# Automatic classification of respiratory patterns involving missing data imputation techniques

**Elena M. Hernández-Pereira [a,*], Diego Álvarez-Estévez [b], Vicente Moret-Bonillo [a]**

[a] Department of Computer Science, Faculty of Informatics, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

[b] Sleep Center, Medisch Centrum Haaglanden, Lijnbaan 32, 2512VA The Hague, The Netherlands

**Abstract:**

A comparative study of the respiratory pattern classification task, involving five missing data imputation techniques and several machine learning algorithms is presented in this paper. The main goal was to find a classifier that achieves the best accuracy results using a scalable imputation method in comparison to the method used in a previous work of the authors. The results obtained show that in general, the Self-Organising Map imputation method allows non-tree based classifiers to achieve improvements over the rest of the imputation methods in terms of the classification accuracy, and that the Feedforward neural network and the Random Forest classifiers offer the best performance regardless of the imputation method used. The improvements in terms of accuracy over the previous work of the authors are limited but the Feed Forward neural network model achieves promising results.

**Keywords:**

Respiratory pattern classification, Missing data imputation, Machine, learning

## 1. Introduction

Medical decision-support systems (MDSS) have always played an important role in medical practice. The MDSS can help physicians in the diagnosis of any disorder using clues ob-tained from signals or images taken from the subject having the disorder. The objective of this work is in the field of the diagnosis of the Sleep Apnoea-Hypopnoea Syndrome (SAHS). In particular a machine learning MDSS is designed to distin-guish sleep apnoeic events (apnoeas and hypopnoeas) from normal breathing.

Sleep apnoea is defined as a pause in breathing, or cessa-tion of the airflow in the respiratory tracts, of at least 10 s in duration. The event is described as a hypopnoea when, rather than a complete cessation, a considerable reduction occurs in the airflow accompanied by a desaturation of oxygen levels in arterial blood. In addition, a micro-arousal happens during sleep that is related to the resolution of these apnoeic events. Since these micro-arousals happen at each event, the physi-ological structure of sleep becomes fragmented. The invol-untary periodic repetition of these respiratory pauses constitutes one of the most frequent sleep disorders: the sleep apnoea-hypopnoea syndrome. The most effective method for SAHS diagnosis is made on the basis of the analysis of a nocturnal polysomnogram, which means a continuous and simultaneous recording during sleep of a set of variables

including airflow in the upper air tracts, oxygen saturation ($SaO_2$) in arterial blood and respiratory effort (both abdominal and thoracic). Following conventional clinical criteria, the apnoeic episodes are detected in the airflow signal, using the information derived from the electrophysiological and oxygen saturation signals as context for interpretation (Berry et al., 2012).

Diagnosis models in SAHS are usually constructed from records that include the polysomnogram information. However, clinical information databases commonly contain missing values or incomplete data where the simple and commonly-used strategy to deal with these gaps is to directly ignore them. Such deletion reduces the number of available cases for analysis and can introduce substantial biases in the study, especially when missing data are not randomly distributed. In this sense, missing data imputation is an area of statistics that has attracted much attention in recent decades.

When imputing missing values, assumptions about their true distribution have to be made. The most favourable form of missingness is missing completely at random (MCAR), which means that the probability of a value being missing is independent of all values in the data set, observed and unobserved. Missing at random (MAR) is less restrictive, as it arises if the probability of missing data of a particular variable could depend on other variables in the data set but not on the variable's value itself. The most severe form of missingness is missing not at random (MNAR), which allows missingness to depend on missing values. The probability of missing data is related to the value of the variable even if other variables in the analysis are controlled (Dahl, 2007; Little & Rubin, 2002).

Different strategies inspired in statistics and machine learning have been developed to address the data imputation problem. A review of the literature reveals that the efficacy of the proposed methods depends strongly on the problem domain (e.g., number of cases, number of variables, missingness patterns), and thus there is no clear indication that favours one method over the others (Ribelles, Martin, & Franco, 2010). Once the missing data are imputed, it is important to evaluate the performance of the imputation method through determining the effect of the imputation on subsequently performed classification. A desirable characteristic for an imputation method is that the missing data estimation is aimed at improving the classification accuracy results. Recent studies have investigated the impact of imputation on the accuracy of the subsequently performed classification. Acuña and Rodriguez (2004) have investigated the effect of four methods that deal with missing values—case deletion, mean imputation, median imputation, and k-nearest neighbours (KNN)—. The classification was performed using linear discriminant analysis and KNN. Their results show that imputation does not have a significant effect on the accuracy of classification. Batista and Monard (2003) tested three imputation methods— mean, mode and KNN— with two classifiers, namely, C4.5 decision tree and CN2 rule induction algorithm. The results show that KNN imputation results in good accuracy, but only when attributes are not highly correlated to each other.

Ribelles et al. (2010) evaluated the performance of several statistical and machine learning imputation methods that were used to predict early breast cancer relapse. The imputation methods used were the mean, hot-deck, three multiple imputation methods using software packages, multilayer perceptron (MLP), KNN and self-organising map (SOM). Once the unknown data were imputed, a prognostic model was created based on artificial neural networks. All imputation methods except for the hot-deck method led to an improvement in prediction accuracy. The machine learning-based techniques outperformed statistical imputation methods in the prediction of patient outcome and were significantly different from those methods in which records with missing values were eliminated.

Rahman and Davis (2013) explored the use of different missing value imputation techniques for incomplete cardiovascular data. Mean imputation, fuzzy unordered rule induction algorithm imputation, decision tree imputation and support vector machine (SVM) imputation were the imputation models studied and the final data sets were classified using several machine learning-based techniques— decision tree, fuzzy unordered rule induction, KNN and K-mean clustering. The final classifier performance was improved when the fuzzy unordered rule induction algorithm was used to predict missing attribute values for K-mean clustering, and for most of the cases, machine learning techniques were found to perform better than mean imputation.

Ritthipravat, Kumdee, and Bhongmakapat (2013) investigated efficient missing data techniques— complete-case analysis, mean imputation, KNN imputation and Expectation Maximisation (EM)— for prediction of nasopharyngeal carcinoma recurrence. Three predictive models, i.e. single point, multiple-point and sequential neural network models were used in the investigation. The results showed that the EM imputation was superior to the other missing data techniques particularly when the sequential neural network was employed.

Garca-Laencina, Sancho-Gmez, and Figueiras-Vidal (2013) presented a Multi-Task Learning (MTL) approach using MLP networks to impute missing values. In this work, classification and imputation were combined in one neural architecture where classification was used as the main task and the imputation of each incomplete feature as a secondary task. The performance of the MTL network has been compared with four imputation procedures—KNN, SOM, MLP and a Gaussian Mixture Model— to solve some synthetic and real problems. Experimental results showed that the proposed method was never worse than the other imputation techniques and also showed the capacity to provide better results when the effects of missing values are considerable.

