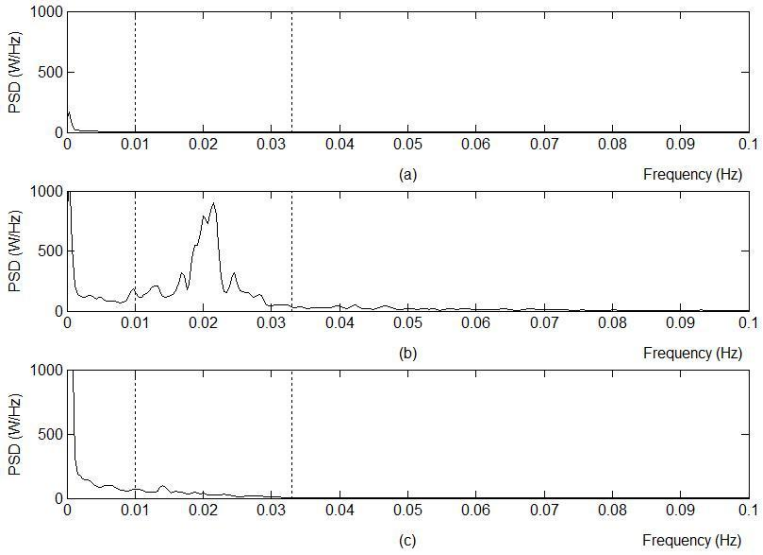

Figure 2. Power spectral density (PSD) computed from oximetry recordings corresponding to (a) an OSAS negative subject (AHI = 0 events/h), (b) an OSAS positive subject (AHI = 44 events/h) and (c) an uncertain OSAS positive subject (AHI = 14 events/h). The dashed line indicates the limits for the band between 0.010 and 0.033 Hz.

Hanning window with a length of 300 samples (50% overlapping) (Welch 1967). Initially, we analysed the statistical properties of the variable representing the frequency component of oximetry signals. The normalised PSD was used as the probability density function of this variable. The following features were computed:

- Feature 8. First statistical moment in the frequency domain (SMF1). SMF1 estimates the expected value of the variable defined by the frequency component. It is ~~expected to be~~usually ~~higher~~ larger for OSA ~~S~~ positive patients since their SaO_2 signals tends to ~~present~~ have a ~~higher~~ larger bandwidth.

- Feature 9. Second statistical moment in the frequency domain (SMF2). SMF2 represents the variance of the variable defined by the frequency component. Similarly, higher values of this feature are associated to patients affected by OSAS.

- Feature 10. Third statistical moment in the frequency domain (SMF3). SMF3 measures the asymmetry of the variable defined by the frequency component. It is positive for both groups of subjects. High values are associated to OSAS ~~-~~negative subjects since the PSD of their signals tend to be concentrated on frequencies close to zero.

- Feature 11. Fourth statistical moment in the frequency domain (SMF4). SMF4 evaluates the sharpness of the distribution defined by the frequency component. It is expected to be higher for OSAS ~~-~~negative subjects since the power of the signal is concentrated in low frequency components.

In addition, other three spectral features were computed from SaO_2 signals. These are directly related to the analysis of the PSD in the frequency band between 0.010 and 0.033 Hz (Zamarrón *et al* 2003). We computed the following features:

- Feature 12. Total area under the PSD (S_T). S_T evaluates the power of the signal under study. High values are associated to signals from positive subjects due to frequent changes and variability.

- Feature 13. Area enclosed in the band of interest (S_B). S_B measures the signal power corresponding to frequencies in the band between 0.010 and 0.033 Hz. As indicated ~~before~~above, it is usually higher for OSAS ~~-~~positive patients.

- Feature 14. Peak amplitude of the PSD in the band of interest (PA). PA represents the most significant frequency component contained in the band between 0.010 and 0.033 Hz. It is expected to be higher in OSAS ~~-~~positive patients due to periodic changes in SaO_2 signals corresponding to these frequencies.

3.2 Feature preprocessing

The preprocessing stage avoids possible differences between the magnitudes of the input features (Bishop 1995). Each of them was normalised to have zero mean and unit variance. The following linear transformation was applied:

$$x_i^n = \frac{\tilde{x}_i^n - \mu_i}{\sigma_i}, \quad (1)$$

where n and i are the sample and feature indices, respectively, x_i^n is the normalised value of feature i for sample n, $\tilde{x}_i^n$ is its corresponding raw value, μ_i is the mean value of feature i and σ_i is its standard deviation.

