Abstract—This study focused on prospectively testing diagnostic performance of different logistic regression (LR) models in the context of sleep apnea hypopnea syndrome (SAHS) detection from blood oxygen saturation (SaO2) recordings. Feature extraction, selection and classification procedures were applied. Time, frequency, linear and nonlinear analyses were carried out to compose the initial feature set. Forward stepwise logistic regression (FSLR) was applied for feature selection. LR was used to measure performance classification of single features and an optimum feature subset from FSLR. A training set composed of 148 recordings from patients suspected of suffering from SAHS was used to obtain LR models, which were further validated on a dataset composed of 50 recordings from normal healthy subjects and 21 recordings from SAHS patients, all derived from an independent sleep unit. Diagnostic performance of one-feature LR models from oximetry in the training set significantly changed on further assessments in the test set. On the other hand, FSLR provided a more general LR model in the context of SAHS, which reached an accuracy of 89.7% on the training set and 87.3% on the test set.

Keywords- sleep apnea hypopnea syndrome, oximetry, blood oxygen saturation, stepwise feature selection, logistic regression.

I. INTRODUCTION

The sleep apnea hypopnea syndrome (SAHS) is a sleep-related disorder characterized by frequent breathing pauses, which lead to deep oxygen desaturations, acute blood pressure and heart rate changes, increased sympathetic activity and cortical arousals [1]. It is estimated that 20% of adults have at least mild SAHS [2], although 90% of cases in men and 98% of cases in women may go undiagnosed for many years [3]. Daytime hypersomnolence, neurocognitive dysfunction, metabolic deregulation and/or cardiovascular and cerebrovascular diseases could affect people having undiagnosed SAHS [1,3].
University Hospital” of Marburg, were studied. Statistical, spectral and nonlinear analyses were carried out to compose an initial feature set from SaO₂ recordings. Forward stepwise logistic regression (FSLR) was applied to compose an optimum feature subset. A LR classifier was used to investigate classification performance.

II. DATA SET

One hundred and eighty five (185) subjects composed the overall population under study. Patients were recruited from two independent sleep units: the “Río Hortega Hospital” (RHH) from Valladolid (Spain) and the “Philipps University Hospital” (PUH) from Marburg (Germany). Firstly, a population set composed of 148 consecutive subjects suspected of suffering from SAHS derived to the sleep unit of the RHH was studied. This population set was used to compose LR models from oximetric features. In order to test whether the predicted models from oximetry will fit recordings from another sleep laboratory, LR models were further assessed on an independent test set. The Marburg subset of the SIESTA database from the PUH was used. Healthy subjects with no sleep disturbances composed the control group, whereas patients with a positive diagnosis of SAHS composed the SAHS-positive group.

The standard apnea – hypopnea index (AHI) from PSG was used to diagnose SAHS. Apnea was defined as a drop in the airflow signal greater than or equal to 90% from baseline lasting at least 10s, whereas hypopnea was defined as a drop greater than or equal to 50% during at least 10 s accompanied by a desaturation greater than or equal to 3% and/or an arousal. Subjects with an AHI ≥ 10 events per hour (e/h) were diagnosed as suffering from SAHS. Regarding the population under study from the RHH, a positive diagnosis of SAHS was confirmed in 100 patients (48 SAHS-negative and 100 SAHS-positive). Every subject contributed one PSG each (7.2 ± 0.4 hours of recording, mean ± SD). On the other hand, nocturnal PSG was carried out during two consecutive nights at the PUH sleep unit (7.7 ± 0.8 hours). In the test set from the PUH, 50 PSG studies from 26 healthy subjects composed the control group (24 subjects contributed two recordings each and 2 subjects contributed one recording each), whereas 21 PSG studies from 11 sleep apnea patients composed the SAHS-positive group (10 patients contributed two recordings each and one patient contributed with a single recording).

All SaO₂ recordings from PSG were saved to separate files and processed offline to compose the initial oximetric feature set. SaO₂ was recorded at a sampling rate of 1 Hz. SaO₂ signals presented zero samples at the beginning of the acquisition process and drops to zero due to patient movements along the recording time. An automatic preprocessing stage was carried out to remove these artifacts.

