Feature selection for detecting patients with weaning failures in Intensive Medicine

Sérgio Oliveira, Filipe Portela, Manuel Filipe Santos, José Neves, Álvaro Silva and Fernando Rua

Abstract**—** In Intensive Care Units most of the admitted patients are mechanically ventilated. The process of ventilator weaning is delicate and it is conducted by following a set of steps. Normally a weaning tentative is executed based in the patient condition (by analyzing ventilation parameters) and physician's knowledge. In some cases this process fails and it causes long term injuries to the patients. The main goal of this work it is to detect patterns to non-successful weaning in order to avoid a wrong tentative and consequently improve patient condition. Clustering data mining was used to select and identify the features and the patterns associated to failures. As result an Index-Davies Bouldin of 0.9819 was achieved. This result represents the better variables symmetric among the clusters created.

Keywords**—** Ventilation Weaning, Intensive Medicine, Intensive Care Unit, Respiratory Diseases, Data Mining, Clustering, INTCare, Mechanical Ventilation, Extubation, Feature Selection.

I. Introduction

Respiratory failure is one of the most common causes of Intensive Care Unit (ICU) admission. 75% of the patients admitted in an ICU require mechanical ventilation. Despite its benefits, these procedures might have some serious drawbacks contributing to promote lungs injury.

Mechanical Ventilation (MV) is one of the most delicate processes in Intensive Medicine. Although MV had an important role in patient life support, a wrong procedures or configuration can provoke long injuries to the patients.

Mechanical ventilation can have negative effects and its mortality rate ranges is from 41% to 65% [1]. The number of re-intubations vary from 2% to 25% [2].

Automatic control of mechanical ventilation can significantly improve patient care in the ICUs, reduce the mortality and morbidity rates associated with provision of inappropriate ventilator treatments and reduce healthcare costs.

The main goal of this work it is characterize patients whom it is not advised to make a weaning. In this work the objective is not to predict if a patient can be or not extubated but create patterns of patients that should not be put in a weaning process.

This work was supported by the use of Clustering Data Mining techniques (K-Means and K-Medoids) to create patters of weaning failures. Clustering is considered a task of grouping a set of data in such a way that objects in the same group (called a cluster) are more similar.

The work is framed in INTCare project and it used real data collected in real-time from the ICU of Hospital Santo António, Centro Hospitalar do Porto (CHP), Portugal.

As result it was possible to make a feature selection and identify a set of patient characteristics associated to weaning failures. The Index-Davies Bouldin achieved a symmetry inter-clusters of 0.9819. The features with most impact and identified by the better cluster were: CDYN, CSTAT, Flow and Support Pressure.

The paper is divided in seven sections after introduce the work a set of related concepts are presented in the second section. In the third section is presented the work context. Section four present the work developed following CRISP-DM methodology. In section five the results achieved are analyzed having in consideration the main target. Finally some conclusion are written and future work defined.

II. Background

A. Intensive Care Units

The patients who are admitted to Intensive Care Units (ICU) are constantly monitored. They have a set of sensors connected from the body to the bedside monitors. The most common monitoring process is patient vital signs and mechanical ventilation. In fact these patients need intensive care and typically they are in a risk-life condition, being their life condition supported by ventilators.

The main goal of intensive medicine it is use the artefacts available in ICU and the intensivists knowledge to diagnose and treat patients with serious illnesses, restoring them to their previous health condition [3].

Most recently became more difficult to make decision in ICU taking in attention all the data collected from the patient.

This work was FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

Sérgio Oiveira is with Algoritmi Research Centre, University of Minho, Portugal.

Filipe Portela is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (Corresponding author to provide phone: +351253510319; fax: +351253510300; e-mail: cfp@dsi.uminho.pt).

Manuel Filipe Santos, is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (e-mail: mfs@dsi.uminho.pt).

José Maia Neves is with Algoritmi Research Centre, University of Minho, Braga, Portugal. (e-mail: {jneves@di.uminho.pt).

Álvaro Silva and Fernando Rua are with Intensive Care Unit of Centro Hospitalar do Porto, Portugal (e-mail: {moreirasilva@me.com; fernandorua.sci@hgsa.min-saude.pt).

The existence of vital signs monitors and ventilators allow a continuous data streaming. This situation difficult the data analysis in a short period of time due to a high number of data collected. ICU is a potential source of implicit knowledge. This knowledge can be used to improve the decision making process.

B. Mechanical Ventilation and Weaning

Respiratory failure is a syndrome in which the respiratory system fails in one or both of its gas exchange functions: oxygenation and carbon dioxide elimination [4].

