Success/Failure Prediction of Noninvasive Mechanical Ventilation in Intensive Care Units*

Using Multiclassifiers and Feature Selection Methods

Félix Martín-González1; Javier González-Robledo1; Fernando Sánchez-Hernández2; María N. Moreno-García3

1Intensive Care Unit, University Hospital of Salamanca, Salamanca, Spain; 2School of Nursing and Physiotherapy, University of Salamanca, Prehospital Emergency Services, Salamanca, Spain; 3Department of Computing and Automation, University of Salamanca, Salamanca, Spain

Keywords Noninvasive ventilation, respiration disorders, respiratory insufficiency, data mining, feature selection methods, classifiers, multiclassifiers

Summary

Objectives: This paper addresses the problem of decision-making in relation to the administration of noninvasive mechanical ventilation (NIMV) in intensive care units.

Methods: Data mining methods were employed to find out the factors influencing the success/failure of NIMV and to predict its results in future patients. These artificial intelligence-based methods have not been applied in this field in spite of the good results obtained in other medical areas.

Results: Feature selection methods provided the most influential variables in the success/failure of NIMV, such as NIMV hours, PaCO
d at the start, PaO
d/FiO
t o ratio at the start, hemocrit at the start or PaO
t o/FiO
t o ratio after two hours. These methods were also used in the preprocessing step with the aim of improving the results of the classifiers. The algorithms provided the best results when the dataset used as input was the one containing the attributes selected with the CFS method.

Conclusions: Data mining methods can be successfully applied to determine the most influential factors in the success/failure of NIMV and also to predict NIMV results in future patients. The results provided by classifiers can be improved by preprocessing the data with feature selection techniques.

Correspondence to:
Maria N. Moreno-García
University of Salamanca
Department of Computing and Automation
Plaza de los Caídos s/n
37008 Salamanca
Spain
E-mail: mmg@usal.es

* Supplementary material published on our website http://dx.doi.org/10.3414/ME14-01-0015

http://dx.doi.org/10.3414/ME14-01-0015
received: January 28, 2014
accepted: March 18, 2015
epub ahead of print: April 30, 2015

1. Introduction

1.1 Noninvasive Mechanical Ventilation

Noninvasive mechanical ventilation (NIMV) exists since the creation of the first negative pressure respirators in the early nineteenth century. Currently positive pressure NIMV is the most widely used system, being the use of other types practically marginal. NIMV with positive pressure uses an oro-facial mask to transmit positive airway pressure without need of orotracheal intubation of the patient.

The introduction of NIMV in the treatment of patients with acute respiratory failure represents one of the great advances in the care of the critical patients as well as a substantial change in the initial way multiple pathologies are addressed. The search for the least aggressive measures is one of the principal characteristics of medical research. It is evident that the treatment of critical patients is usually very aggressive and requires a multitude of measures to keep the patient alive and to monitor the treatment correctly; however, these measures are sometimes responsible for complications that can be fatal. For the last three decades modern NIMV has allowed an alternative treatment in patients who previously could only be handled by orotracheal intubation (OTI) and connection to invasive mechanical ventilation (IMV).

Nevertheless, one of the risks of the less invasive technologies is the delay in the application of measures that may be necessary. In the case of respiratory failure (RF), the delay in the application of OTI and IMV in patients who are not going to improve with NIMV increases their mortality. The identification and correct selection of patients who respond to the treatment therefore poses an important challenge in the application of NIMV. However, one of the major difficulties in Intensive Care Medicine that makes it complicated to obtain significant results is the great heterogeneity of the population. The patients treated in intensive care units are diverse: there are both surgical and medical types of patients; they have multiple pathologies, and many of them need NIMV.

Usually, the data gathered from these patients are processed by means of statistic-
1.2 Data Mining Techniques

Data mining includes a large variety of algorithms (clustering, association rules, decision trees, multiclassifiers, etc.) that can be applied for knowledge discovery and prediction. These methods allow us to manage huge volumes of data in order to induce patterns, discover trends, find similarities, etc. that can be useful for different purposes. In the biomedical field they have been widely applied [1] for diagnosis, processing of medical signals, image processing, patient monitoring, analysis of survival data [2] and analysis of causes of death [3], among other things. Both health care and biomedicine currently operate in data-intensive environments; therefore, a multidisciplinary approach is necessary for suitable information processing that will lead to effective decision-making [4]. A recent study [5] proves the high demand of knowledge-based systems in medicine and points out pharmaco-vigilance, intensive care monitoring, and support for guidelines and clinical pathways as the most promising fields for their application. Most recently medical informatics research domains have been extended to innovative aspects of computer science as data mining, usage, evaluation, or visualization [6].