3.3 Classification

We used multilayer perceptron (MLP) neural network classifiers to process the normalised features. The purpose of this stage is to classify patterns from SaO_2 recordings into one of two possible groups: OSAS positive or negative. MLP neural networks suitably adapt to classification tasks since they are capable of estimating posterior probabilities (Bishop 1995, Haykin 1999). Classifiers based on MLP present some advantages in comparison with other conventional techniques such as discriminant analysis. Mainly, they avoid prior assumptions about the statistical distribution of the input data (Zhang 2000). Moreover, MLP networks can establish complex nonlinear decision boundaries in the input space (Bishop 1995, Haykin 1999).

For a given problem, designing neural networks requires to take some decisions about network architecture and training. In this study, input patterns must be labelled as OSAS positive or negative, which represents a two-class classification problem. A single node is required in the output layer of the network. A logistic activation function was used for this node in order to interpret network outputs as probabilities (Bishop 1995). On the other hand, we decided to evaluate MLP networks with a single hidden layer. As indicated in (Hornik 1991), networks with this architecture are capable of universal approximation. The hyperbolic tangent activation function was used for hidden ~~nodes~~ units since it provides faster convergence of the training algorithms (Haykin 1999).

We trained our MLP networks using a coding scheme $t = 1$ if the input pattern belongs to class C_1 (OSAS ~~-~~positive group) and $t = 0$ if it belongs to class C_0 (OSAS ~~-~~negative group), with t representing the target value and C_j the membership variable. As a result, the network output can be interpreted as the probability of having OSAS given the input pattern. Therefore, the Bayes decision rule can be applied to perform pattern classification in order to minimise the probability of misclassification (Bishop 1995). It states the following:

$$\text{Decide } C_j \text{ for } \mathbf{x} \text{ if } p(C_j|\mathbf{x}) = \max_{j=0,1} p(C_j|\mathbf{x}), \quad (2)$$

where $p(C_j|\mathbf{x})$ represents the posterior probability of class C_j for the input pattern $\mathbf{x}$.

Training MLP networks involves ~~to~~ adjusting network weights ~~from~~ based on a finite set of data that represents the statistical properties of the problem. In this context, two different techniques can be applied: the maximum likelihood criterion and the Bayesian inference approach. In this study, we propose to compare the performance of MLP classifiers from both frameworks to assist in OSAS diagnosis using oximetry data.

3.3.1 Maximum likelihood

In classification problems, the probability of observing the target value t (represented by a 0-1 encoding) is modelled with a Bernoulli distribution (Bishop 1995). It is expressed as:

$$p(x) = y^{t}(1-y)^{1-t} \quad (3)$$

where y is the network output value, i.e. the estimate for the posterior probability $p(C_1|\mathbf{x})$. According to this, the likelihood (L) of observing the training set is given by the following expression:

$$L = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n} \quad (4)$$

where N is the size of the training set and it has been assumed that training samples are statistically independent. The aim of the maximum likelihood approach is to adjust network weights in order to maximise L. Usually, the negative logarithm of the likelihood is considered. Then, the optimisation process is equivalent to minimise the expression resulted from that transformation. It is referred to as the cross-entropy error function (E_D) and is expressed as (Bishop 1995):

$$E_D = -\sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \quad (5)$$

~~As a result, the maximum likelihood approach allows to obtain the weight vector **w** that best fits the training data~~The optimal weights cannot be found directly since there is a non-linear dependence on the weights. Instead, an iterative approach is used: partial derivatives of the error function with respect to the weights can be determined analytically, and these are used in second-order non-linear optimisation algorithms. ~~Subsequently~~After weight optimisation, network predictions for new input patterns are computed as:

$$y = f(\mathbf{x},\mathbf{w}) = g_l\left\{\sum_{h=1}^{H}\left[w_{ho}\, g_t\left(\sum_{i=1}^{I} w_{ih}x_i + b_h\right) + b_o\right]\right\} \quad (6)$$

where I is the number of features in the input vector, H is the number of hidden ~~neurons~~units, w_{ho} is the weight connecting hidden ~~neuron~~ unit h with output ~~neuron~~ unit o, b_o is the bias associated to output ~~neuron~~ o, w_{ih} is the weight connecting the feature i of the input pattern with hidden ~~neuron~~ unit h, b_h is the bias associated to hidden ~~neuron~~ unit h, $g_t(\cdot)$ is the activation function for ~~neurons~~ units in the hidden layer and $g_l(\cdot)$ is the activation function for the output layer ~~neuron~~unit.

3.3.2 Bayesian inference

The Bayesian approach ~~suggests to~~ models the posterior probability density function of the weight vector rather than determining an optimum set of network weights (Bishop 1995, Nabney 2002). When the maximum likelihood approach is applied, d~~D~~ifferent (representative) training sets ~~representing the problem under study~~ lead to different network weights ~~when the maximum likelihood approach is applied~~. Bayesian techniques aim to ~~consider~~ account for this uncertainty by representing the degrees of belief in the values of the weight vector (Bishop 1995) with a probability distribution. According to ~~the~~ Bayes' theorem, the posterior distribution of the weights (**w**) given the training set (D) is expressed as:

Formatted: Indent: First line: 0 cm

Comment [n4]: I would move this equation and the description befote equation (2).