The goal is to reduce lung injury due to over distention. However, the efficacy of this approach has not been established [5]. To overcome this problem the patients are ventilated using artificial ventilation.

Nowadays mechanical ventilator, are only used by the clinicians to consult the patient values. The data observed are not stored in a database. This situation results in a wasting of data that could be transformed in knowledge and it could be very useful to the decision-making process.

In addition, the process of ventilator weaning is based in a medical assumption [6] and in a tentative-error procedure, which sometimes seriously compromises the patient condition.

Mechanical ventilation is commonly used in ICU and it is very important to treat many different illnesses, however is relatively costly [2].

Weaning is a gradual process of liberation from, or discontinuation of, mechanical ventilator support resulting in an extubation. In Intensive Medicine an extubation process is considered successful when a patient can breathe from himself for a period upper than one hour.

The Intelligent Decision Support Systems (IDSS) for mechanical ventilation can be grouped in two types: an expert advisory systems or an automatic control of ventilation or weaning [7]. In the ICUs there is a set of IDSS system to ventilators, however, most of them are rule-based system. They are not adaptive and they do not use the results obtained to improve the models.

After an overview [7] it is also possible to verify that most of the existing systems is not using data mining to predict the results. The most far as they can go it is in the input data that can be based in clinical rules and guidelines. Many of its rules can be adaptive and can be derived on the basis of physiological models.

C. INTCare

In the ICU of CHP was deployed a Pervasive Intelligent Decision Support System (PIDSS). This PIDSS is in a continuous developing and test and it is a result of INTCare project [8].

INTCare system is able to monitoring the patient condition in real-time by collecting, processing and displaying the information collected from the bedside monitors and other hospital sources in an intuitive and easy way.

It also has a module to support the decision process through Data Mining models. This module can induce in real-time and using online learning several models able to predict clinical events, as is for example patient outcome [9], organ failure [9], length of stay [10], readmission [11, 12] and barotrauma [13], among others.

INTCare uses intelligent agents [15, 16] to perform their tasks automatically and without human intervention.

This work is inserted in the second phase of the project where the main concern is the respiratory system.

After make a first research to predict barotrauma [13, 14] now it is time to explore a new field: weaning and extubation.

D. Data Mining

Data Mining is the process of using artificial intelligence techniques and statistical and mathematical functions to extract knowledge from the data stored in the database. The achieved knowledge can be presented in multiple forms: business rules, similarities, patterns or correlations [17]. Clustering is inserted in the group of Data Mining problems.

Clustering has as main goal divides the data collected in datasets with similar values. The groups created by the clusters represents a natural catchment of the data and data aggregations. The groups created should make sense, be helpful or both. Clustering rules are not pre-defined. They are discovered along the clustering process. The clusters are characterized by a great internal homogeneity and external heterogeneity [18].

The use of cluster to identify groups of variables is an important asset in many areas like psychology and social sciences, biology, statistical, pattern recognition, information recovery, machine learning and data mining [19, 20].

Clustering offers a high number of algorithms. The choice of the best algorithm to use it is depending from the data collected and project goal. The majority of the clustering methods are grouped into five categories.

The hierarchical methods execute a hierarchical decomposition of the data. These methods can be divisive or agglomerative.

Divisive methods behave the other way. The density-based methods are useful to filter outliers or discovering data with arbitrary form.

The agglomerative methods start with singular objects to create an isolated group. Then the groups or objects are successively merged until a group is missing.

The Partition Methods build a set of partitions on the data, where each partition represents a cluster.

Grid-based methods restrict the space of objects to a finite number of cells forming a grid structure. The Model-based methods formulate a model hypothesis for each cluster and find the best fit the data to the model [21].

Clustering assessment can be done by laying on two factors: compactness and separability. The compactness expresses how much the cluster elements are near. How lesser the variance value it is, greater it will be the cluster compactness. The calculation of the intra cluster distance is very useful to assess this characteristic. The separability evaluates how diverse the clusters are. This can be evaluated by the inter-cluster distance. How higher the distance is better the clusters are [22].

III. Material And Methods

Cross Industry Standard Process for Data Mining (CRISP-DM) was the methodology chosen to conduct this work. CRIS-DM is a cycle process divided in six steps: Business Understanding, Data Understanding, Data preparation, Modelling, Evaluation and Deployment. These steps provide a structured approach to planning a data mining project. In this work only the last phase was not performed.