Mitra and Samanta (2015) proposed an intelligent system for hepatitis disease diagnosis using a multiple imputation technique for managing missing values, performed by a bootstrap-based algorithm. The outputs of this technique were different sets of imputed data that were combined by arithmetic mean to give final results. Once missing data were imputated, a reduction phase by rough-set-based selection was applied and finally, the classification phase was performed using incremental back propagation neural networks and the Levenberg–Marquardt algorithm. The method offers comparable results with other studies in terms of classification accuracy.

To the best knowledge of the authors of this paper, there is no attempt in the literature to take advantage of imputation methods to improve classification performance in the identification of respiratory patterns. Several works for dealing with the identification of individual apnoeic episodes have been found but none of them mentions the treatment of missing values. Várady, Micsik, Benedek, and Benyó (2002) introduced an on-line signal classification method for the detection of the presence or absence of normal breathing. Four different artificial neural networks were presented for the recognition of three different patterns in the respiration signals (normal breathing, hypopnoea, and apnoea). Bystricky and Safer (2004) combined neural networks with dynamic Markov models to assign each instant in the electrocardiogram (ECG) signal recording to one of the following four states: "no apnoea", "onset of apnoea", "apnoea" and "end of apnoea". In this proposal, a neural network is employed to extract a set of morphological characteristics from the beats on the basis of the ECG signal. These characteristics constitute the input to a dynamic Markov model which only contemplates a sequence of transitions permitted between the four aforementioned states. Tian and Liu (2005) have used a time delay network to identify apnoeas on the basis of respiratory airflow signal and peripheral capillary oxygen saturation ($SpO_2$) signal which is an estimation of the oxygen saturation level. The neural network inputs are the area and the standard deviation of the respiratory airflow signal; the basal level and desaturation level of the $SpO_2$ signal; and a correlation coefficient between the $SpO_2$ and respiratory airflow signals. Fontenla-Romero, Guijarro-Berdiñas, Alonso-Betanzos, and Moret-Bonillo (2005) proposed an *ad hoc* technique for identifying apnoeas based on the respiratory airflow signal. They used a mobile window to calculate the absolute value of the difference between the instantaneous value of the respiratory airflow signal and its average value in the window. An adaptive threshold is then applied to the samples of the signal generated in the mobile window to determine whether they correspond with apnoea or normal breathing. Polat, Yosunkaya, and Gunes (2008) compared different classifier algorithms to detect the obstructive sleep apnoea syndrome, which is a particular type of SAHS. The classifier algorithms included C4.5 decision tree, artificial neural network, artificial immune recognition system, and adaptive neuro-fuzzy inference system. The clinical features used were arousals index, apnoea-hypopnoea index, $SaO_2$ minimum value in stage of rapid eye movement, and percent sleep time in stage of $SaO_2$ intervals bigger than 89%. Maali and Al-Jumaily (2012) proposed a genetic fuzzy approach for detecting apnoeic events by using airflow, thoracic and abdominal respiratory movement signals and oxygen desaturation as the inputs. In this approach fuzzy rules and weights are generated by genetic algorithms.

The system MIASOFT (Intelligent Monitoring of the Sleep apnoea-hypopnoea Syndrome), developed by the authors, is a comprehensive medical decision-support system for the diagnosis of SAHS (Álvarez Estévez, 2012). MIASOFT is knowledge-based, and it has been designed to allow explanative capabilities of its results. For that purpose, and with the aim to mimic human handling of generalisation and reasoning procedures, MIASOFT has been implemented using a fuzzy logic inference engine to provide judgments on the basis of similarity and approximation. In MIASOFT, to walk round the problem of missing values, the inference engine makes use of a chaining of different knowledge-bases to account for the situations where different attributes can be missing (Álvarez Estévez & Moret-Bonillo, 2009; Moret-Bonillo, Álvarez-Estévez, Fernández-Leal, & Hernández-Pereira, 2014). Such a solution is far from being optimal and complicates the design when the number of features increases. The scope of this work is to develop a machine learning model that can learn from examples and effectively handle the occurrence of missing values. This approach represents a more straightforward and scalable solution than the one presented in MIASOFT. However the question remains as to whether such an approach can outperform the results of the first solution, and thus we include the MIASOFT system as an additional benchmark.

Five well-known methods, i.e. mean imputation, multiple linear regression, hot-deck, k nearest neighbours and self-organising maps are used to impute absent values in the data set and several linear and non-linear models are applied to classify respiratory patterns as apnoeas, hypopnoeas or normal breathing. The objective of this work is to obtain a machine learning model that achieves the most accurate results in the respiratory pattern identification task. Another goal is to analyse the improvements in identification accuracy against the MIASOFT system results when different algorithms are applied to impute missing data values.

The paper is structured as follows: a description of the materials and methods used in this research is given in Section 2, Section 3 presents the results obtained and finally, a discussion and the conclusions are presented in Section 4.

## 2. Materials and methods

### 2.1. *Data processing*

Patient data which correspond to Polysomnographic (PSG) recordings were gathered from the Sleep Health Heart Study (SHHS) (Quan et al., 1997). This prospective cohort study was originally implemented to analyse the consequences of obstructive sleep apnoea and other sleep-disordered breathing on the development of cardiovascular diseases. The resulting database was then enabled as a resource for subsequent studies. For the purpose of this work, a sample of 95 and 68 recordings have been randomly selected from this database as training and validation set, and test set respectively. Patient demographics from the resulting samples are shown in Table 1. Each recording contains expert consensus on the different events scored by clinicians during the manual offline analysis of the recordings. Annotations regarding the scoring of apnoeic events include hypopnoeas, obstructive apnoeas and central apnoeas for which onset and duration for each event are specified. These annotations will be used as the standard reference for the validation of our approach.

For the construction of the data sets, features are extracted from a subset of PSG signals that involve both respiratory and neurophysiological information. Specifically a total of 9 features are used which are described in Table 2. The process to

**Table 1 – Patient demographics including Apnoea-Hypopnoea Index (AHI) and Body Mass Index (BMI) for training and validation (Train. & Val.) and test data sets.**

| Data set | Number | Male | Age -mean (std*)- | AHI -mean (std*)- | BMI -mean (std*)- |
|---|---|---|---|---|---|
| Train. & Val. | 95 | 49 | 66.27 (10.02) | 46.21 (27.63) | 30.29 (6.03) |
| Test | 68 | 29 | 68.01 (11.27) | 35.09 (19.34) | 29.14 (5.00) |

*Std: standard deviation.

automatically extract these features from the raw biomedical signals contained in the PSG is described in:

- (Álvarez Estévez & Moret-Bonillo, 2009) and (Moret-Bonillo et al., 2014) for the extraction of features 1 to 8. In these references, an explanation can be found of how the individual features that are extracted from each of the different PSG respiratory channels are then related in time to form what has been called an apnoeic pattern (AP), that is, a set of features that together characterise a certain time interval of the PSG and point to the possible occurrence of an apnoeic event.
- (Álvarez-Estévez, Sánchez-Maroño, Alonso-Betanzos, & Moret-Bonillo, 2011) for the detection of Electroencephalogram (EEG) arousals (feature 9). For the association of an EEG arousal to the AP, the criterion described in Sleep Health Heart Study (2002) is used as reference. Specifically, an EEG arousal is associated with an AP if the arousal begins less than 5 s after the end of the AP.