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})\,p(\mathbf{w})}{p(D)}, \tag{7}$$

where $p(\mathbf{w})$ is the prior probability function over weight space, $p(D|\mathbf{w})$ is the likelihood of the training data (computed using equation (4)) and $p(D)$ is a normalisation factor known as the evidence (Nabney 2002). Once the posterior has been calculated, it can be used to infer the distribution of output values by computing the following expression:

$$p(\mathbf{x}, D) = \int p(\mathbf{x}, \mathbf{w})\,p(\mathbf{w}|D)\,d\mathbf{w} \tag{8}$$

where $p(t|\mathbf{x},\mathbf{w})$ is the model for the distribution of the noise on the target for a given weight vector. In pattern classification problems, it is given by the expression in (3). Therefore, the probability of membership of an input pattern to the OSAS positive group is obtained as:

$$p(C_1|\mathbf{x}, D) = \int p(C_1|\mathbf{x}, \mathbf{w})\,p(\mathbf{w}|D)\,d\mathbf{w} = \int y(\mathbf{x}, \mathbf{w})\,p(\mathbf{w}|D)\,d\mathbf{w} \tag{9}$$

We used the evidence procedure (Mackay 1992) to implement Bayesian MLP networks. A Gaussian approximation to the posterior (Laplace approximation) is used to ~~solve~~ compute integrals such as those ~~mentioned before (Nabney 2002)~~in equations (8) and (9). In the absence of data, the prior probability distribution is chosen ~~in order~~ to favour small weights. Smooth mappings are preferred since they provide better generalisation (Bishop 1995). Thus, the prior is modelled using the following zero-mean ~~exponential~~ Gaussian function:

$$p(\mathbf{w}) = \frac{1}{Z_W}\exp(-\alpha E_W) = \frac{1}{Z_W}\exp\left(-\frac{\alpha}{2}\|\mathbf{w}\|^2\right), \tag{10}$$

where Z_W is a normalisation factor and α, ~~is~~ referred to as a hyperparameter, is the inverse variance of the Gaussian. It controls the distribution of ~~other~~ the network parameters, i.e. network weights and biases. On the other hand, the likelihood function for the training data given in (4) can be written as:

$$L = p(D|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n} = \exp(G(D|\mathbf{w})) \tag{11}$$

where $G(D|\mathbf{w}) = -E_D$ is the cross-entropy error function as in equation (5). According to the expression in (7), ~~the~~ Bayes' theorem is applied to compute the posterior probability from $p(\mathbf{w})$ and $p(D|\mathbf{w})$. It is given by:

$$p(\mathbf{w}|D) = \frac{1}{Z_S}\exp(G - \alpha E_W) = \frac{1}{Z_S}\exp(-S(\mathbf{w})), \tag{12}$$

where Z_S is a normalisation factor for the Gaussian, $E_W = \frac{1}{2}\|\mathbf{w}\|^2$ and $S(\mathbf{w}) = G(D|\mathbf{w}) + E_W(\mathbf{w})$.

In practice, this distribution is approximated by a Gaussian centred on the maximum posterior weight vector ($\mathbf{w}_{MP}$):

$$p(\mathbf{w}|D) \approx \frac{1}{Z_S^*}\exp\left[-S(\mathbf{w}_{MP}) - \frac{1}{2}\Delta\mathbf{w}^T\mathbf{A}\Delta\mathbf{w}\right] \tag{13}$$

where Z_S^* is the normalisation factor for the Gaussian, $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{MP}$ and $\mathbf{A}$ is the Hessian matrix of the error function S($\mathbf{w}$). The implementation of the Bayesian MLP network requires the use of a nonlinear optimisation algorithm to find $\mathbf{w}_{MP}$, which is found by minimising S($\mathbf{w}$). Additionally, the hyperparameter α is periodically updated during this optimisation process. The maximum posterior value for the hyperparameter is found by maximising the likelihood $p(D|\alpha)$ (Bishop 1995, Nabney 2002).

The hyperparameter controls the shape of the prior distribution of weights. Complex forms of this hyperparameter can be specified. Specifically, different values of α can be used to model the distribution of the weights associated to each input feature. This procedure is known as automatic relevance determination (ARD) (Neal 1996, Nabney 2002). Once the network has been trained, the importance of an input feature (x_i) can be evaluated by analysing its associated hyperparameter (α_i). A low value of α_i corresponds to a large variance prior, which allows weights of large magnitude. Thus, the feature x_i is very relevant for predicting the output. Conversely, a large value of α_i is interpreted as low influence of the input feature on network output. As a result, feature selection is implemented in the network training process.