To do this work R tool were used. R is presented as an environment of statistical programming language for development [23]. The library "cluster" was used primarily to implement the k-means algorithms and Partitioning Around Medoids (PAM) and then for graphing. For optimum number of cluster number it was used the library "fpc". To evaluate the clusters the library "clusterSim" and the Davies-Bouldin Index were used.

IV. Knowledge Discovering Process

The process of knowledge discovery using DM techniques is very complex. As mentioned the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was followed to guide the present study.

A. *Business Understanding*

To identify the patient ventilation variables (feature selection) which interfered in a non-successful weaning / extubation is the main goal of this work. The goal is not to predict if a patient is prepared to be weaned but to identify patterns and clusters of data associated to weaning failures. The clusters were designed using only data monitored by ventilators; the values used were numeric and they were from discrete quantitative type.

Clinically it is expected to create new knowledge to the intensivists helping them to take the better decision in the moment of weaning a patient.

B. *Data Understanding*

The data used to conduct this study were exclusively collected from ventilators in the ICU of CHP. This data corresponds to 50 patients comprising a period between 2014-09-19 and 2015-02-03 in a total of 15325 records. Each record contains thirteen fields:

- CDYN – (F_1): Dynamic compliance in mL/ cmH_2O;
- CSTAT – (F_2): Static compliance from inspiratory pause measured in mL/ cmH_2O;
- FIO2 – (F_3): Fraction of inspired oxygen (%);
- Flow – (F_4): Peak flow setting in litters per minute;
- RR – (F_5): Respiratory rate setting in berths per minute;
- PEEP – (F_6): Positive End-Expiratory Pressure in cmH_2O;
- PMVA – (F_7): Mean airway pressure in cmH_2O;
- Plateau pressure –(F_8): End inspiratory pressure in cmH_2O;
- Peak pressure – (F_9): Maximum circuit pressure in cmH_2O;
- RSTAT – (F_10): Static resistance from inspiratory pause measured in cmH_2O/L/s;
- Volume EXP – (F_11): Exhaled tidal volume in litters;
- Volume INS – (F_12): Tidal volume settings in litters;
- Support Pressure – (F_13): Exhaled minute volume litters;

The values obtained with the coefficient of variation show that the distributions of most fields are mixed. Only the fields: Plateau Pressure, PMVA, Peak Pressure and FIO2 have a coefficient of variation lower than 20%. This measure of dispersion is the ratio between the standard deviation and the average.

Table 1 provides a statistical analysis of the fields used in this study. For each of the numerical values it was analyzed their minimum (MIN) and maximum (MAX), average (AVG), standard deviation (SD) and coefficient of variation (CV).

Table 1 – Distribution of variables

	MAX	MIN	AVG	SD	CV
Plateau Pressure	36	6.2	19.98	3.83	19.18
CDYN	200	0	46.31	40.45	87.34
CSTAT	71	0	20.88	20.93	100.22
PMVA	18	3.1	10.52	1.84	17.48
Peak Pressure	40	9	20.53	3.85	18.75
RSTAT	29	0	8.89	8.60	96.74
FIO2	100	35	49.68	7.43	14.96
FLOW	60	0	23.83	23.61	99.06
RR	24	0	1.69	5.50	324.73
PEEP	10	3	5.09	1.04	20.45
Volume EXP	2.46	0	0.53	0.17	31.83
Volume INS	0.56	0	0.26	0.25	98.06
Support Pressure	27	4	14.1	3.56	25.24

C. *Data Preparation*

To carry out this study it was necessary to identify the patients who were not able to be submitted to the extubation process, even though they were very close to being subjected to the process.

In order to identify these occurrences it was necessary to determinate the respective scenario. Patients who were not submitted to extubation process were patients who had mechanical ventilation variations or the support pressure level was continuous for more than one hour but they never were an attempt to extubation under this scenario. In this process five levels were identified:

- -2: Support pressure is constantly changing (less than 3 minutes);
- -1: Support pressure is continuous between 4 minutes and 30 minutes;
- 0: Support pressure is constant from 30 minutes to 60 minutes
- 1: Maintains the same support pressure for more than 60 minutes but the patient is not extubated;

- 2: Maintains the same support pressure for more than 60 minutes and the patient is extubated;

To this work only the patient with level equal to 1 were considered. After this first selection the records which presented at least one null value were deleted. This operation was of utmost importance, because the models cannot be induced if the dataset has null values. At same time clinical ranges were used to validate the values collected. Based in those information it was possible to consider only correct values (after be validated by an intelligent agent). In this process the records without quality and inconsistences (null values and values out of the normal range) were eliminated.