Following the previously described procedures, a total of 39,539 and 27,500 patterns (train, validation and test) have been collected, each one with one possible output namely: (i) normal-respiration, (ii) hypopnoea, or (iii) apnoea. For the training and validation set, the number of each class is 5436 apnoea patterns, 12,078 hypopnoea patterns and 22,025 normal-respiration patterns. For the test set, the number of each class is 1,796, 6619 and 19,085 for apnoea, hypopnoea and normal-respiration patterns respectively. Occurrence of missing values in the data sets is certainly non-missing at random (NMAR). The missingness originate from the situations in which a certain feature cannot be evaluated in the context of the corresponding AP. Such a situation is actually common and may be caused by several reasons including presence of artifacts, inaccuracy of the detection algorithm or simply the current physiological condition (for example, a reduction in breathing may manifest differently across the individual respiratory channels). Characterisation of the features and their related missing rate can be found in Table 2. EEG arousal is not included in the table as it has no missing values in any train, validation and test sets. It is a qualitative and nominal feature with 0 mode.

### 2.2. *Data imputation methods*

Imputation is the process used to determine and assign replacement values for missing data items (Little & Rubin, 2002). Imputation methods are especially useful in situations where a complete data set is required for the analysis. A wide range of methods and tools for data imputation is available. Some methods try to make use of the available information, for example, Listwise or casewise data Deletion techniques (LD), based on the omission of all those records that contain a missing value for one or more variables. Other methods are proper imputation techniques as they compute appropriate values to replace the missing data. So, according to their degree of complexity, we have implemented four of these methods: three statistical methods (mean, multiple linear regression and hot deck) and two machine learning based methods (self-organising maps and k nearest neighbours).

- Mean/mode imputation

This is a method where any missing value of a quantitative variable is replaced by the mean of the observed values for that variable. If the variable is qualitative, the missing values are replaced by the mode.

- Multiple Linear Regression, MLR

Given a missing value for a variable $X$, suppose that $q$ variables have been observed for that record. The records

**Table 2 – Feature characterisation of the data sets.**

| Feature | Range | Scale | Training and validation | | Test | |
|---|---|---|---|---|---|---|
| | | | Mean | Missing rate (%) | Mean | Missing rate (%) |
| Desaturation | 0–100 | Ratio | 2.4673 | 2.77 | 1.5690 | 2.6982 |
| Airflow red. | 0–100 | Ratio | 49.7118 | 44.18 | 43.0438 | 51.5491 |
| Abdominal respiration red. | 0–100 | Ratio | 54.7780 | 33.80 | 48.6391 | 39.1564 |
| Thoracic respiration red. | 0–100 | Ratio | 55.0571 | 37.55 | 49.2979 | 43.9964 |
| Desaturation | 0–400 | Secs. | 14.2083 | 2.77 | 12.0752 | 2.6982 |
| Airflow reduction | 0–400 | Secs. | 21.3744 | 44.18 | 20.9191 | 51.5491 |
| Abdominal respiration red. | 0–400 | Secs. | 24.3508 | 33.80 | 23.7060 | 39.1564 |
| Thoracic respiration red. | 0–400 | Secs. | 25.0748 | 37.55 | 24.0440 | 43.9964 |

Red. = reduction, secs. = seconds.

where these $q+1$ variables are available define a training set, and a regression model to predict $X$ from the $q$ predictors is fitted. Finally, the fitted model provides a prediction for the initial missing value of X. Multiple linear regression has been considered in this study. A number $p>1$ of independent variables $X_1, X_2, \ldots, X_p$ is considered, so a population model $Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p+\varepsilon$, is assumed where Y denotes the dependent variable or response, $X_1, X_2, \ldots, X_p$ are the independent or predictor variables, $\varepsilon$ is a random disturbance or error whose presence represents the absence of an accurate relationship, and $\beta_0, \beta_1, \ldots, \beta_p$ are unknown coefficients or parameters that define the regression hyperplane $\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p+\varepsilon$. If a qualitative variable is considered with c categories, $c-1$ dummy dichotomous variables are introduced into the model:

$$z_{i1}=\begin{cases}0 & \text{if } i\notin\text{category 1}\\ 1 & \text{if } i\in\text{category 1}\end{cases} \tag{1}$$

$$z_{i2}=\begin{cases}0 & \text{if } i\notin\text{category 2}\\ 1 & \text{if } i\in\text{category 2}\end{cases} \tag{2}$$

$$z_{i,c-1}=\begin{cases}0 & \text{if } i\notin\text{category } c-1\\ 1 & \text{if } i\in\text{category } c-1\end{cases} \tag{3}$$

The category c is the base category. Any variable for which the category is built, defined and identified, are all individuals that have value 0 for the other $c-1$ variables. Thus, considering these $c-1$ new variables:

$$Y=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_kX_{ik}+\alpha_1z_{i1}+\alpha_2z_{i2}+\ldots+\alpha_{c-1}z_{i,c-1}+\varepsilon_i$$

$$i=1,2,\ldots,n$$

- Hot-deck imputation

Given an incomplete pattern, this method estimates missing values from similar but complete records of the same data set. The similarity criterion used is the heterogeneous Euclidean-overlap metric (HEOM) (Wilson & Martinez, 1997), which uses the so-called overlap metric for categorical attributes and a normalised city-block distance for linear numeric quantitative attributes. The overlap metric is a normalised Hamming distance given as the percentage of coordinates that differ. The HEOM distance is intended to remove the effects of the arbitrary ordering of categorical values, and it constitutes an overly simplistic approach to handling these kinds of attributes.

Consider that a patient case is represented by an n-dimensional input vector, $x=[x_1,x_2,\ldots,x_n]^T$, and that $m$ is a vector of binary variables such that $m_j=1$ if $x_j$ is unknown and $m_j=0$ if $x_j$ is present. Given a pair of patient cases, represented by $x_a$ and $x_b$, the HEOM distance between them is:

$$d(x_a,x_b)=\sqrt{\sum_{j=1}^{n}d_j(x_{aj},x_{bj})^2} \tag{4}$$

where $d_j(x_{aj},x_{bj})$ is the distance between $x_a$ and $x_b$ on its jth attribute:

$$d_j(x_{aj},x_{bj})=\begin{cases}1 & \text{if } (1-m_{aj})(1-m_{bj})=0\\ d_O(x_{aj},x_{bj}) & \text{if } x_j \text{ is a categorical attribute}\\ d_N(x_{aj},x_{bj}) & \text{if } x_j \text{ is a quantitative attribute}\end{cases} \tag{5}$$

Unknown data are handled by returning a distance value of 1 (i.e., maximal distance) if either of the input values is unknown. The overlap distance function $d_O$ assigns a value of 0 if the discrete attributes are the same; otherwise, the value is 1. The range normalised difference distance function $d_N$ is given by:

$$d_N(x_{aj},x_{bj})=\frac{|x_{aj}-x_{bj}|}{max(x_j)-min(x_j)} \tag{6}$$

where $max(x_j)$ and $min(x_j)$ are the maximum and minimum values, respectively, observed in the training set for the numerical attribute $x_j$; thus, the normalisation attempts to scale the attribute down to the point where differences are almost always less than one, and the resulting distance matrix is set to range between 0 and 1. The difference $|x_{aj}-x_{bj}|$ is the city-block distance (Ribelles et al., 2010).