III. METHODOLOGY

Our methodology was divided into three stages. Feature extraction was accomplished in the first stage. A total of 16 features composed the initial feature set from oximetry, which was the input to the subsequent stage. In the second stage, forward stepwise logistic regression (FSLR) was applied for feature selection, in order to obtain the optimum feature subset. Additionally, a LR classifier was used to assess classification performance in the third stage. Independent training and test populations were used to assess feature subsets and LR models.

A. Feature extraction stage

Oximetric recordings were parameterized by means of 16 features from four feature subsets: time domain statistics, frequency domain statistics, spectral and nonlinear features. All features were computed for each whole overnight recording. In the time domain, an averaged feature value per subject was obtained from non-overlapping 512-sample segments. In the frequency domain, a single PSD estimate per recording was computed from overlapping 512-sample segments.

1) Time domain statistics. The amplitude (%) of each SaO₂ signal was used to compute the normalized histogram. First to fourth order statistical moments were computed. Arithmetic mean (M1t), variance (M2t), skewness (M3t) and kurtosis (M4t) in the time domain were applied to quantify central tendency, amount of dispersion, asymmetry and peakedness, respectively [13]:

\[ f_{1\text{st \ statistical \ moment}} = E[x] = \mu = \frac{1}{N} \sum_{n=1}^{N} x_n, \] (1)

\[ n^{th \ - \ order \ statistical \ moment} = E[(x-\mu)^n]. \] (2)

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

i) First to fourth-order moments (M1f–M4f) in the frequency domain [13]. The amplitude (W/Hz) of the PSD function at each single spectral component was used to obtain the normalized histogram.

ii) Median frequency (MF), which is defined as the spectral component which comprises 50% of the total signal power [14]:

\[ 0.5 \sum_{f=0 \text{Hz}}^{0.5f_s} PSD(f) = \sum_{f=0 \text{Hz}}^{MF} PSD(f). \] (3)

iii) Spectral entropy (SE), which is a disorder quantifier related to the flatness of the spectrum [14]:

\[ SE = - \sum_{f} p_f \ln(p_f), \] (4)

where \( p_f \) is the normalized value of the PSD at each frequency component.

3) Spectral features. The frequency band 0.014 – 0.033 Hz proposed by Zamarrón et al. was parameterized [15]:

i) Total spectral power (Pt), which is computed as the total area under the PSD.
ii) Peak amplitude (PA) in the apnea frequency band, which is the local maximum of the spectral content in the apnea frequency range 0.014–0.033 Hz.

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

4) Nonlinear features.

i) Sample entropy (SampEn), which quantifies irregularity in time series, with larger values corresponding to more irregular data [16]:

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

where \( A^m \) and \( B^m \) are the average number of \( (m) \)-length and \( (m+1) \)-length segments \( X_{i0}(i) \) \( 1 \leq i \leq N-m \) with \( d[X_{i0}(i),X_{i1}(j)]=\min_{k=0,\ldots,m-1} \max |x(i+k)-x(j+k)| \), respectively, and

\[ d[X_{i0}(i),X_{i1}(j)] = \max |x(i+k)-x(j+k)|. \]

ii) Central tendency measure (CTM), which provides a variability measure from second order difference plots, assigning larger values to lower variability [17]:

\[ CTM = \frac{1}{N-2} \sum_{i=1}^{N-2} \delta(d_i), \]

where

\[ \delta(d_i) = \begin{cases} 1 & \text{if } \left[ (x(i+2)-x(i+1))^2 + (x(i+1)-x(i))^2 \right]^{1/2} < \rho, \\ 0 & \text{otherwise} \end{cases} \]

iii) Lempel–Ziv complexity (LZC). LZC is a non-parametric measure of complexity linked with the rate of new subsequences and their repetition along the original sequence [18]. The complexity counter \( c(n) \) is increased every time a new subsequence is encountered:

\[ LZC = \frac{c(n)}{b(n)}, \]

where \( b(n) \) is a normalization parameter [18].