After all the changes are applied to the initial dataset, the number of records decreased to 13135 records, which is associated to 28 patients.

D. Modelling

The algorithms used to create clusters were: k-means and K-medoids. The choice of these two algorithms was fundamentally based on two characteristics: the principle of partition method and the difference in sensitivity to outliers.

K-means is a simple iterative clustering algorithm which partitions the dataset into a priori number. The algorithm is simple to implement and run, being their implementation faster [25]. The K-means is sensitive to outlier, because the objects are far from the majority, which it can significantly influence the average value of the set. This inadvertently affects the allocation of other objects to the clusters. This effect is particularly exacerbated by the use of the squared error function [24].

On the other hand the k-medoids instead of using the value of a cluster object as a reference point, takes on real objects and represents the clusters, dare one object at a cluster. The remaining objects are assigned to the most similar cluster. The partitioning method is then performed based on the principle teaches the sum of the differences between each of p and its corresponding object representative object [24]. K-medoids algorithm is similar to the K-means, except that the centroids must belong to the set of data that are grouped [25].

The implementation of the algorithms required some settings. The identification of the appropriate number of K was achieved by implementing the calculation of Square Error Sum (SSE). The SSE determines the squared distances between each member of the cluster and the cluster centroid.

$$SSE(o_i) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ji}^{p}\, dist(o_i, c_j)^2$$

In general, when the number of clusters increases SSE value decreases. When it is identified a dramatic decrease in the SEE value, the number of K can be considered as being great [24].

There were ten plays for each dataset used for each model. The metric used to calculate the dissimilarity being the observations was "Euclidean". The developed models can be represented by the following expression:

$$M_n = \{A_f;\ F_i;\ D_x;\ AG_y\}.$$

The model M_n belongs to an approach (A) and it is composed by a set of fields (F), a type of variable (TV) and an algorithm (AG):

$$A_f = \{Discription\ (Clustering)_1\}$$

$$F_i = \{F\_1_1, F\_2_2, F\_3_3, F\_4_4, F\_5_5, F\_6_6, F\_7_7, F\_8_8, F\_9_9, F\_10_{10}, F\_11_{11}, F\_12_{12}, F\_13_{13}\}$$

$$TV_x = \{Qualitative\ variables\ ordinal_1\}$$

$$AG_y = \{K - means_1, K - medoids(PAM)_2\}$$

E. Evaluation

In this phase, the Davies-Bouldin Index (DBI) was used to evaluate the models generated.

Among the algorithms used the one which presented the most satisfactory results was the K-means algorithm.

Some of the models developed had interesting results but only one appears to present satisfactory results (DBI). Table 2 shows the most relevant models.

Table 2 – Models for clustering

Model	Fields	Number Clusters	Algorithm	DBI
M_1	$F_{\{1,2,3,4,5,6,7,8,9,10,11,12,13\}}$	8	AG_2	1.31
M_2	$F_{\{2,3,4,5,6,13\}}$	5	AG_2	1.10
M_3	$F_{\{1,2,4,13\}}$	7	AG_1	0.98
M_4	$F_{\{1,2,3,4,13\}}$	10	AG_1	1.10
M_5	$F_{\{1,2,3,4,11,13\}}$	10	AG_1	1.20

Analyzing Table 2 it is possible to identify that the model M_3 is best model generated. The Davies Bouldin Index tends to +∝ however M_3 model has an index of 0.98. This is not the ideal case but it is the model closest to 0.

In order to present a representation of the data segments generated by the best model, the figure 1 was designed.

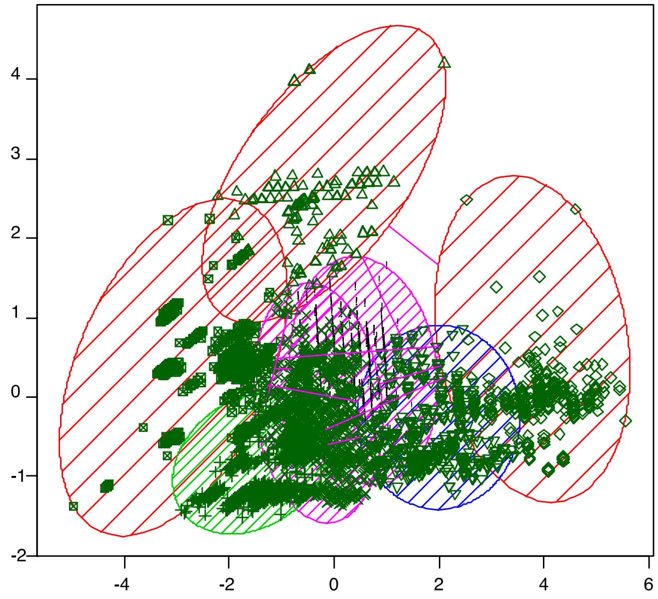

Figure 1: Graphical representation of the clusters presented in table 2

Since the goal is select and identify patient features which were not capable of being subjected to extubation process, it was necessary to identify the cluster that best characterizes these patients.