Data mining techniques have been successfully used to infer knowledge in very diverse medical areas; however, in spite of their great interest and promising use in the aforementioned fields, these methods have not yet been exploited in the specific domain of the treatment of respiratory diseases and NIMV administration in intensive care units.

2. Objectives

The aim of the study presented in this paper was to identify early factors influencing the success/failure of NIMV and to build models for automatically predicting results of NIMV in future patients by means of data mining techniques. Here we used multiclassifiers, a relatively recent category of algorithms that usually provide better results than other traditional data mining methods.

3. Background

Mechanical ventilation (MV) refers to any mechanical procedure used for totally or partially replacing the need of air flow to the lungs [7].

Invasive mechanical ventilation (IMV) is the kind of MV that requires an endotracheal device (usually an orotracheal tube or tracheotomy) for its administration, unlike NIMV, which does not need a tracheal tube. This includes both ventilation with positive pressure and with negative pressure as well as other breathing support systems that nowadays are incidental [8].

IMV is essential in a great number of patients, and it is in charge of keeping them alive until the pathology that made its use necessary has been overcome. But we have known for a long time that IMV is not innocuous. In spite of the great advances achieved over the last 100 years, which have allowed complications to be reduced, they are still present. These complications, such as airway injuries or infections such as sinusitis and ventilator associated pneumonia, as well as the frequent need for deep sedation for its administration, led to a progressive improvement in NIMV throughout the entire 20th century. The benefit regarding quality of life, survival and improvement of respiratory failure demonstrated its great usefulness. But in particular it was the success achieved in the treatment of obstructive sleep apnea syndrome by means of continuous positive airway pressure that gave the definitive impulse to NIMV. In later years its use gradually spread to patients with acute respiratory failure (ARF). In the last three decades numerous studies have demonstrated its usefulness in the reduction of orotracheal intubation, mortality rate and the length of hospital stay, increasing the number of pathologies capable of being treated in this way.

However, applying NIMV in patients who are not going to improve with it and who really need IMV, with the consequent delay in OTI, can heighten the prognosis [9]. Some authors [10, 11] have provided certain indications concerning the clinical/physiological parameters of NIMV (hypoxemia, dyspnea, ventilation failure, tachypnea...) and pathologies (chronic obstructive pulmonary disease, postextubation RF, hypoxemic RF...); however, the level of evidence is very different depending on the pathologies. Whereas in acute exacerbations of chronic obstructive pulmonary disease (COPD) the evidence is overwhelming [12], in other types of pathologies there is just a series of documented cases and poor quality studies. Other kinds of studies have focused on factors influencing the results of NIMV in specific diseases such as acute heart failure [13] or chest wall disease [14]. All of them use classical statistical techniques but data mining methods have not been applied in this field with the single exception of one study [15] in which a fuzzy clustering method was used but with a different purpose: the estimation of respiratory parameters and the dependence of the expiratory time constant on the volume in patients with and without COPD.

4. Data Mining Methods

In this work two categories of data mining algorithms were used: feature selection techniques and classification methods. The former were necessary in order to select the most influential attributes in the success/failure of NIMV from the large number of attributes available, before applying the classification algorithms. The latter were used for inducing the models that allow us to predict NIMV results in future patients. A complete description of the methods used in this work is provided in the Appendix III.

4.1 Feature Selection Techniques

Feature selection techniques are widely applied in both statistic and data mining fields. While methods as lasso, and elastic nets have proved its effectiveness in linear and logistic regression [16], particularly when the number of predictor variables is greater than the sample size, in the field of machine learning other kind of method are used. In this field, well-known algorithms
whose efficacy has been demonstrated [17] are CFS (Correlation-based Feature Subset Selection) [18] and methods based on the information gain (IG) with respect to the class [19]. There are other strategies, called wrappers, which use specific induction algorithms to estimate the importance of feature subsets [18]. Wrappers methods have not been considered to select attributes given that these methods have a high computational cost and they must be designed for each particular learning algorithm, which complicates the process. In this work CFS and IG methods have been applied due to the good balance between simplicity and performance showed by them. CFS evaluates the significance of a subset of features (attributes), taking into account the individual predictive ability of each feature and the degree of redundancy between them. This method selects the subsets of attributes that are highly correlated with the class while having low inter-correlation between them.