4 Results

We computed the proposed 14 features from each SaO_2 signal in our database. Table 2 summarises the mean value of each feature for the complete set of signals. Subsequently, the preprocessing stage was applied on each feature. The obtained patterns were fed into a MLP classifier. We developed MLP networks using the two optimisation frameworks described before: maximum likelihood (ML) and Bayesian inference (BY). The Netlab software was used to implement these algorithms (Nabney 2002).

4.1 Training

The training set with 74 subjects was used to develop our classification algorithms. The scaled conjugate gradient was used to optimise both types of MLP classifiers (Moller 1993). It was applied to minimise the cross-entropy error function given by E_D in order to determine the optimum weight vector for networks corresponding to the ML criterion. In the BY framework, the optimisation algorithm was used to find the most probable set of weights by minimising the error function S(**w**); this is alternated with re-estimation of hyperparameters using the evidence framework (Mackay 1992).

Different network architectures of both types were evaluated by varying the number of hidden nodes from 2 to 20. The generalisation performance of each network configuration was measured with sensitivity, specificity and accuracy. Additionally, we used receiver operating characteristics (ROC) analysis. The area under the ROC curve (AUROC) was computed as a measure of classification ability (Hanley and McNeil 1982). Given the random nature of network initialisation, the performance measures were averaged for a total of 10 runs, i.e. a total of 10 different networks were trained for each configuration. Table 3 summarises the results achieved on the training set for both types of MLP networks.

As ~~it~~ can be observed, ML-MLP networks can correctly classify all the samples in the training set. This is achieved for network configurations with more than 4 nodes in the hidden layer. In contrast, BY-MLP networks provided lower classification ability on the training data, reaching an accuracy value around 90% and an AUROC close to 0.95.

Table 2. Time and spectral features extracted from nocturnal oximetry recordings for all subjects under study. Data are presented as mean ± standard deviation. *n*: number of subjects; SMT1: first statistical moment in the time domain; SMT2: second statistical moment in the time domain; SMT3: third statistical moment in the time domain; SMT4: fourth statistical moment in the time domain; ApEn: approximate entropy; CTM: central tendency measure; LZC: Lempel-Ziv complexity; SMF1: first statistical moment in the frequency domain; SMF2: second statistical moment in the frequency domain; SMF3: third statistical moment in the frequency domain; SMF4: fourth statistical moment in the frequency domain; S_T: total area under the power spectral density; S_B: area enclosed in the band of interest; PA: peak amplitude of the power spectral density in the band of interest.

TIME FEATURES			
	All (*n* = 187)	OSAS Positive (*n* = 111)	OSAS Negative (*n* = 76)
SMT1	93.36 ± 5.29	91.78 ± 5.37	95.67 ± 4.26
SMT2	2.29 ± 2.02	3.18 ± 2.08	0.98 ± 0.93
SMT3	-0.50 ± 0.50	-0.60 ± 0.46	-0.35 ± 0.54
SMT4	4.92 ± 2.16	4.25 ± 1.64	5.89 ± 2.44
ApEn	0.80 ± 0.38	1.03 ± 0.28	0.47 ± 0.25
CTM	0.47 ± 0.28	0.30 ± 0.20	0.71 ± 0.18
LZC	0.49 ± 0.18	0.60 ± 0.13	0.34 ± 0.14
SPECTRAL FEATURES			
	All (*n* = 187)	OSAS Positive (*n* = 111)	OSAS Negative (*n* = 76)
SMF1	0.0104 ± 0.0050	0.0121 ± 0.0051	0.0079 ± 0.0037
SMF2	0.0135 ± 0.0031	0.0139 ± 0.0031	0.0130 ± 0.0031
SMF3	2.85 ± 1.56	2.27 ± 1.08	3.69 ± 1.76
SMF4	17.03 ± 23.41	12.11 ± 9.71	24.22 ± 33.64
S_T	15.21 ± 25.75	23.31 ± 29.54	3.37 ± 11.22
S_B	5.88 ± 11.51	9.51 ± 13.78	0.59 ± 1.53
PA	840.75 ± 1841.80	1363.05 ± 2239.52	77.93 ± 251.64

Table 3. Classification results achieved by maximum likelihood and Bayesian MLP classifiers on the training set. Se: sensitivity; Sp: specificity; Ac: accuracy; AUROC: area under the ROC curve.