- K Nearest Neighbours, KNN

The K nearest neighbours algorithm is a method for classifying objects based on closest training examples in the feature space. It is part of a family of learning methods known as instance-based (Aha, Kibler, & Albert, 1991; Cover & Hart, 1967) or lazy learning. These methods are based on the principle that the instances within a data set will generally exist in close proximity with other cases that have similar properties. Learning in these algorithms consists of simply storing the presented training data set. When a new instance is encountered, a set of similar training instances is retrieved from memory and used to make a local approximation of the target function (Mitchell, 1997). In this work, the performance of the KNN algorithm to impute missing values is studied. This procedure will be referred as KNNimpute (Garca-Laencina, Sancho-Gmez, Figueiras-Vidal, & Verleysen, 2009). Given an incomplete pattern, this method selects its K closest cases from the training cases with known values in the attributes to be imputed, such that they minimise some distance measure. Once the K nearest neighbours have been found, a replacement value to substitute for the missing attribute value must be estimated. How the replacement value is calculated depends on the type of data; the mode can be used for qualitative data and the mean for continuous data. Several methods exist to determine the distance between training cases with the Euclidean measure being the most popular (Fujikawa, 2001; Mitchell, 1997).

- Self-Organising Maps, SOM

A Self-organising map is a neural network model made out of a set of nodes that are organised on a 2D grid and fully connected to the input layer. Each node has a specific topological position in the grid, as well as a vector of weights of the same dimension used for the input vectors (Kohonen, 2001). After the SOM model has been trained, it can be used to

estimate missing values. When an incomplete observation is presented to the SOM, the missing input variables are ignored during the selection of the best matching unit (BMU). This selection is made by minimising the distance between the observation and the nodes. The incomplete data are imputed by the feature values of the BMU in the missing dimensions (Ribelles et al., 2010). The SOM imputation approach is implemented using the SOM toolbox. To determine the number of map units, a heuristic formula which depends on the number of observations is used (Vesanto, Himberg, Alhoniemi, & Parhankangas, 2000).

### 2.3. *Classification methods*

In this section, we provide an overview of the methods used in the research for respiratory pattern classification: apnoea, hypopnoea or normal breathing. Several approaches were considered, two linear models – linear discriminant analysis and a proximal support vector machine–, and four non linear ones – a multilayer feedforward neural network, a classification tree, a Random Forest and a deep neural network–.

- Linear Discriminant Analysis, LDA

The linear discriminant analysis is a classification method originally developed by Fisher (1936). It is simple, mathematically robust and often produces models whose accuracy is as good as more complex methods. It consists of searching some linear combinations of selected variables, which provide the best separation between the considered classes. These different combinations are called discriminant functions. It assumes that different classes generate data based on different Gaussian distributions (Srivastava & Carter, 1983).

- Proximal Support Vector Machine, pSVM

The proximal Support Vector Machine (Fung & Mangasarian, 2001) is a method that classifies points assigning them to the closest of two parallel planes (in input or feature space) that are pushed as far apart as possible. The difference with a Support Vector Machine (SVM) is that this one classifies points by assigning them to one of two disjoint half-spaces. The pSVM leads to an extremely fast and simple algorithm by generating a linear or nonlinear classifier that merely requires the solution of a single system of linear equations.

- Multilayer Feedforward Neural Network, FNN

The multilayer feedforward neural network is one of the most commonly used neural network classification algorithms (Bishop, 1995). The architecture used for the classifier consisted of a two layer feed-forward neural network: one hidden and one output layer. It has been demonstrated that, with an appropriate number of hidden neurons, one hidden layer is enough to model any continuous function (Hornik, Stinchcombe, & White, 1989). The optimal number of hidden neurons for this problem was empirically obtained. Logistic transfer functions were used for each neuron in both the hidden and the output layers. The learning algorithm used was the conjugate gradient (Moller, 1993) with the mean squared error cost function. A maximum number of 3000 epochs were performed on the training set.

- Classification Trees

Classification trees are used to predict membership of cases or objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels (Breiman, Friedman, Olshen, & Stone, 1984). Each internal (non-leaf) node of the tree is labelled with an input feature. The arcs coming from a node labelled with a feature are labelled with each of the possible values of the feature. Each leaf of the tree is labelled with a class or a probability distribution over the classes. A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees is by far the most common strategy for learning decision trees from data (Quinlan, 1986).

- Random Forests

Random Forests (Breiman, 2001) are an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes. For an ensemble of decision trees for a multiclass classification function, one of the general methods is Bagging. This method is the simpler, more robust and more highly parallel technique. In the Bagging version used, a fixed-sized fraction of the training data is employed to construct each classifier in the ensemble. The Bagging method simply produces an ensemble of N decision trees constructed from N random subsets of the training data, where each subset is of the fixed-size mentioned in the previous sentence. With Bagging, the original method from the literature (Breiman, 1996) of choosing a subset of points from a complete training set of N points was to choose a *bootstrap* sample (Efron, 1979). Simply put, this means randomly choosing N points with equal probability from the set with replacement, so that some points may be chosen more than once or not at all.

To compute prediction of an ensemble of trees for unseen data, the Random Forest model takes an average of predictions from individual trees. To estimate the prediction error of the bagged ensemble, predictions for each tree are computed on its out-of-bag observations, are averaged over the entire ensemble for each observation and then the predicted out-of-bag response is compared with the true value at this observation.

- Deep Neural Network, DNN

A deep neural network is an artificial neural network (ANN) with multiple hidden layers of units between the input and

output layers (Bengio, 2009; Schmidhuber, 2015). Similar to shallow ANNs, DNNs can model complex non-linear relationships. The extra layers enable composition of features from lower layers, giving the potential of modelling complex data with fewer units than a similarly performing shallow network. When performing supervised learning on a multi-class classification problem, common choices for the activation function and cost function are the softmax function and cross entropy function, respectively. Backpropagation and gradient descent have been the preferred method for training these structures due to the ease of implementation and their tendency to converge to better local optima in comparison with other training methods. Another training parameter to be considered with a DNN is the size (number of layers and number of units per layer), which has been empirically established.

### 2.4. *Performance measures*

After the classifiers were trained, the performance of the system is evaluated in terms of the following measures:

- The classification accuracy, computed as the percentage of correctly classified positive and negative instances.
- The sensitivity which quantifies the ability to correctly identify positive instances. It is the proportion of true positives that are correctly identified.
- The specificity which quantifies the ability to correctly identify negative instances. It is the proportion of true negatives that are correctly identified.