The information gain based methods evaluate the importance of an attribute by obtaining the information gain (IG) with respect to the class. The attributes providing the most IG are the most relevant and therefore the most influential in the classification.

4.2 Classification Methods

Several classification algorithms were applied to induce models for predicting the class of non-classified patients from other attributes known about them. In this study the classes are the "success" and "failure" of NIMV, and the models will thus allow us to predict the results of NIMV in patients before and during its administration and, consequently, to decide whether it is suitable for these patients or not. Although the extensive description of the methods is provided in the Appendix III, in the following subsections we summarize the foundations of the algorithms involved in the study.

4.2.1 Decision Trees

A decision tree consists of a set of conditions that are organized in a hierarchical structure, so that the final decision can be determined following the true conditions from the root to the leaves of the tree. In this study, two well known and generally used tree induction algorithms were employed, J48 and REPTree. J48 is an advanced version of C4.5 [20], one of the most used and well-known algorithms. J48 is an information gain-based method with pruning procedures that uses rules. REPTree is a fast decision tree learning algorithm, also based on information gain, which uses reduced-error pruning with back fitting. Given that the enclosed procedure and the performance of these methods are different, the application of both of them has been carried out in order to find out which of them provides the best results, when used as single classifiers as well as when used as base classifiers of multiclassifiers.

4.2.2 Bayesian Networks

Bayesian networks are structures that relate variables and that include distributions of conditional probabilities associated with these variables. The learning process for a dataset lies in finding, among all possible graphs, the graph that best represents the set of dependencies/independencies between data. The problem does not have an exact solution and it is necessary to resort to heuristic search methods. The K2 search algorithm [21] was applied in this study.

4.2.3 k-Nearest Neighbors

The classification of an item by means of this algorithm is the result of the majority vote of its neighbors, that is, the class most frequent among its k-nearest neighbors [22]. The search of the neighbors is carried out according to some distance metric such as euclidean, manhattan, etc.

4.2.4 Multiclassifiers

Multiclassifiers are methods that combine several individual classifiers induced with different basic methods or obtained from different training datasets with the purpose of improving the precision of the predictions. Another additional advantage of these techniques is the reduction of the overfitting problem, which takes place when the learning process finds a regularity in the data that is distinctive of the training set but cannot be extended to other datasets.

The methods for building multiclassifiers are divided into two groups. The first methods, such as Bagging [23], Boosting [24] and Random Forest [25], induce models that merge classifiers with the same learning algorithm, but introducing modifications in the training data set. The second type of methods, named hybrids, such as Stacking [26] and Cascading [27], create new hybrid learning techniques from different base learning algorithms.

In this study the methods used are Bagging, Random Forest and AdaBoost. The latter is a variant of the Boosting method.

5. Data Description

The study was carried out with data from all patients hospitalized in the Intensive Care Unit (ICU) of the University Hospital of Salamanca from 2006 to 2011 who were treated with NIMV (both kinds: BPAP – bi-level positive airway pressure, and CPAP – continuous positive airway pressure). During this time interval there were 4661 hospitalizations of 4249 patients. The study comprised 389 patients who received treatment with NIMV, some of them on several occasions, representing a total of 410 cases with a mean age of 66.69 ± 13.38 years. The most common cause of use of NIMV was acute hypoxemic RF (38.05%), followed by postextubation RF (21.22%) and by re-worsening of chronic obstructive pulmonary disease (19.02%). The cause of the smallest group was acute pulmonary edema (6.83%).

The information was gathered from the clinical records of these patients stored in the database of the University Hospital of Salamanca. This is one of the four great reference hospitals in Spain according to TOP 20 benchmark report of the year 2012 (http://www.iasist.com/es/top-20/top-20-2012).

A great number of variables were analyzed, ranging from demographic data to physiological, gasometrical, biochemical, hematologic or ventilatory variables, among others. A total number of 93 vari-
variables were used in the study as input attributes to the algorithms. A detailed description of them can be found in Appendix I.

The involved variables were of several types (binary, categorical and numeric). They were not transformed since the algorithms used are able to manage all types of input data. However, in the preprocessing step some variables that do not influence the classification (dates, patient number ... and class dependent variables (place of death, cause of death ...) were eliminated.