B. Feature selection stage

FSLR is a filter approach for feature selection that chooses the strongest variables in a data set. The likelihood ratio test was used to assess statistical differences (\( p \)-value) between nested LR models differing in one degree of freedom. Iterative binary LR processes were applied to describe the relationship between a dichotomous dependent variable (SAHS-negative vs. SAHS-positive) and the independent variables. At each iteration, the stepwise method performs a test for backward elimination followed by a forward selection procedure [19]. Features that contribute with significant information are added to the model, whereas no significant features are removed:

\[ p_{\text{feature}}^{(\text{step})} = \min(p_{j}^{(\text{step})}) < \alpha_E \rightarrow \text{add feature}. \]

The FSLR algorithm stops when all variables have entered the model or when all variables in the model have \( p \)-values lower than the upper limit of minimal significance (\( \alpha_0 \)) and the variables not included in the model have \( p \)-values greater than the lower limit of minimal significance (\( \alpha_0 \)) [19].

C. Feature classification stage

LR relates a categorical dependent variable with a set of input features. For dichotomous problems, input patterns are classified into one of two mutually exclusive categories (SAHS-positive or SAHS-negative in the context of SAHS diagnosis) and the probability density for the response variable can be modeled by a Bernoulli distribution [20]. LR classifiers assign an input vector to the class with the maximum a posteriori probability value. The maximum likelihood criterion is used to optimize coefficients of the independent input features in the logistic model [20].

D. Statistical analysis

Statistical differences were evaluated by means of the nonparametric Mann-Whitney U test. Training and testing processes were carried out using population groups from two independent sleep units. The training set was used to perform feature extraction and feature selection processes, where a number of LR models were computed. All one-feature LR models and the optimum model from FSLR were subsequently assessed in a test set from an independent sleep unit. Sensitivity (SAHS-positive patients correctly classified), specificity (SAHS-negative subjects rightly classified) and accuracy (the total percentage of subjects correctly classified) were computed to quantify classification performance.

IV. RESULTS

Oximetric features in our initial feature set showed statistical significant differences (\( p \)-value < 0.001) between databases from both independent sleep units in the SAHS-negative group (M1t, M3t, M4t, M3f, M4f, MF, SE, PA, SampEn, CTM), in the SAHS-positive group (M1t, M2t, M3t, M1f, M2f, P_r, PA, CTM) or in both groups (M1t, M3t, CTM). Firstly, 16 single LR models were computed using data in the training set. Table I shows the classification performance for each one-feature LR model applying leave-one-out cross-validation. Every LR model was subsequently validated on the test set. Table II summarizes the diagnostic performance of each model on the second sleep unit. Regarding time domain statistics, the diagnostic accuracy of single LR models for M1t and M3t significantly decreased in the test set, whereas the accuracy of the LR model composed of M2t slightly decreased and the accuracy of the LR model composed of M4t significantly increased in the test set. Regarding the frequency domain statistics, the accuracy of the LR model composed of M1f slightly increased in the test set, whereas the accuracy of the models for the remaining features significantly decreased. Regarding conventional
spectral features, the performance of LR models for $P_T$ and $P_R$ significantly decreased in the test set, whereas the accuracy of the model composed of $PA$ significantly increased. In the nonlinear feature set, a drastically higher accuracy was reached by the model composed of $SampEn$. Additionally, the accuracy obtained with the model composed of $CTM$ significantly decreased. On the other hand, almost the same accuracy was obtained both in the training and test sets when the LR model composed of $LZC$ was used.

The FSLR algorithm ran through 4 steps. The diagnostic accuracy increased from 79.7% at step one to 89.7% at the end of the process. $M2t$, $M4t$, $P_R$ and $LZC$ were automatically selected. The following LR model was obtained:

$$\Pi = \frac{1}{1 + e^{-\left(-9.48 + 3.32 M2t - 1.03 M4t + 23.94 P_R + 26.55 LZC\right)}}$$  \hspace{1cm} (12)

Table I shows the classification accuracy of the model in the training set applying leave-one-out cross-validation. An accuracy of 89.7% (92.0% sensitivity and 85.4% specificity) was reached in the training set. Finally, the optimum LR model was further tested. A similar accuracy (87.3%) was reached on the population from an independent sleep unit (95.2% sensitivity and 84.0% specificity).