### 2.5. *Experimental procedure*

The experimental procedure is detailed as follows:

1. For the imputation methods, establish the parameters where necessary. The KNN uses the Euclidean distance and a number of 5 neighbours and the SOM adapts the number of map units to the data set size.
2. For each nonlinear classifier, establish its architecture. For the FNN a one hidden layer architecture with 40 units was chosen. For the Random Forest, the number of trees chosen was 15 and for the DNN, two hidden layers with 800 and 400 units respectively were used.
3. Take the whole data set and generate 10 different 10-fold cross validation sets in order to better estimate the true error rate of each model.
4. Train each model and obtain $10 \times 10$ performance measures over the validation sets.
5. Apply a Kruskal–Wallis test (Hollander & Wolfe, 1973) to check if there are significant differences among the means of the trained models for a level of significance $\gamma = 0.05$.
6. If there are differences among the medians, then apply a multiple comparison procedure (Hsu, 1996) to find the simplest model (lowest complexity) whose error is not significantly different from that of the model with the best mean accuracy rate. In this work, a Tukey's honestly significant criterion (Hsu, 1996) was used as multiple comparison test.
7. Apply the best model to the test set and obtain the final performance measures.

The experiments performed in this work were executed using the software tool Matlab (MATLAB, 2013).

## 3. Experimental results

In this section, the results obtained after applying missing data imputation techniques and several classifiers are shown and compared in terms of the effectiveness measures described in Section 2.4. To compare and study the convenience of imputing data, the reference model was first estimated by simply removing missing values from the original data set; this process is usually described as Listwise or case Deletion (LD). Then, the methods described in Section 2.2 were applied to input absent values, and the classification methods (Section 2.3) were used to predict the respiratory patterns.

### 3.1. *Training and validation data set results*

Table 3 shows the accuracy measures obtained by the selected models over a $10 \times 10$-fold cross validation for the respiratory pattern classification. These results are yield against the standard reference, i.e. the medical expert scores.

The LD method is improved by all the imputation methods for the Random Forest and the deep neural network. For the rest of the classifiers, LD does not improve imputation. The mean method offers better results than the hot-deck except for the neural network based classifiers. The reason for the mean method to be slightly better than the hot-deck method could be because using the mean/mode value for replacing missing values is more appropriate for the input variables than the HEOM distance. This distance is obtained taking into account all the variables of the example and it seems that not all of them are equally related. Finally, the SOM method is the best method only for the FNN.

If we analyse these results from the classifier point of view, the Random Forest gets the best accuracy result using the mean imputation method. The number of trees employed was 15. For the rest of the imputation methods, the FNN results—achieved with a 9-40-3 model— are better than the remainder of the classifiers. Several tests were made on the FNN architecture. For the mean imputation method, the best results were obtained with a 9-100-3 FNN, but the improvement over the 9-40-3 model was very small. Among the linear models tested (LDA and pSVM), the LDA performs better using any

**Table 3 – Respiratory pattern classification results. Mean validation set accuracy (%) of a $10 \times 10$-fold cv. Best values marked in bold font.**

| | LD | Mean | MLR | Hot-deck | KNN | SOM |
|---|---|---|---|---|---|---|
| LDA | 75.85 | 76.04 | 76.85 | 73.88 | 75.85 | 76.37 |
| pSVM | 71.85 | 74.21 | 74.83 | 72.76 | 71.73 | 74.76 |
| FNN | **80.04** | 81.19 | **81.13** | **80.43** | **80.00** | **81.17** |
| Classification tree | 73.38 | 76.77 | 71.72 | 74.82 | 73.36 | 75.83 |
| Random Forest | 79.54 | **81.43** | 79.76 | 79.56 | 79.63 | 80.80 |
| DNN | 78.61 | 79.03 | 80.39 | 79.63 | 78.77 | 80.38 |

imputation method. For the non-linear classifiers, between the decision tree models, the use of an ensemble improves the individual accuracy results significantly.

Besides validation against the standard reference comprising expert annotations, results from the presented approach are compared against the performance achieved by the expert system MIASOFT, previously developed by the authors. The accuracy of the results obtained by MIASOFT was 78.67% and the sensitivity and specificity results, for each respiratory pattern are shown in Table 4.

The results obtained for the different classifiers with the five imputation methods used were not better than the MIASOFT results in terms of accuracy except for the neural networks based classifiers and the Random Forest. For the deep neural network, any of the imputation methods outperforms the MIASOFT results, except the LD. Nevertheless, the FNN and the Random Forest outperform MIASOFT accuracy results no matter what imputation method is used and the Random Forest achieves the best accuracy results (81.43%) with the mean imputation method.

Analysing the balanced accuracy (mean of the sensitivity and specificity measures)—which is equivalent to the area under the ROC curve with one operation point ($AUC_1$)—against the MIASOFT results and over the three respiratory patterns, the following can be stated. For the apnoea pattern (Table 5a), the LDA performs better than MIASOFT no matter what imputation method was used. The non linear classifiers, except the classification tree, improve the MIASOFT results with all the imputation methods except the hot-deck and SOM. For the hypopnoea pattern (Table 5b), the Random Forest and the FNN models outperform the MIASOFT results with any of the imputation methods. Nevertheless, the lineal models are slightly worse than MIASOFT with any imputation method. Finally, for the normal breathing pattern (Table 5c) none of the classifiers with any of the imputation methods improve on the MIASOFT values.

Table 5 shows the Area Under ROC curve with one operation point value, obtained for each of the respiratory pattern.

To verify if the models are significantly different, a Kruscal–Wallis test was applied. Figure 1 shows the accuracy for each model using a box-whisker plot. In this figure, y-axis represents the classification accuracy and x-axis is formed by a duo indicating the imputation method and the classifier used, respectively. In order to rigorously select the final model, the Kruskal–Wallis test was applied to check if there are statistical differences among the mean validation accuracies. The p-value obtained was 0 for a significance level of 95%. Therefore, the null hypothesis (all means are equal) can clearly be rejected. Afterwards, a multiple comparison procedure was performed to make all-pairwise comparisons among each model.

Figure 2 graphically represents the comparison for those models whose mean accuracy is significantly different from the best, that is: Random Forest with mean imputation method. Those combinations whose interval is not crossing the dashed line are significantly different from the best model, therefore, can be discarded. There are nine models whose accuracy is not significantly different from the best model which are: the FNN with LD, mean, MLR, hot-deck and SOM imputation methods, Random Forest with SOM imputation method and DNN with MLR and SOM imputation methods. Therefore these are the models that were applied to the test data set.