Appendix I also includes statistical values of some of these variables as well as the value distribution of some important variables in relation to the two classes.

6. Data Mining Study and Results

Several data mining algorithms were applied to the described dataset containing 410 examples and 93 attributes. In the study the label to be predicted (that is, the class attribute) is the attribute which takes the “success” or “failure” values. The dataset contains 202 examples with the “success” value for the label and 208 examples with the “failure” value. NIMV was considered successful when there was no need of MV in the 72 hours following its removal.

The large number of attributes in the dataset may cause a loss of accuracy in the classifiers induced by the data mining algorithms. This problem of high dimensionality was addressed by means of feature selection methods. First, the two feature selection methods described in section 4.1, CFS and an IG based method, were applied in order to determine the most influential attributes for the classification. A threshold value was established for the metrics of each one of these methods. In the case of IG the chosen attributes are those with a value of information gain greater than 0.01, and for the CFS method we have settled to 5 the number of consecutive non-improving nodes of the search process. A forward best first search was used for CFS since this search strategy performs slightly better than forward selection, and similarly to backward elimination.

The attributes, shown in Appendix II, were selected to be used in the next step, which involves the application of the classification algorithms in order to build models for predicting the class of unclassified patients.

Several classification algorithms were applied; however, only those that provided the best results are reported in the paper. These algorithms were two decision trees (J48 and REPTree), Bayes Net, k-Nearest Neighbors, Random forest and the multiclassifiers, Bagging and Adaboost, with J48 and REPTree as base classifiers in both of them.

The confidence factor set to decision tree was 0.25. Although higher values of confidence allow to improve the accuracy by about 1%, they lead to a low pruning and can cause overfitting problems. The search algorithms used for Bayes Net were K2 and TAN, providing both the same results. Regarding the number of neighbors used in KNN algorithm we chose those providing the best accuracy values, which were 21 for the dataset with all the attributes, 5 for the dataset with CFS selected attributes and 9 for the dataset with IG selected attributes. The number of iterations of the multiclassifiers were set analogously in order to obtain the best results. The highest accuracy values were obtained with a number of iterations between 30 and 50 with hardly impact on the time taken to build the models. The configuration parameters for Random Forest are maximum depth of the trees, number of attributes and number of trees. After an exhaustive study of the impact of the setting values on the results we found out that unlimited depth of trees yielded the best results. On the other hand, the maximum difference of accuracy values by varying the last two parameters for runs using the same seed was 1.9%, however the changes in the precision are not in the same direction for different runs using different seeds, so that, after multiple runs these differences are compensated. Thus we decided not to limit the two first parameters and to set to 75 the number of trees since it is the number that resulted in a better behavior in most of the conducted experiments.

The study was carried out using, as input of every classification algorithm, first the dataset containing all the attributes,

\[
\begin{align*}
\text{IF VMNI hours} \leq 2 & \quad \text{balance after 12 hours} \leq 1000 \quad \text{THEN SUCCESS} \\
& \quad \text{balance after 12 hours} > 1000 \\
& \quad \text{RR change} \leq 4 \\
& \quad \text{VMNI hours} \leq 1 \quad \text{THEN SUCCESS} \\
& \quad \text{VMNI hours} > 1 \quad \text{THEN FAILURE} \\
& \quad \text{RR change} > 4 \quad \text{THEN FAILURE} \\
\text{IF VMNI hours} > 2 & \quad \text{RR2} \leq 27 \\
& \quad \text{sedation} = \text{NO} \\
& \quad \text{balance after 24 hours} \leq 900 \quad \text{THEN SUCCESS} \\
& \quad \text{balance after 24 hours} > 900 \\
& \quad \text{hematocrit} \leq 34.1 \\
& \quad \text{RR2} \leq 21 \\
& \quad \text{PAFI2} \leq 131 \quad \text{THEN FAILURE} \\
& \quad \text{PAFI2} > 131 \\
& \quad \text{RR change} \leq 3 \quad > 1 \quad \text{THEN SUCCESS}
\end{align*}
\]