V. DISCUSSION AND CONCLUSIONS

This study assessed the performance of a number of LR models from oximetry in the context of SAHS diagnosis. An initial feature set composed of 16 features (time vs. frequency and linear vs. nonlinear) was developed to characterize $SaO_2$ dynamics. A filter stepwise selection approach previous to classification was applied to obtain the optimum LR model from oximetry. All one-variable LR models and the optimum LR model from FSLR were subsequently assessed in a dataset from a different sleep unit. Our results showed that the diagnostic performance of almost all individual LR models from each single feature significantly changed in the test set. This agrees with heterogeneity of population derived to sleep units. On the other hand, $M2t$ and $LZC$ reached similar accuracies in the training and test sets. These two features contributed significantly to the optimum model from FSLR. Thus, our LR model achieved similar diagnostic performance on both datasets. Furthermore, our optimum model reached high classification accuracy in the training set (92.0% sensitivity, 85.4% specificity and 89.7% accuracy) and in the test set (95.2% sensitivity, 84.0% specificity and 87.3% accuracy).

In this study, significantly different population databases with different inclusion criteria were used to prospectively test our methodology. The SAHS-negative group in the RHH database is composed of subjects showing previous symptoms of daytime somnolence and disturbed sleep, who finally obtain a negative SAHS diagnosis, whereas the control group from the Marburg subset of the SIESTA database is composed of healthy subjects without previous symptoms of disturbance breathing. Additionally, there are features showing statistical significant differences between both databases also in the SAHS-positive group. Furthermore, SAHS-positive patients were predominant in the RHH database, which could influence the model design, whereas normal subjects are predominant in the PUH database. Our results showed that diagnostic performance is expected to change significantly when individual LR models composed of single features are further assessed on independent sleep units. On the other hand, our optimum LR model from FSLR achieved higher accuracies on both data sets. This agrees with the aim of forward stepwise selection: features are selected taking into account the amount of information added to the model, instead of maximizing classification accuracy.

Other researchers attempted to improve SAHS diagnosis by means of multivariate analysis. Using conventional oximetric indexes, 88.0% sensitivity and 70.0% specificity were reached from stepwise linear