**Table 4 – Respiratory pattern classification results for MIASOFT. Validation set sensitivity, specificity and Area Under ROC (Receiver Operating Characteristic) curve with one operation point ($AUC_1$) values (%).**

| | Apnoea | Hypopnoea | Normal respiration |
|---|---|---|---|
| Sensitivity | 81.22 | 64.65 | 85.73 |
| Specificity | 96.41 | 86.68 | 79.72 |
| $AUC_1$ | 88.81 | 75.66 | 87.72 |

**Table 5 – (a) Apnoea, (b) Hypopnoea and (c) Normal breathing classification results. Area under ROC curve with one operation point ($AUC_1$) values (%). Best values marked in bold font.**

| | $AUC_1$ | | | | | |
|---|---|---|---|---|---|---|
| | LD | Mean | MLR | Hot-deck | KNN | SOM |
| (a) | | | | | | |
| LDA | **90.55** | **90.67** | **89.98** | **89.04** | **90.55** | **90.68** |
| pSVM | 89.17 | 89.62 | 86.06 | 87.38 | 89.08 | 88.10 |
| FNN | 90.30 | 89.30 | 88.80 | 87.91 | 90.36 | 88.77 |
| Classification tree | 86.08 | 85.58 | 84.88 | 84.26 | 86.10 | 84.99 |
| Random Forest | 90.12 | 88.97 | 88.87 | 88.03 | 90.11 | 88.55 |
| DNN | 89.98 | 89.10 | 88.83 | 87.52 | 89.96 | 88.54 |
| (b) | | | | | | |
| LDA | 75.96 | 66.07 | 70.07 | 66.77 | 75.96 | 69.13 |
| pSVM | 72.49 | 62.85 | 66.33 | 63.27 | 72.37 | 66.08 |
| FNN | **80.39** | 77.04 | **77.79** | **76.89** | **80.35** | **77.63** |
| Classification tree | 74.00 | 73.22 | 70.40 | 71.72 | 73.98 | 72.42 |
| Random Forest | 79.91 | **77.96** | 76.86 | 76.46 | 80.00 | 77.55 |
| DNN | 79.01 | 73.16 | 76.72 | 75.85 | 79.16 | 76.63 |
| (c) | | | | | | |
| LDA | 78.70 | 78.51 | 81.28 | 77.76 | 78.70 | 80.48 |
| pSVM | 68.53 | 76.01 | 79.17 | 75.47 | 68.39 | 78.57 |
| FNN | **80.54** | 84.14 | **84.35** | **83.73** | **80.54** | **84.43** |
| Classification tree | 76.82 | 81.27 | 76.97 | 79.66 | 76.77 | 80.47 |
| Random Forest | 80.00 | **84.56** | 83.22 | 83.16 | 80.18 | 84.20 |
| DNN | 78.35 | 81.80 | 83.85 | 83.13 | 78.76 | 83.90 |

### 3.2. *Test data set results*

Once the best models in terms of accuracy have been selected, the results over the test data set were obtained and compared against the standard reference and the expert system MIASOFT. Tables 6 and 7 show the classification results in terms of the effectiveness measures described in Section 2.4.

The results obtained were slightly better than the MIASOFT results, with the FNN with the mean imputation method being the best model in terms of accuracy. As has been stated in the previous subsection, the results of these models are not statistically different which is confirmed with the accuracy values over the test set. Nevertheless, it is surprising that the model which has achieved the best accuracy value in the validation set, is now one of the models that offers the lowest accuracy value.

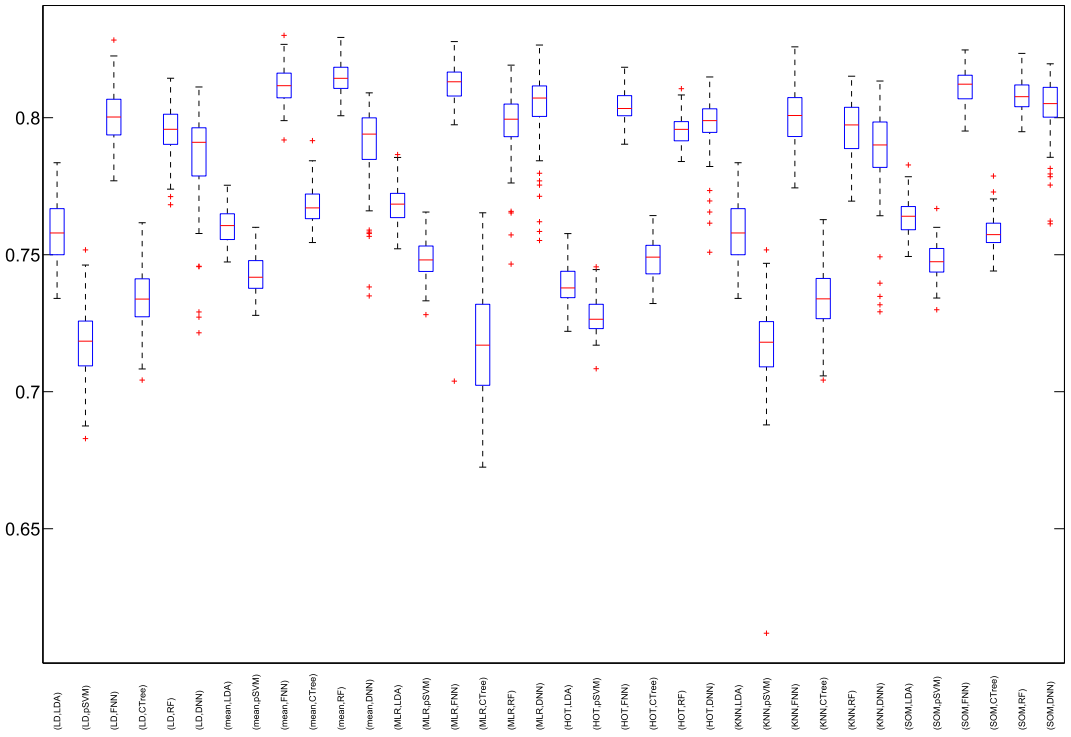


**Fig. 1 – Box-whiskers plots for the validation data using a 10-fold cross validation and 10 different experiments. CTree = Classification Tree, Hot = hot-deck, RF = Random Forest.**