Figure 1 An example of rules induced by the tree based algorithms, where PAFI2 is PaO\textsubscript{2}/FiO\textsubscript{2} ratio after two hours, RR2 is respiratory rate in breaths a minute two hours after admission and RR change is the change of respiration rate two hours after admission. PaO\textsubscript{2} is measured in mmHg and FiO\textsubscript{2} in percentage.
then the one containing the attributes selected by the CFS method and finally the dataset with the attributes selected by the Information Gain method. As commented before, in the first case the dataset have 93 attributes. After applying the CFS method the number of attributes was reduced to 17 and the application of IG method allowed us to select the best 44 attributes (the most significant are provided in Appendix II). In all experiments carried out with the classification algorithms 10-fold cross validation was applied. The generated models cannot be showed since they are very large in spite of setting smaller values of dimensionality aiming at improving the results of the classifiers; however, they also provided important information about the effectiveness of the feature selection methods, the values of accuracy (percentage of correctly classified instances) provided by the applied algorithms were represented in Figure 2. The comparison between the datasets obtained by different feature selection methods can be observed in this graph. Additionally, a pair-wise comparison of the classifiers using a paired t-test was performed with the aim of evaluating the statistical significance of the results. First, the test was applied to compare the accuracy of every algorithm for the three datasets (Table 2) and then to determine the best algorithms for the dataset providing the best results of accuracy (Table 2). The established level of significance (p) was 0.05.