	Maximum likelihood				Bayesian inference			
Hidden ~~nodes~~units	Se (%)	Sp (%)	Ac (%)	AUROC	Se (%)	Sp (%)	Ac (%)	AUROC
2	99.77 ± 0.72	90.33 ± 3.67	95.95 ± 1.56	0.98 ± 0.01	92.50 ± 1.53	86.70 ± 0.00	90.10 ± 0.91	0.96 ± 0.00
4	99.77 ± 0.72	100.00 ± 0.00	99.86 ± 0.43	1.00 ± 0.00	90.20 ± 2.64	87.00 ± 1.05	88.90 ± 1.66	0.95 ± 0.01
6	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	90.90 ± 1.86	87.70 ± 2.25	89.60 ± 1.28	0.95 ± 0.00
8	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	89.80 ± 1.20	89.00 ± 1.61	89.50 ± 0.57	0.95 ± 0.00
10	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	88.90 ± 1.29	88.70 ± 1.72	88.80 ± 1.11	0.95 ± 0.00
12	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	88.60 ± 1.07	88.00 ± 2.33	88.40 ± 0.70	0.95 ± 0.00
14	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	88.00 ± 1.53	87.70 ± 1.61	87.80 ± 1.27	0.95 ± 0.01
16	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	88.60 ± 0.00	88.00 ± 1.72	88.40 ± 0.70	0.95 ± 0.00
18	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	88.90 ± 0.72	87.00 ± 1.05	88.10 ± 0.57	0.95 ± 0.00
20	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00 ± 0.00	88.40 ± 0.72	86.70 ± 0.00	87.70 ± 0.43	0.95 ± 0.00

Formatted Table

4.2 Test

The trained networks were also evaluated on previously unseen data to obtain a more reliable measure of generalisation capability. The test set with 113 subjects was used. The results achieved by ML-MLP and BY-MLP networks are summarised in table 4.

As expected, the results on the test set were lower than those achieved on the training set by both types of MLP networks. However, the decrease was significantly marked for ML-MLP networks. This was not very surprising. The networks have $16H+1$ parameters (where H is the number of hidden units) and hence when $H>4$ there are more parameters than training data samples. In addition, it was observed that there were slight differences between network configurations with a different number of hidden nodes~~. It~~ which was observed for both ML-MLP and BY-MLP networks. Thus, less complex algorithms such as networks with 2 hidden nodes are preferred. The ML-MLP algorithm with this configuration provided a mean accuracy of 76.81% (86.42% sensitivity and 62.83% specificity) and a mean AUROC of 0.86. The accuracy reached by BY-MLP classifiers on the test set was substantially higher. The BY-MLP network with 2 nodes in the hidden layer achieved a mean accuracy of 85.58% (87.76% sensitivity and 82.39 % specificity) and a mean AUROC of 0.90.

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Comment [n5]: I would also include results from logistic regression as comparison.

Table 4. Classification results achieved by maximum likelihood and Bayesian MLP classifiers on the test set. Se: sensitivity; Sp: specificity; Ac: accuracy; AUROC: area under the ROC curve.

	Maximum likelihood				Bayesian inference			
Hidden ~~nodes~~units	Se (%)	Sp (%)	Ac (%)	AUROC	Se (%)	Sp (%)	Ac (%)	AUROC
2	86.42 ± 1.79	62.83 ± 3.90	76.81 ± 0.91	0.86 ± 0.04	87.76 ± 0.63	82.39 ± 2.39	85.58 ± 0.84	0.90 ± 0.00
4	82.39 ± 5.53	67.39 ± 3.24	76.28 ± 3.68	0.84 ± 0.02	85.07 ± 4.39	84.57 ± 0.69	84.87 ± 2.87	0.89 ± 0.01
6	84.48 ± 5.37	68.04 ± 2.91	77.79 ± 3.08	0.84 ± 0.03	86.42 ± 0.47	84.13 ± 1.05	85.49 ± 0.62	0.90 ± 0.01
8	84.33 ± 3.32	66.74 ± 3.56	77.17 ± 1.71	0.81 ± 0.03	86.27 ± 0.63	84.78 ± 0.00	85.66 ± 0.37	0.90 ± 0.00
10	85.37 ± 3.57	66.74 ± 2.72	77.79 ± 2.45	0.82 ± 0.03	86.42 ± 1.10	84.78 ± 0.00	85.75 ± 0.65	0.90 ± 0.00
12	85.97 ± 1.75	65.87 ± 3.56	77.79 ± 1.93	0.82 ± 0.02	85.82 ± 1.61	84.78 ± 0.00	85.40 ± 0.96	0.90 ± 0.00
14	84.93 ± 2.77	66.74 ± 1.79	77.52 ± 1.68	0.82 ± 0.01	85.37 ± 1.37	84.78 ± 0.00	85.13 ± 0.81	0.90 ± 0.00
16	84.78 ± 4.15	65.87 ± 2.06	77.08 ± 2.75	0.81 ± 0.03	85.82 ± 1.27	84.78 ± 0.00	85.40 ± 0.75	0.90 ± 0.00
18	84.33 ± 3.39	67.39 ± 3.69	77.43 ± 1.78	0.81 ± 0.03	85.37 ± 1.54	84.78 ± 0.00	85.13 ± 0.91	0.90 ± 0.00
20	85.67 ± 1.04	64.57 ± 2.30	77.08 ± 1.14	0.80 ± 0.01	85.37 ± 1.54	85.00 ± 0.69	85.22 ± 0.84	0.90 ± 0.00