In order to identify the cluster that best identifies such patients, it was necessary to find the cluster which presented the lowest variability.

Cluster 3 was identified as being the best cluster. It has the lowest sum of squares with a ratio of 1247.02. This cluster contains about 2382 occurrences. Thus cluster 3 has the lowest average of the sum of the squares value: 0.52.

Table 3 shows the patient characteristics / features of Cluster 3. For example in the case of CDYN (F1) the minimum value is 13 and the maximum values is 20.

When a patient presents a value for each one of the four fields between the minimum and the maximum values of the cluster, he is associated to cluster 3.

Table 3 – Distributions of cluster 3

Cluster	Fields	MIN	MAX	MEAN	SD	CFV
Cluster 3	$F_{\{1\}}$	13	20	16.49	1.27	7.73%
	$F_{\{2\}}$	0	48	7.61	14.47	190.30%
	$F_{\{4\}}$	6.9	12	9.14	0.73	7.94%
	$F_{\{13\}}$	8	14	10.72	1.28	11.88%

V. Discussion

The best model M_3 created 7 clusters. Cluster 3 is the only able to properly identify the properties of the patients that should not be weaned / extubated.

The amount of Cluster 3 registers corresponds to 18.13% of the dataset used. The third cluster presents a significant amount of records and it displays relevant characteristics of the distribution of its variables.

There is great heterogeneity in the variables that make up the Cluster 3. $F_{\{2\}}$ is the variable that demonstrates to have more homogeneous values. In general the cluster presents variables with a little dispersion.

Although the clusters generated by the M_3 model have at least one interception with another cluster, cluster 3 is better prepared to target the patients who do not meet the conditions to be extubated. Since it has an average of sum of squares lesser than all the other clusters, presenting a value of 0.52.

VI. Conclusion

After conclude this study it was possible to identify a set of patient features / variables that have great similarity on weaning / extubation failures.

The most satisfactory result is attained by the model M3 obtaining a Davies-Bouldin Index of 0.98.

M3 is the model which presented results closest to the optimum value: zero. It should be noted that most of the variables used in cluster 3 has an acceptable dispersion, since its value is 13.47. This result was obtained with the implementation of the k-means algorithm, thus demonstrating to be the most suitable algorithm to carry out the data segmentation.

The achieved results gives confidence in continues this research work in order to make a better feature selection of the variables associated to a wrong weaning.

The better cluster Davies Bouldin Index is acceptable being it above to the maximum acceptable level (1). It presents a result of 0.98 which means that there is similarities in the clusters created.

Presenting the cluster in an intuitive way and easy to understand and combining it with the clinical experts it is possible to provide best care to the patients.

This work is a starting point to avoid wrong extubation and avoid long injuries associated to respiratory system (e.g. lungs injuries).

VII. Future Work

Having in consideration the achieved results it is expected to improve these results by adding other variables and explore other clustering techniques.

At same time it will try induce data mining models to predict the probability of a patient have a weaning and extubation.

In the future all the results achieved will be included in INTCare system in order to give more options to the intensivist in the moment of deciding.

Acknowledgment

The authors would like to FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EEI-SII/1302/2012 (INTCare II). This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

References

[1] A. S. Fauci, "Harrison's Principles of Internal Medicine, 17e," ed: Silverchair Science: Minion, 2008.

[2] F. T. Tehrani, "Automatic control of mechanical ventilation. Part 2: the existing techniques and future trends," *Journal of clinical monitoring and computing,* vol. 22, pp. 417-424, 2008.

[3] F. Portela, M. F. Santos, Á. Silva, F. Rua, A. Abelha, and J. Machado, "Adoption of Pervasive Intelligent Information Systems in Intensive Medicine," *Procedia Technology,* vol. 9, pp. 1022-1032, 2013.