### Table 1  Quality metrics of the classifiers built from the three datasets

<table>
<thead>
<tr>
<th>All attributes</th>
<th>TPR (sensitivity)</th>
<th>FPR (specificity)</th>
<th>PPV</th>
<th>NPV</th>
<th>ROC area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision tree J48</td>
<td>0.71±0.12</td>
<td>0.33±0.10</td>
<td>0.69±0.07</td>
<td>0.69±0.06</td>
<td>0.73±0.09</td>
</tr>
<tr>
<td>REPTree</td>
<td>0.69±0.12</td>
<td>0.31±0.10</td>
<td>0.70±0.07</td>
<td>0.69±0.06</td>
<td>0.71±0.07</td>
</tr>
<tr>
<td>Bayes net</td>
<td>0.77±0.10</td>
<td>0.34±0.09</td>
<td>0.70±0.07</td>
<td>0.74±0.07</td>
<td>0.80±0.06</td>
</tr>
<tr>
<td>k-nearest neighbors</td>
<td>0.80±0.15</td>
<td>0.45±0.12</td>
<td>0.65±0.09</td>
<td>0.73±0.06</td>
<td>0.79±0.08</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.78±0.09</td>
<td>0.28±0.09</td>
<td>0.74±0.08</td>
<td>0.76±0.08</td>
<td>0.81±0.05</td>
</tr>
<tr>
<td>Bagging with J48</td>
<td>0.76±0.10</td>
<td>0.29±0.09</td>
<td>0.73±0.08</td>
<td>0.75±0.08</td>
<td>0.80±0.06</td>
</tr>
<tr>
<td>Bagging with REPTree</td>
<td>0.76±0.11</td>
<td>0.27±0.09</td>
<td>0.75±0.07</td>
<td>0.75±0.06</td>
<td>0.83±0.06</td>
</tr>
<tr>
<td>AdaBoost with J48</td>
<td>0.77±0.11</td>
<td>0.30±0.10</td>
<td>0.72±0.08</td>
<td>0.75±0.07</td>
<td>0.81±0.06</td>
</tr>
<tr>
<td>AdaBoost with REPTree</td>
<td>0.71±0.10</td>
<td>0.28±0.09</td>
<td>0.72±0.08</td>
<td>0.71±0.07</td>
<td>0.78±0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attributes selected by CFS</th>
<th>TPR (sensitivity)</th>
<th>FPR (specificity)</th>
<th>PPV</th>
<th>NPV</th>
<th>ROC area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision tree J48</td>
<td>0.67±0.11</td>
<td>0.24±0.10</td>
<td>0.74±0.07</td>
<td>0.69±0.06</td>
<td>0.75±0.08</td>
</tr>
<tr>
<td>REPTree</td>
<td>0.72±0.12</td>
<td>0.27±0.11</td>
<td>0.73±0.08</td>
<td>0.72±0.07</td>
<td>0.76±0.07</td>
</tr>
<tr>
<td>Bayes net</td>
<td>0.74±0.09</td>
<td>0.23±0.09</td>
<td>0.77±0.07</td>
<td>0.74±0.07</td>
<td>0.82±0.06</td>
</tr>
<tr>
<td>k-nearest neighbors</td>
<td>0.60±0.08</td>
<td>0.13±0.10</td>
<td>0.82±0.06</td>
<td>0.68±0.05</td>
<td>0.76±0.05</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.75±0.09</td>
<td>0.26±0.08</td>
<td>0.75±0.07</td>
<td>0.74±0.06</td>
<td>0.82±0.05</td>
</tr>
<tr>
<td>Bagging with J48</td>
<td>0.79±0.09</td>
<td>0.26±0.09</td>
<td>0.76±0.07</td>
<td>0.78±0.07</td>
<td>0.83±0.06</td>
</tr>
<tr>
<td>Bagging with REPTree</td>
<td>0.75±0.09</td>
<td>0.29±0.08</td>
<td>0.73±0.07</td>
<td>0.74±0.06</td>
<td>0.83±0.06</td>
</tr>
<tr>
<td>AdaBoost with J48</td>
<td>0.75±0.09</td>
<td>0.26±0.09</td>
<td>0.75±0.07</td>
<td>0.74±0.07</td>
<td>0.82±0.06</td>
</tr>
<tr>
<td>AdaBoost with REPTree</td>
<td>0.70±0.10</td>
<td>0.23±0.08</td>
<td>0.76±0.08</td>
<td>0.72±0.06</td>
<td>0.81±0.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attributes selected by Information Gain method</th>
<th>TPR (sensitivity)</th>
<th>FPR (specificity)</th>
<th>PPV</th>
<th>NPV</th>
<th>ROC area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision tree J48</td>
<td>0.71±0.12</td>
<td>0.32±0.11</td>
<td>0.70±0.07</td>
<td>0.70±0.06</td>
<td>0.70±0.07</td>
</tr>
<tr>
<td>REPTree</td>
<td>0.70±0.11</td>
<td>0.29±0.11</td>
<td>0.72±0.07</td>
<td>0.70±0.07</td>
<td>0.74±0.07</td>
</tr>
<tr>
<td>Bayes net</td>
<td>0.76±0.10</td>
<td>0.34±0.09</td>
<td>0.70±0.07</td>
<td>0.73±0.06</td>
<td>0.80±0.06</td>
</tr>
<tr>
<td>k-nearest neighbors</td>
<td>0.65±0.12</td>
<td>0.40±0.11</td>
<td>0.63±0.08</td>
<td>0.62±0.07</td>
<td>0.66±0.08</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.75±0.09</td>
<td>0.28±0.08</td>
<td>0.74±0.07</td>
<td>0.73±0.06</td>
<td>0.80±0.05</td>
</tr>
<tr>
<td>Bagging with J48</td>
<td>0.76±0.10</td>
<td>0.32±0.09</td>
<td>0.71±0.08</td>
<td>0.73±0.06</td>
<td>0.79±0.06</td>
</tr>
<tr>
<td>Bagging with REPTree</td>
<td>0.75±0.10</td>
<td>0.25±0.08</td>
<td>0.75±0.07</td>
<td>0.74±0.05</td>
<td>0.82±0.06</td>
</tr>
<tr>
<td>AdaBoost with J48</td>
<td>0.72±0.10</td>
<td>0.29±0.09</td>
<td>0.72±0.07</td>
<td>0.71±0.06</td>
<td>0.80±0.06</td>
</tr>
<tr>
<td>AdaBoost with REPTree</td>
<td>0.74±0.09</td>
<td>0.25±0.09</td>
<td>0.76±0.08</td>
<td>0.74±0.08</td>
<td>0.82±0.06</td>
</tr>
</tbody>
</table>

7. Discussion

In the conducted study, the feature selection methods were used to reduce the dimensionality aiming at improving the results of the classifiers; however, they also provided important information about the high pruned models. In addition, multiclassifier models are composed by several trees thus is impossible to show them.

However, we have selected to show as an example those rules that appear in most of the trees and that are short, that is, they represent a classification of the instances with a small number of attributes. These representative rules are shown in Table 1. For instance, the first rule in the example says that if the time of VMNI is lesser or equal than two hours and the fluid balance after 12 hours of VMNI application is lesser or equal than 1000 ml then VMNI administration is successful. However, another rule says that for the same time of VMNI, if fluid balance after 12 hours is greater than 1000 ml and the respiratory rate change is greater than 4 breaths a minute then VMNI administration fails.