Formatted Table

5 Discussion

Classifiers based on Bayesian MLP networks were evaluated as a~~n assistant~~ tool to assist in OSAS diagnosis from nocturnal oximetry data. A total of 14 features computed from time-domain and frequency-domain analysis of SaO_2 signals were used as inputs. The performance of these classifiers was compared with that achieved by common ML-MLP networks. We found that the classification ability of these algorithms was improved by ~~BY-MLP. Our results indicate that~~using Bayesian inference ~~represents a more effective optimisation technique to implement our MLP classifiers~~. It is often preferred in applications with a ~~reduced~~ small dataset such as that presented in this study.

BY-MLP networks provided the best classification performance with an accuracy of 85.58% and an AUROC of 0.90 on the test set. These networks significantly outperformed classifiers based on ML-MLP (76.81% accuracy and 0.86 AUROC). The opposite situation was observed for data in the training set. The accuracy reached on this set by BY-MLP networks ranged from 87.70% (BY-MLP with 20 hidden nodes) to 90.10% (BY-MLP with 2 hidden nodes). For ML-MLP networks, it was close or equal to 100%. The comparison of the results achieved on both sets indicates that ML-MLP networks overfitted the training data. These networks are more likely to be affected by overfitting since their inherent complexity is greater than that of BY-MLP networks (Bishop 1995). For a given training set, the bias-variance trade-off requires the best compromise between a good representation of the data and network complexity in order to achieve high generalisation capability. A simple or inflexible model (large bias) can lead to underfit the data. The model may not have the ability to learn enough the underlying distribution. On the other hand, a too complex or flexible model (large variance) could capture the noise present in the training data, leading to overfitting. Both situations result in poor generalisation (Bishop 1995, Haykin 1999).

The main advantage of BY-MLP networks is that the effect of the bias-variance dilemma is ~~no relevant~~reduced. They implement regularisation by including the hyperparameter α, which controls the distribution of the weights and the network complexity during training (Nabney 2002). Many real applications such as the proposed OSAS diagnosis problem present a training set with a limited size. Thus, regularisation techniques are required to develop effective neural network algorithms. In our preceding work, an MLP network with nonlinear features from oximetry data was developed to help in OSAS diagnosis (Marcos *et al* 2008). It provided an accuracy of 85.5% and an AUROC of 0.90 on a test set with 83 subjects using weight decay regularisation. These results were similar to those obtained with BY-MLP networks developed in this study. However, weight decay requires the user to adjust an additional regularisation parameter~~,~~ which controls the trade-off between reducing training error and favouring small weight values. A ~~high~~ large number of experimental runs are needed to optimise this parameter. Therefore, the Bayesian approach represents a more efficient regularisation technique since the hyperparameter is automatically updated in the training process. Moreover, it allows us to specify complex priors to implement automatic relevance determination. This procedure provides a means to perform feature selection in the network training algorithm (Nabney 2002). From our experiments, we found that ApEn, LZC, SMT1 and SMT4 were determined as the most relevant features to classify SaO_2 recordings as OSAS ~~–~~positive or negative. This result suggests that time-domain features from oximetry data provided the most relevant information to detect OSAS.

Other methodologies to analyse oximetry signals for OSAS diagnosis ~~were~~ have been ~~previously~~ proposed ~~to analyse oximetry signals for OSAS diagnosis~~. Visual inspection of these recordings provided a sensitivity of 91% and a specificity of 69% (Rodríguez *et al* 1996). However, this is a time-consuming technique and the interpretation of the signals may differ from one expert to another. Conventional oximetry indices were proposed for automated analysis of SaO_2 recordings. Most of the commercial oximeters can provide ODI2, ODI3, ODI4 and the cumulative time spent below 90% of saturation. The diagnostic capability of these indices has been evaluated in other studies. The reported results varied among different researchers: the sensitivity ranged from 31% to 98% and the specificity from 41% to 100% (Netzer *et al* 2001). Vázquez *et al* (2000) reported 98% sensitivity and 88% specificity by

Comment [n6]: i.e. time-domain features were better than frequency-domain features

Comment [n7]: Include a table of final alpha values for each feature. Ideally, it might also be good to train some models on a reduced set of inputs to see if this improves performance.