**TABLE I. DIAGNOSTIC PERFORMANCE FOR EACH SINGLE FEATURE AND FOR THE OPTIMUM LR MODEL IN THE TRAINING SET**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Training set (RHH)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Se (%)</td>
<td>Sp (%)</td>
<td>Ac (%)</td>
<td></td>
</tr>
<tr>
<td>$M1t$</td>
<td>98.0</td>
<td>6.3</td>
<td>68.2</td>
<td></td>
</tr>
<tr>
<td>$M2t$</td>
<td>86.0</td>
<td>70.8</td>
<td>81.1</td>
<td></td>
</tr>
<tr>
<td>$M3t$</td>
<td>93.0</td>
<td>27.1</td>
<td>71.6</td>
<td></td>
</tr>
<tr>
<td>$M4t$</td>
<td>91.0</td>
<td>39.6</td>
<td>74.3</td>
<td></td>
</tr>
<tr>
<td>$M1f$</td>
<td>100</td>
<td>0.0</td>
<td>67.6</td>
<td></td>
</tr>
<tr>
<td>$M2f$</td>
<td>78.0</td>
<td>54.2</td>
<td>70.3</td>
<td></td>
</tr>
<tr>
<td>$M3f$</td>
<td>78.0</td>
<td>52.1</td>
<td>69.9</td>
<td></td>
</tr>
<tr>
<td>$M4f$</td>
<td>76.0</td>
<td>75.0</td>
<td>76.4</td>
<td></td>
</tr>
<tr>
<td>$SE$</td>
<td>87.0</td>
<td>54.2</td>
<td>73.2</td>
<td></td>
</tr>
<tr>
<td>$PA$</td>
<td>86.0</td>
<td>18.8</td>
<td>72.3</td>
<td></td>
</tr>
<tr>
<td>$P_R$</td>
<td>85.0</td>
<td>68.8</td>
<td>79.7</td>
<td></td>
</tr>
<tr>
<td>$SampEn$</td>
<td>93.0</td>
<td>4.2</td>
<td>64.2</td>
<td></td>
</tr>
<tr>
<td>$CTM$</td>
<td>81.0</td>
<td>87.5</td>
<td>83.1</td>
<td></td>
</tr>
<tr>
<td>$LZC$</td>
<td>91.0</td>
<td>31.3</td>
<td>71.6</td>
<td></td>
</tr>
<tr>
<td>LR ($M2t$, $M4t$, $P_R$, $LZC$)</td>
<td>92.0</td>
<td>85.4</td>
<td>89.7</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE II. DIAGNOSTIC PERFORMANCE FOR EACH SINGLE FEATURE AND FOR THE OPTIMUM LR MODEL IN THE TEST SET**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Test set (PUH)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Se (%)</td>
<td>Sp (%)</td>
<td>Ac (%)</td>
<td></td>
</tr>
<tr>
<td>$M1t$</td>
<td>100.0</td>
<td>22.0</td>
<td>45.1</td>
<td></td>
</tr>
<tr>
<td>$M2t$</td>
<td>100.0</td>
<td>70.0</td>
<td>78.9</td>
<td></td>
</tr>
<tr>
<td>$M3t$</td>
<td>100.0</td>
<td>0.0</td>
<td>29.6</td>
<td></td>
</tr>
<tr>
<td>$M4t$</td>
<td>76.2</td>
<td>92.0</td>
<td>87.3</td>
<td></td>
</tr>
<tr>
<td>$M1f$</td>
<td>100.0</td>
<td>62.0</td>
<td>73.2</td>
<td></td>
</tr>
<tr>
<td>$M2f$</td>
<td>100.0</td>
<td>0.0</td>
<td>29.6</td>
<td></td>
</tr>
<tr>
<td>$M3f$</td>
<td>85.7</td>
<td>32.0</td>
<td>47.9</td>
<td></td>
</tr>
<tr>
<td>$M4f$</td>
<td>81.0</td>
<td>32.0</td>
<td>46.5</td>
<td></td>
</tr>
<tr>
<td>$SE$</td>
<td>85.7</td>
<td>14.0</td>
<td>35.2</td>
<td></td>
</tr>
<tr>
<td>$PA$</td>
<td>100.0</td>
<td>2.0</td>
<td>31.0</td>
<td></td>
</tr>
<tr>
<td>$P_R$</td>
<td>100.0</td>
<td>66.0</td>
<td>76.1</td>
<td></td>
</tr>
<tr>
<td>$SampEn$</td>
<td>90.5</td>
<td>92.0</td>
<td>91.5</td>
<td></td>
</tr>
<tr>
<td>$CTM$</td>
<td>100.0</td>
<td>0.0</td>
<td>29.6</td>
<td></td>
</tr>
<tr>
<td>$LZC$</td>
<td>90.5</td>
<td>64.0</td>
<td>71.8</td>
<td></td>
</tr>
<tr>
<td>LR ($M2t$, $M4t$, $P_R$, $LZC$)</td>
<td>95.2</td>
<td>84.0</td>
<td>87.3</td>
<td></td>
</tr>
</tbody>
</table>
regression [8]. PCA was applied to a small set of 3 spectral and 3 nonlinear features [11]. First-to-fifth principal components were selected and 93.0% accuracy (97.0% sensitivity and 79.3% specificity) was reached on the test set from the same sleep unit. Forward stepwise LR was also previously applied to a wide feature set from oximetry, reaching 89.7% accuracy (92.0% sensitivity and 85.4% specificity) using cross-validation [12]. In the present study, we analyzed datasets from two different sleep units to prospectively assess LR models from oximetry.

We should take into account some limitations regarding the general application of our methodology. Desaturations in the overnight SaO2 profile could not be exclusively due to apnea events typical of SAHS, which could influence our results. In addition, further work is required to test the performance of our methodology from ambulatory portable monitoring at patient’s home. Additional feature selection techniques could be assessed in the context of OSA diagnosis, such as variable ranking, factor analysis, subspace clustering or genetic algorithms. Moreover, more complex classifiers could be used to deal with uncertainty in the data and to assess the generalization of our optimum feature subset, such as Fisher’s discriminant, neural networks or support vector machines.

In summary, LR models from oximetry have been assessed using populations from independent sleep units. Significant statistical differences between databases were found for many oximetric features, which decrease diagnostic performance of single LR models. The FSLR-derived model, composed of a reduced number of features automatically selected, achieved a balanced sensitivity – specificity pair and high accuracy in both datasets from different sleep units.

ACKNOWLEDGMENT

This work has been partially supported by Ministerio de Ciencia e Innovación and FEDER grant TEC 2008-02241 and the grant project from the Consejería de Sanidad de la Junta de Castilla y León GRS 337/A/09. D. Álvarez was in receipt of a PIRTU grant from the Consejería de Educación de la Junta de Castilla y León and the European Social Fund.

REFERENCES