Table 1 shows the results obtained by these classification algorithms with the three different datasets. The reported values are: TPR (True Positive Rate), also known as sensitivity or recall, FPR (False Positive Rate) or specificity, PPV (Positive Predictive Value), NPV (Negative Predictive Value) and area under the ROC curve.

In order to evaluate the effectiveness of the feature selection methods, the values of accuracy (percentage of correctly classified instances) provided by the applied algorithms were represented in Figure 2. The comparison between the datasets obtained by different feature selection methods can be observed in this graph. Additionally, a pair-wise comparison of the classifiers using a paired t-test was performed with the aim of evaluating the statistical significance of the results. First, the test was applied to compare the accuracy of every algorithm for the three datasets and then to determine the best algorithms for the dataset providing the best results of accuracy. The established level of significance (p) was 0.05.
most influential factors in NIMV outcome. Although CFS and IG methods did not give exactly the same results, they were co-incident for some variables. The most significant are: NIMV hours, PaCO\(_2\) at the start, PaO\(_2\)/FiO\(_2\) ratio at the start, hemato-crit at the start and PaO\(_2\)/FiO\(_2\) ratio after two hours.

Several works in the literature identify some of these factors as predictors of NIMV failure. In [28] high APACHE II and RR > 35 are pointed as the main factors. pH and PaCO\(_2\) at admission and after one hour are marked in [11] as the best predictors for patients suffering hypercapnic respiratory failure. On the other hand, PaO\(_2\)/FiO\(_2\) ratio, albumin and GCS (Glasgow Coma Scale), all of them at the start, are the most important factors marked in [29]. In general, these findings show that an overall bad situation or a poor initial response to NIMV are markers of poor prognosis. In addition, fluid balance after 24 hours and fluid balance after 12 hours were selected by the IG method with a great weight of information gain, 16.62% and 8.46% respectively. This fact indicates the importance of avoiding very positive balances for these patients.

Regarding the individual behavior of the variables with respect to the classes we found that is similar for demographic variables such as age and sex, which presents similar proportion of both classes for its different values, contrarily to the behavior of other important variables. We highlight RF etiology, which presents a dissimilar distribution for its different types. While re-worsening of chronic obstructive pulmonary disease (COPD), acute hypercapnic RF and acute pulmonary edema show a

## Table 2
T-test level of significance (p values) of the accuracy differences for the datasets of selected attributes, CFS and IG (columns 1 and 2) against the all attributes dataset and the significance of the accuracy results of Bagging with J48 against the remainder algorithms for the CFS dataset (column 3)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CFS dataset vs all attributes dataset</th>
<th>IG dataset vs all attributes dataset</th>
<th>CFS dataset Bagging with J48 vs other algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree J48</td>
<td>3.18E-02</td>
<td>4.36E-03</td>
<td>4.49E-06</td>
</tr>
<tr>
<td>REPtree</td>
<td>4.71E-02</td>
<td>4.41E-01</td>
<td>7.28E-05</td>
</tr>
<tr>
<td>Bayes net</td>
<td>2.96E-05</td>
<td>1.91E-01</td>
<td>4.39E-01</td>
</tr>
<tr>
<td>k-nearest neighbors</td>
<td>6.98E-06</td>
<td>1.39E-07</td>
<td>2.90E-05</td>
</tr>
<tr>
<td>Random forest</td>
<td>3.30E-02</td>
<td>2.13E-01</td>
<td>7.82E-01</td>
</tr>
<tr>
<td>Bagging with J48</td>
<td>3.17E-03</td>
<td>2.98E-02</td>
<td>–</td>
</tr>
<tr>
<td>Bagging with REPtree</td>
<td>1.52E-04</td>
<td>2.70E-03</td>
<td>1.16E-01</td>
</tr>
<tr>
<td>AdaBoost with J48</td>
<td>1.61E-03</td>
<td>8.28E-01</td>
<td>4.84E-02</td>
</tr>
<tr>
<td>AdaBoost with REPtree</td>
<td>1.04E-02</td>
<td>1.29E-03</td>
<td>1.51E-03</td>
</tr>
</tbody>
</table>
low rate of NIMV failure, the remaining types, acute hypoxemic RF and postextubation R, present a higher rate, especially in the case of acute hypoxemic RF. Illustration of these relationships between the values of some significant variables and the classes is shown in Figure 1 of ▶Appendix I.