[4] A. Kaynar and S. Sharma. (2010, Respiratory Failure. 39. Available: http://emedicine.medscape.com/article/167981-print

[5] T. E. Stewart, M. O. Meade, D. J. Cook, J. T. Granton, R. V. Hodder, S. E. Lapinsky, *et al.*, "Evaluation of a Ventilation Strategy to Prevent Barotrauma in Patients at High Risk for Acute Respiratory Distress Syndrome," *New England Journal of Medicine,* vol. 338, pp. 355-361, 1998.

[6] S. P. Stawicki, "Mechanical ventilation: weaning and extubation," 2007.

[7] F. T. Tehrani and J. H. Roum, "Intelligent decision support systems for mechanical ventilation," *Artificial Intelligence in Medicine,* vol. 44, pp. 171-182, 2008.

[8] F. Portela, M. F. Santos, J. Machado, A. Abelha, Á. Silva, and F. Rua, "Pervasive and intelligent decision support in Intensive Medicine–the complete picture," in *Information Technology in Bio-and Medical Informatics*, ed: Springer, 2014, pp. 87-102.

[9] F. Portela, M. F. Santos, J. Machado, A. Abelha, and Á. Silva, "Pervasive and Intelligent Decision Support in Critical Health Care Using Ensembles," in *Information Technology in Bio-and Medical Informatics*, ed: Springer Berlin Heidelberg, 2013, pp. 1-16.

[10] R. V. Filipe Portela, Sérgio Oliveira, Manuel Filipe Santos, António Abelha, José Machado, Álvaro Silva and Fernando Rua, "Predict hourly patient discharge probability in Intensive Care Units using Data Mining," *ScienceAsia Journal (ICCSCM 2014)*, 2014.

[11] Pedro Braga, F. Portela, and M. F. Santos, "Data Mining Models to Predict Patient's Readmission in Intensive Care Units," in *ICAART - International Conference on Agents and Artificial Intelligence*, Angers, France, 2014.

[12] R. Veloso, F. Portela, M. F. Santos, Á. Silva, F. Rua, A. Abelha, *et al.*, "A clustering approach for predicting readmissions in intensive medicine," *Procedia Technology,* vol. 16, pp. 1307-1316, 2014.

[13] F. P. Sérgio Oliveira, Manuel Filipe Santos, José Machado, António Abelha, Álvaro Silva and Fernando Rua, "Intelligent Decision Support to predict patient Barotrauma risk in Intensive Care Units," in *Procedia Technology - HCIST 2015 - Healthy and Secure People*, Elsevier, Ed., ed, 2015.

[14] S. Oliveira, F. Portela, M. F. Santos, J. Machado, A. Abelha, Á. Silva, *et al.*, "Predicting Plateau Pressure in Intensive Medicine for Ventilated Patients," in *New Contributions in Information Systems and Technologies,* ed: Springer, 2015, pp. 179-188.

[15] L. Cardoso, F. Marins, F. Portela, M. Santos, A. Abelha, and J. Machado, "The Next Generation of Interoperability Agents in Healthcare," *International Journal of Environmental Research and Public Health,* vol. 11, pp. 5349-5371, May 2014.

[16] M. F. Santos, F. Portela, M. Vilas-Boas, J. Machado, A. Abelha, and J. Neves, "INTCARE - Multi-agent approach for real-time Intelligent Decision Support in Intensive Medicine," in *3rd International Conference on Agents and Artificial Intelligence (ICAART)*, Rome, Italy, 2011.

[17] E. Turban, R. Sharda, and D. Delen, *Decision Support and Business Intelligence Systems*, 9th Edition ed.: Prentice Hall, 2010.

[18] S. Tufféry, *Data mining and statistics for decision making*: John Wiley & Sons, 2011.

[19] H. C. Koh and G. Tan, "Data mining applications in healthcare," *Journal of Healthcare Information Management—Vol,* vol. 19, p. 65, 2011.

[20] P.-N. Tan, Steinbach, M., & Kumar, V, *Introduction to Data Mining* 1ed.: Addison-Wesley Longman Publishing Co., Inc., 2005.

[21] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*: Morgan kaufmann, 2006.

[22] J. C. Krzysztof, W. S. Roman, and A. Lukasz, "Data Mining: A Knowledge Discovery Approach," ed: Springer, 2007.

[23] L. Torgo, *Data mining with R: learning with case studies*: Chapman & Hall/CRC, 2010.

[24] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*: Elsevier, 2011.

[25] X. Wu and V. Kumar, *The top ten algorithms in data mining*: CRC Press, 2009.