Comment [n8]: It would be very helpful to include Accuracy (and where posible AUROC) values in this paragraph. Perhaps a summary table of the algorithm results would also help.

means of ODI4 when a threshold of 15 events/h was applied to define OSAS. It represents the highest diagnostic accuracy reported by these indices. However, a definition of arousal different to the criteria proposed by the Atlas Task Force was applied (The Atlas Task Force 1992). Lévy *et al* (1996) proposed an additional index (Δ index) from oximetry signals to detect OSAS. It is a measure of the variability of the SaO_2 recordings. This parameter achieved 98% sensitivity and 46% specificity using a threshold of 15 events/h on the AHI to determine the presence of OSAS. Magalang *et al* (2003) achieved similar results using the Δ index, with a sensitivity of 91% and a specificity of 59%. These results were improved by using the Δ index together with the other conventional indices, which yielded a sensitivity of 90% and a specificity of 70%. Finally, Roche *et al* (2002) suggested to combine information from oximetry data with clinical features using logistic regression. This model achieved an accuracy of 62.1%.

The presented algorithms improved the classification accuracy reported in the cited studies. Moreover, we used our database of SaO_2 recordings to compare the diagnostic capability of our networks and conventional oximetry indices. In order to do this, we computed the classification results of these indices on our signal database. For each index, the threshold (l) that provided the highest accuracy on the training set was selected. Subsequently, it was applied on signals in the test set to compute sensitivity, specificity and accuracy values. In addition, we computed the AUROC from data in the test set. The obtained results are summarised in table 5.

The best diagnostic performance was provided by ODI2 and ODI3, which achieved a sensitivity of 76.12%, a specificity of 93.48% and an accuracy of 83.19%. In addition, CT90 provided the best AUROC value with 0.77. As it can be observed, the BY-MLP classifiers clearly outperformed conventional oximetry indexes. The proposed algorithms represent a more effective technique for automated OSAS diagnosis from SaO_2 data.

Some limitations can be found in our methodology. A larger population would be desirable to obtain a better representation of the statistical properties of our classification problem. The generalisation capability of our networks could be improved by using a larger training set. Similarly, a more accurate estimate of our classification results could be obtained through a larger test set. Moreover, our results were computed as the average value of several runs. A model selection stage is required in order to select an optimum classification algorithm from those trained in our experiments.

In summary, BY-MLP classifiers have shown to be a useful tool to assist in OSAS diagnosis from oximetry data. They achieved an accuracy of 85.58% (87.76% sensitivity and 82.39% specificity) and an AUROC of 0.90. They outperformed ML-MLP networks, which provided an accuracy of 76.81% and an AUROC of 0.86. Our results suggest that Bayesian techniques are preferred for the optimisation of our MLP networks. In addition, the proposed BY-MLP algorithms improved the diagnostic capability of visual inspection of SaO_2 recordings and conventional oximetry indices. Our BY-MLP networks could be used to assist medical experts for early detection of OSAS only using oximetry signals. These algorithms could be applied as an effective technique for OSAS screening, contributing to reduce the number of required PSG tests.

Table 5. Diagnostic results achieved by conventional oximetry indices. ODI2: oxygen desaturation index over 2%; ODI3: oxygen desaturation index over 3%; ODI4: oxygen desaturation index over 4%; CT90: cumulative time spent below 90% of saturation; l: threshold value determined on the training set; Se: sensitivity; Sp: specificity; Ac: accuracy; AUROC: area under the ROC curve.

Index	l	Se (%)	Sp (%)	Ac (%)	AUROC
ODI2	9	76.12	93.48	83.19	0.706
ODI3	8	76.12	93.48	83.19	0.725
ODI4	7.6	73.13	95.65	82.30	0.758
CT90	11	68.66	95.65	79.65	0.774

Other possible future work – classification of segments of dataset so that a more refined diagnosis can be given; it might also increase performance.

Acknowledgment

This work has been partially supported by Ministerio de Ciencia e Innovación and FEDER under project TEC2008-02241, and by Consejería de Sanidad de la Junta de Castilla y León under project GRS 337/A/09.

References

Álvarez D, Hornero R, Abásolo D, Del Campo F and Zamarrón C 2006 Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection *Physiol. Meas.* **27** 399-412

Álvarez D, Hornero R, García M, Del Campo F and Zamarrón C 2007 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** 13-24

Bennet JA and Kinnear WJM 1999 Sleep on the cheap: the role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome *Thorax* **54** 958-9

Bishop CM 1995 *Neural networks for pattern recognition* (Oxford University Press, Oxford, UK)

Cohen ME, Hudson DL and Deedwania PC 1996 Applying continuous chaotic modelling to cardiac signals *IEEE Eng. Med. Biol. Mag.* **15** 97-102

Del Campo F, Hornero R, Zamarrón C, Abásolo DE and Álvarez D 2006 Oxygen saturation regularity analysis in the diagnosis of obstructive sleep apnea *Artif. Intell. Med.* **37** 111-8

El-Solh AA, Magalang UJ, Mador MJ, Dmochowski J, Veeramachaneni S, Saberi A, Draw AM, Lieber BB and Grant BJB 2003 The utility of neural network in the diagnosis of Cheyne-Stokes respiration *J. Med. Eng. Technol.* **27** 54-8