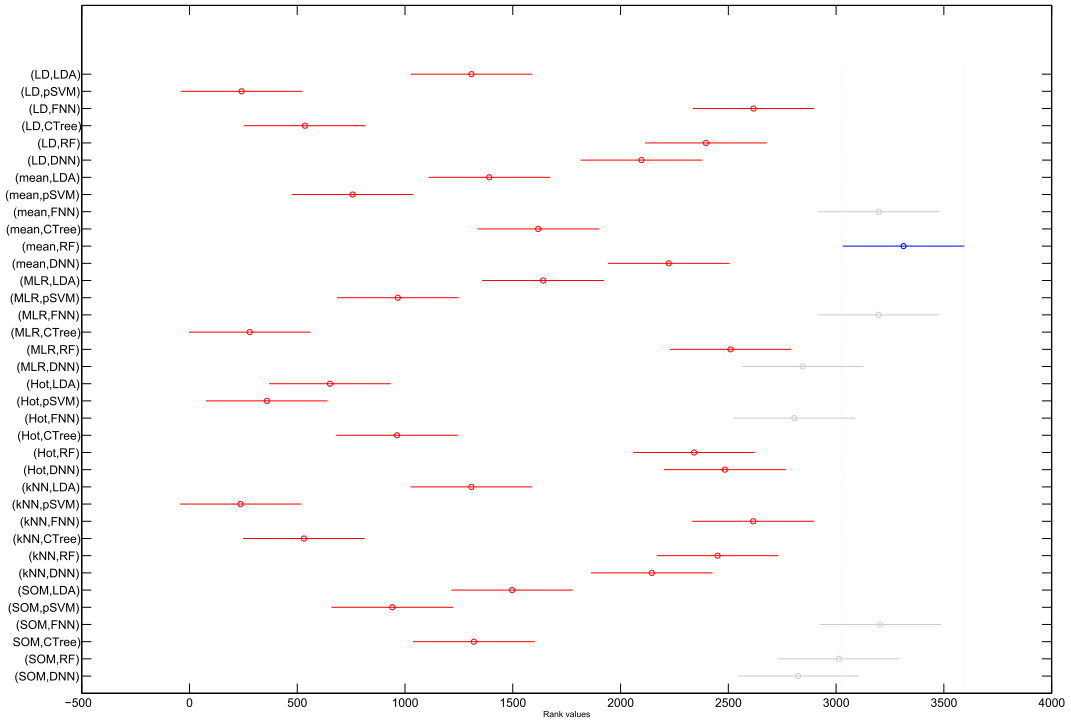


**Fig. 2 – Multiple comparison procedure plot (the best model is marked). CTree = Classification Tree, Hot = hot-deck, RF = Random Forest.**

**Table 6 – Respiratory pattern classification results. Test set accuracy values (%).**

| | Accuracy (%) |
|---|---|
| MIASOFT | 76.60 |
| Random Forest + mean | 72.76 |
| Random Forest + SOM | 77.43 |
| FNN + LD | 71.54 |
| FNN + mean | 80.14 |
| FNN + MLR | 79.67 |
| FNN + Hot-deck | 79.05 |
| FNN + SOM | 79.03 |
| DNN + MLR | 79.77 |
| DNN + SOM | 79.26 |

Taking into account the $AUC_1$ measure, and analysing the results for each respiratory pattern, the following can be expounded. For the apnoea pattern, only the Random Forest with the mean method outperforms MIASOFT results. The rest of the models do not offer good results with the test set. For the hypopnoea and normal breathing patterns, all of the models except Random Forest with the mean imputation method, outperform MIASOFT results. In this case, the FNN with the MLR imputation method is the best one.

**Table 7 – (a) Apnoea, (b) Hypopnoea and (c) Normal breathing performance measurements values (%) over the test set. Best $AUC_1$ values marked in bold font.**

| | Sensitivity | Specificity | $AUC_1$ |
|---|---|---|---|
| (a) | | | |
| MIASOFT | 69.32 | 97.53 | 83.43 |
| Random Forest + mean | 63.08 | 98.72 | 80.90 |
| Random Forest + SOM | 62.86 | 98.11 | 80.49 |
| FNN + LD | 74.00 | 96.12 | **85.06** |
| FNN + mean | 65.70 | 98.80 | 82.25 |
| FNN + MLR | 65.87 | 98.58 | 82.23 |
| FNN + Hot-deck | 63.53 | 98.09 | 80.81 |
| FNN + SOM | 63.42 | 98.35 | 80.89 |
| DNN + MLR | 66.31 | 98.59 | 82.45 |
| DNN + SOM | 63.75 | 98.32 | 81.04 |
| (b) | | | |
| MIASOFT | 56.56 | 84.73 | 70.65 |
| Random Forest + mean | 63.32 | 76.79 | 70.06 |
| Random Forest + SOM | 58.18 | 85.40 | 71.79 |
| FNN + LD | 77.49 | 68.32 | 72.90 |
| FNN + mean | 51.56 | 90.54 | 71.05 |
| FNN + MLR | 58.47 | 87.82 | **73.14** |
| FNN + Hot-deck | 53.80 | 89.16 | 71.48 |
| FNN + SOM | 57.50 | 87.53 | 72.52 |
| DNN + MLR | 56.41 | 88.51 | 72.46 |
| DNN + SOM | 56.81 | 87.96 | 72.38 |
| (c) | | | |
| MIASOFT | 84.23 | 68.96 | 76.60 |
| Random Forest + mean | 76.94 | 72.45 | 74.70 |
| Random Forest + SOM | 85.48 | 68.24 | 76.86 |
| FNN + LD | 63.57 | 87.84 | 75.70 |
| FNN + mean | 91.41 | 62.23 | 76.82 |
| FNN + MLR | 88.33 | 68.12 | **78.22** |
| FNN + Hot-deck | 89.27 | 64.28 | 76.78 |
| FNN + SOM | 87.97 | 67.44 | 77.70 |
| DNN + MLR | 89.14 | 66.71 | 77.93 |
| DNN + SOM | 88.50 | 67.24 | 77.87 |

## 4. Discussion and conclusions

This paper presents a comparative study of the respiratory pattern classification task involving five missing data imputation techniques, and six different machine learning algorithms. The main goal was to find a classifier that achieves the most accurate results using a scalable imputation method in comparison to the method used by MIASOFT. As we pointed out, in contrast to the data-driven approach followed in this work, MIASOFT is more knowledge-based, and it has been designed to allow explanatory capabilities for their results. But for the respiratory pattern classification task, the developed approach seems to be slightly better.

The imputation techniques include three statistical methods – mean, multiple linear regression and hot-deck– and two machine learning methods – K nearest neighbours (KNN) and self-organising maps (SOM). These techniques were compared with the listwise deletion method and the results show the danger of eliminating records with missing values from the original data set. Such deletion can introduce substantial biases in the study. Once the unknown data were imputed, a classification model was created comparing two linear models, linear discriminant analysis (LDA) and proximal support vector machine (pSVM), and four non linear ones, a feedforward neural network (FNN), a classification tree, a Random Forest and a deep neural network (DNN).

The results obtained show that in general, the SOM imputation method allows non-tree based classifiers to achieve improvements over the rest of the imputation methods in terms of the classification accuracy. For this imputation method, the FNN provides the best result. From the classifier point of view, the FNN model offers the best performance except for the mean imputation method where the Random Forest model achieves the best result. It seems that linear classification methods are less appropriate for the respiratory pattern classification. At this point and taking into account the model comparison carried out, it seems that the FNN is a good solution no matter what imputation method used. So a deeper study must be undertaken into FNN architectures. The DNN provides less good results than the FNN with a more complex architecture and training procedure, so the benefits of using this powerful model are limited for this particular study. Besides, taking into account the promising results with the Random Forest model, it seems that a combination of classification models offers better performance than the individual ones.

Nevertheless, the results obtained in terms of accuracy are not as good as expected. The improvements over MIASOFT results are limited so a deeper study might be done. It would be desirable to analyse the relationship between the input variables used in this work by means of the use of feature selection methods. Although these methods are commonly applied in data sets with a large number of variables, they offer potential benefits such as reducing training and utilisation times and defying the curse of dimensionality to improve prediction performance. Besides, after applying

feature selection, more complex missing data imputation methods could be used.