On the other hand, the classification models provide us with information about what combination of factors can lead to success or failure of the NIMV. These models can be also applied to future patients in order to predict the result of NIMV administration.

▶Table 1 shows some quality metrics for the models induced by several classification algorithms from the three different datasets described in the previous section. For all the datasets, the values of sensitivity are slightly higher for multiclassifiers than for most of the single classifiers. In some cases Bayes Net achieves better results in the classification of positive examples but its results are worst in the classification of negative ones. A positive aspect of multiclassifiers, which can be observed in the table, is the fact that the values of PPV and NPV are very similar in most of the cases, this indicates that they are not biased towards a specific kind of error. Concerning the ROC area, the CFS dataset presented better results than the other two datasets, although for the Bagging multiclassifier with REP Tree only surpasses a little the values obtained with the other datasets and for AdaBoost with REP Tree ROC area value obtained with the CFS dataset was slightly lower than the value obtained with the IG dataset.

▶Figure 2 shows the accuracy obtained by the classification algorithms with the three datasets described previously. The best behavior corresponds to the Bagging algorithm with J48 as base classifier and using as input the dataset containing the attributes selected by the CFS method. At first glance, it can be seen that prior application of the CFS feature selection method to the classification algorithms leads to better results in spite of using only 17 attributes in the induction of the predictive models. All of the algorithms present better accuracy when they work with CFS selected attributes than when they work with all of them. We can see in ▶Table 2 that the accuracy improvement for the dataset containing the CFS selected attributes is statistically significant for all the algorithms since all the p values obtained in the conducted t-test are lesser than 0.05, the established level of significance. However, in the case of IG selected attributes, the accuracy presents a significant improvement only for four algorithms, J48, Bagging with J48 and REP Tree and Adaboost with REP Tree. Besides, for k-nearest neighbors the accuracy obtained with the IG selected attributes is significantly worst. We want to highlight the case of Random Forest where the improvement achieved with the CFS selection was significant despite the fact that this algorithm has its own feature selection procedure.

On the other hand, the paired t-test [30] was conducted in order to evaluate whether the Bagging with J48 algorithm is significantly better than the others algorithms. The last column of ▶Table 2 shows that the improvement in the accuracy is significant for most of the algorithms but for three of them (Bayes net, Random tree and Bagging with REP Tree) is not significant.

These predictive models obtained from data about previous patients can be a useful tool to determine if NIMV administration, under specific conditions (the attributes of the models), to new patients will be successful. In an automated system the necessary information for applying the models is stored in a database, thus the process to obtain the outcome would be fast and simple and decisions based on it could be made quickly.

Due to space limitation, in this work only the models built from all available variables are shown, however, since the predictive models involve some evolutionary variables, several models built from the available variables in different moments can be managed along the time in order to allow make decisions at different times.

8. Conclusions
One of the main challenges in the application of NIMV is correctly identifying the patients who will respond to the treatment. Statistical techniques have generally been used to find out the key indicators for the selection of suitable patients for effective NIMV application; however, data mining techniques have not been exploited in this field until now.

This paper describes a data mining study carried out in an attempt to identify early factors influencing the success/failure of NIMV and to build models that will enable us to predict NIMV results in future patients. We applied several algorithms, including feature selection methods, simple classifiers, and multiclassifiers to data obtained from an Intensive Care Unit (ICU) over 6 years. From among the great variety of variables analyzed, feature selection methods provided the most influential variables in the success/failure of NIMV, such as NIMV hours, PaCO₂ at the start, PaO₂/FiO₂ ratio at the start, hematocrit at the start or PaO₂/FiO₂ ratio after two hours. These methods were also used to select the attributes employed in the classification, thus improving the results of the classifiers. The results prove the effectiveness of applying feature selection techniques before predictive tasks since in most of the experiments the accuracy was improved compared to the results obtained from the dataset containing all of the attributes. The algorithm that initially seemed to provide the best values of precision and of other validation metrics was the Bagging multiclassifier. Specifically, the best result was obtained when J48 was used as base classifier and the dataset used as input was the one containing the attributes selected with the CFS method. However, significance tests showed that the superiority of this algorithm is statistically significant against all the algorithms except Bayes Net and Random Forest.

All these data mining techniques can be useful when deciding whether or not to use NIMV in future patients, thus improving patients’ outcomes.

Acknowledgment
The authors thank the staff of the Medical Records Department of the University Hospital of Salamanca for their support in providing patient data.
References