George CFP 2001 Reduction in motor vehicle collisions following treatment of sleep apnoea with nasal CPAP *Thorax* **56** 508-12

Hanley JA and McNeil BJ 1982 The meaning and use of the area under a receiving operating characteristic (ROC) curve *Radiology* **143** 29-36

Haykin S 1996 Neural networks expand SP's horizons *IEEE Signal Process. Mag.* **13** 24-49

Haykin S 1999 *Neural Networks: A Comprehensive Foundation* (Prentice Hall, New Jersey, US)

Hornero R, Álvarez D, Abásolo D, Del Campo F and Zamarrón C 2007 Utility of aproximate entropy from overnight pulse oximetry data in the diagnosis of the obstructive sleep apnea syndrome *IEEE Trans. Biomed. Eng.* **54** 107-13

Hornik K 1991 Approximation capabilities of multilayer feedforward networks *Neural Net~~w~~.* **4** 251-7

Lattimore JL, Celermajer DS and Wilcox I 2003 Obstructive sleep apnea and cardiovascular disease *J. Am. Coll. Cardiol.* **41** 1429-37

Lempel A and Ziv J 1976 On the complexity of finite sequences *IEEE Trans. Inf. Theory* **22** 75-81

Lévy P, Pépin JL, Deschaux-Blanc C, Paramelle B and Brambilla C 1996 Accuracy of oximetry for detection of respiratory disturbances in sleep apnea syndrome *Chest* **109** 395-99

Mackay DJC 1992 The evidence framework applied to classification networks *Neural Computation* **4** 720-736

Magalang UJ, Dmochowski J, Veeramachaneni S, Draw A, Mador MJ, El-Solh A and Grant BJB 2003 Prediction of the apnea-hypopnea index from overnight pulse oximetry *Chest* **124** 1694:701

Marcos JV, Hornero R, Álvarez D, del Campo F, Zamarrón C and López M 2008 Utility of multilayer perceptron neural network classifiers in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry *Comput. Meth. Programs Biomed.* **92** 79-89

Moller MF 1993 A scaled conjugate gradient algorithm for fast supervised learning *Neural Netw.* **6** 525-33

Nabney IT 2002 *Netlab: algorithms for pattern recognition* (Springer, London~~Berlin, Germany~~)

Formatted: Font: Italic

Formatted: Font: Bold

Neal, RM 1996 *Bayesian learning for neural networks*, Lecture notes in statistics 118 (Springer, New York)
Netzer N, Eliasson AH, Netzer C and Kristo DA 2001 Overnight pulse oximetry for sleep-disordered breathing in adults: a review *Chest* **120** 625-33
Pincus SM 2001 Assessing serial irregularity and its implications for health *Ann. NY Acad. Sci.* **954** 245-67
Polat K, Yosunkaya S and Günes S 2008 Comparison of different classifier algorithms on the automated detection of obstructive sleep apnea syndrome *J. Med. Syst.* **32** 243-50
Qureshi A and Ballard RD 2003 Obstructive sleep apnea *J. Allergy Clin. Immunol.* **112** 643-51
Rechtschaffen A and Kales A 1968 *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects* (Brain Information Services, Brain Research Institute, University of California, Los Angeles, US)
Roche N, Herer B, Roig C and Huchon G 2002 Prospective testing of two models based on clinical and oximetric variables for prediction of obstructive sleep apnea *Chest* **121** 747-52
Rodríguez JM, De Lucas P, Sánchez MJ, Izquierdo JL, Peraíta R and Cubillo JM 1996 Usefulness of the visual analysis of night oximetry as a screening method in patients with suspected clinical obstructive sleep apnea syndrome *Arch Bronconeumol* **32** 437-41
The Atlas Task Force 1992 EEG arousals: scoring rules and examples. A preliminary report from the Sleep Disorder task Force of the American Sleep Disorders Association *Sleep* **15** 173-84
Vázquez JC, Tsai WH, Flemons WW, Masuda A, Brant R, Hajduk E, Whitelaw WA and Remmers JE 2000 Automated analysis of digital oximetry in the diagnosis of obstructive sleep apnoea *Thorax* **55** 302-7
Welch PD 1967 The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodogram *IEEE Trans Audio Electroacoust* **15** 70-3
Zamarrón C, Gude F, Barcala J, Rodríguez JR and Romero PV 2003 Utility of oxygen saturation and heart rate spectral analysis obtained from pulse oximetric recordings in the diagnosis of sleep apnea syndrome *Chest* **123** 1567-76
Zhang GP 2000 Neural networks for classification: a survey *IEEE Trans Syst Man Cybern Part C-Appl Rev* **30** 451-62

Formatted: Font: Italic