We conclude that machine learning techniques may be a better approach to imputing missing values than statistical methods, as they led to improvements in the prediction accuracy of the classifiers, as has been demonstrated in the SAHS diagnosis field. Imputation techniques depend on the available data and the prediction method used; thus, the results obtained might not generalise to different data sets.

## Acknowledgements

We would like to thank members of Case Western University and the Sleep Health Heart Study for their support with the PSG recordings database. The opinions expressed in the paper are those of the authors and do not necessarily reflect the views or the endorsement of the Sleep Heart Health Study. This research was partially funded by the Spanish Ministerio de Economía y Competitividad under project TIN 2013-40686-P, partially supported by the European Union ERDF, and by the Xunta de Galicia under project GRC2014/35.

## REFERENCES

Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In D. Banks, F. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, clustering, and data mining applications studies in classification, data analysis, and knowledge organisation* (pp. 639–647). Berlin Heidelberg: Springer.

Aha, D., Kibler, D., & Albert, M. (1991). Instance-based learning algorithms. *Machine Learning, 6*, 37–66.

Álvarez Estévez, D. (2012). *Intelligent diagnosis of the sleep apnea-hypopnea syndrome: A comprehensive approach through an intelligent system to support medical decision*. Ph.D. thesis. Spain: Department of Computer Science, University of A Coruna.

Álvarez Estévez, D., & Moret-Bonillo, V. (2009). Fuzzy reasoning used to detect apneic events in the sleep apnea-hypopnea syndrome. *Expert Systems with Applications, 36*, 7778–7785.

Álvarez-Estévez, D., Sánchez-Maroño, N., Alonso-Betanzos, A., & Moret-Bonillo, V. (2011). Reducing dimensionality in a database of sleep EEG arousals. *Expert Systems with Applications, 38*, 7746–7754.

Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence, 17*, 519–533.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning, 2*, 1–127.

Berry, R., Brooks, R., Gamaldo, C., Harding, S., Marcus, C., Vaughn, B., et al. (2012). *The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications, Version 2.0*. Darienm, Illinois: American Academy of Sleep Medicine.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. New York: Chapman & Hall.

Bystricky, W., & Safer, A. (2004). *Identification of individual sleep apnea events from the ECG using neural networks and a dynamic Markovian state model* (pp. 297–308).

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on, 13*, 21–27.

Dahl, F. A. (2007). Convergence of random k-nearest-neighbour imputation. *Computational Statistics & Data Analysis, 51*, 5913–5917.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7*, 1–26.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179–188.

Fontenla-Romero, O., Guijarro-Berdiñas, B., Alonso-Betanzos, A., & Moret-Bonillo, V. (2005). A new method for sleep apnea classification using wavelets and feedforward neural networks. *Artificial Intelligence in Medicine, 34*, 65–76.

Fujikawa, Y. (2001). *Efficient algorithms for dealing with missing values in knowledge discovery*. Master degree thesis. Japan Advanced Institute of Science and Technology.

Fung, G., & Mangasarian, O. (2001). Proximal support vector machine classifiers. In F. Provost, & R. Srikant (Eds.), *Proceedings KDD-2001: Knowledge discovery and data mining*. New York: Association for Computing Machinery.

Garca-Laencina, P. J., Sancho-Gmez, J.-L., & Figueiras-Vidal, A. R. (2013). Classifying patterns with missing values using multi-task learning perceptrons. *Expert Systems with Applications, 40*, 1333–1341.

Garca-Laencina, P. J., Sancho-Gmez, J.-L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing, 72*, 1483–1493.

Hollander, M., & Wolfe, D. (1973). *Nonparametric statistical methods*. New York: John Wiley.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359–366.

Hsu, J. (1996). *Multiple comparisons: Theory and methods*. Chapman & Hall/CRC.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Springer-Verlag.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley & Sons.

Maali, Y., & Al-Jumaily, A. (2012). Genetic fuzzy approach based sleep apnea/hypopnea detection. *International Journal of Machine Learning and Computing, 2*, 685–688.

MATLAB. (2013). *Version 8.1.0.604 (R2013a)*. Natick, Massachusetts: The MathWorks Inc.

Mitchell, T. M. (1997). *Machine learning* (1st ed.). New York, NY, USA: McGraw-Hill, Inc.

Mitra, M., & Samanta, R. (2015). Hepatitis disease diagnosis using multiple imputation and neural network with rough set feature reduction. In S. C. Satapathy, B. N. Biswal, S. K. Udgata, & J. Mandal (Eds.), *Volume 327 of advances in intelligent systems and computingProceedings of the 3rd International Conference on Frontiers of intelligent Computing: Theory and Applications (FICTA) 2014* (pp. 285–293). Springer International Publishing.

Moller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks, 6*, 525–533.

Moret-Bonillo, V., Álvarez-Estévez, D., Fernández-Leal, A., & Hernández-Pereira, E. (2014). Intelligent approach for analysis of respiratory signals and oxygen saturation in the sleep apnea-hypopnea syndrome. *The Open Medical Informatics Journal*, 1–19.

Polat, K., Yosunkaya, S., & Gunes, S. (2008). Comparison of different classifier algorithms on the automated detection of obstructive sleep apnea syndrome. *Journal of Medical Systems, 32*, 243–250.

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., et al. (1997). The sleep heart health study: design, rationale, and methods. *Sleep, 20*, 1077–1085.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.
Rahman, M., & Davis, D. (2013). Machine learning-based missing value imputation method for clinical datasets. In G.-C. Yang, S.-l. Ao, & L. Gelman (Eds.), *Volume 229 of lecture notes in electrical engineeringIAENG transactions on engineering technologies* (pp. 245–257). Netherlands: Springer.
Ribelles, N., Martin, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine, 50*, 105–115.
Ritthipravat, P., Kumdee, O., & Bhongmakapat, T. (2013). Efficient missing data technique for prediction of nasopharyngeal carcinoma recurrence. *Information Technology Journal, 12*, 1125–1133.
Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks, 61*, 85–117.
Sleep Health Heart Study, S. (2002). *Sleep health heart study. Reading center manual of operations*. Technical Report Case Western Reserve University, VMLA-039–02.
Srivastava, M. S., & Carter, E. M. (1983). *Applied multivariate statistics*. Amsterdam: North Holland.
Tian, J., & Liu, J. (2005). Apnea detection based on time delay neural network. In *27th IEEE EMB Conference* (pp. 2571–2574).
Várady, P., Micsik, T., Benedek, S., & Benyó, Z. (2002). A novel method for the detection of apnea and hypopnea events in respiration signals. *IEEE Transactions on Biomedical Engineering, 49*, 936–942.
Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). *SOM toolbox for Matlab 5*. Technical Report A57 SOM Toolbox Team. Helsinki University of Technology.
Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research, 6*, 1–34.

* Corresponding author. e-mail address: elena.hernandez@udc.es (E.M. Hernández-Pereira).