Organ Failure Diagnosis by Artificial Neural Networks

ÁLVARO SILVA
Serviço de Cuidados Intensivos
Hospital Geral de Santo António
Porto, Portugal
a.moreirasilva@mail.telepac.pt

PAULO CORTEZ
Dep. de Sistemas de Informação
Universidade do Minho
Guimarães, Portugal
pcortez@dsi.uminho.pt

MANUEL SANTOS
Dep. de Sistemas de Informação
Universidade do Minho
Guimarães, Portugal
mfs@dsi.uminho.pt

LOPES GOMES
Clínica Médica I
Inst. de Ciências Biomédicas Abel Salazar
Porto, Portugal
cardiologia.hgsa@mail.telepac.pt

JOSÉ NEVES
Dep. de Informática
Universidade do Minho
Braga, Portugal
jneves@di.uminho.pt

ABSTRACT
In recent years, *Clinical Data Mining* has gained an increasing acceptance by the research community, due to its potential to find answers that could extend life or give comfort to ill persons. In particular, the use of tools such as *Artificial Neural Networks*, which have been mostly used in *classification* tasks. The present work reports the adoption of these techniques for the prediction of organ dysfunction of *Intensive Care Unit* patients. The novelty of this approach is due to the use intermediate outcomes, defined by the *Out of Range Measurements* of four bedside monitored variables, which obtained an overall accuracy of 70%.

KEY WORDS
Intensive Care Medicine, Classification, Multilayer Perceptrons, Out of Range Measurements, Sequential Organ Failure Assessment

1 Introduction

In *Intensive Care Units (ICUs)*, scoring the severity of illness has become a routine in daily practice. Indeed, several metrics are available such as *Acute Physiology and Chronic Health Evaluation System (APACHE II)* or *Acute Physiology Score (SAPS II)* [1]. However, most of these prognostic models (given by *Logistic Regression*) are static, since they are computed with data collected within the first 24 hours of a patient's admission to the *ICU*. Therefore, a limited impact will occur in clinical decision making, due to the lack of accuracy of the patient's condition, since no intermediate measures are used.

On the other hand, an increasing attention has been set over the *Clinical Data Mining* field, which aims at discovering some structure in large clinical heterogeneous data [2]. In particular, *Artificial Neural Networks (ANNs)* are connectionist models that mimic the central nervous system, being successfully applied for the design of medical intelligent systems. For instance, the number of *ANN* publications in *Medicine* has grown from two in 1990 to five hundred in 1998 [3].

This interest arose due to an ever-increasing load of data, which presents high complexity. Human experts are limited and may overlook important details. *ANNs* have the potential to solve some of these hurdles, due to capabilities such as nonlinear learning, multi-dimensional mapping and noise tolerance [4].

In *ICUs*, optimal time *Organ Failure Diagnosis* is a critical task, since its rapid detection may allow physicians to respond quickly with therapy. Moreover, multiple organ failure will highly increase the probability of the patient's death. The *Sequential Organ Failure Assessment (SOFA)* is a diary index, ranging from 0 to 4, that allows organ failure detection [5]. An organ is considered to fail when its *SOFA* score is equal or higher than 3.

In the present work, *ANNs* will be adopted for organ failure prediction (identified by high *SOFA* values) of: *respiratory*, *coagulation*, *liver*, *cardiovascular*, *central nervous system* and *renal*. Several approaches will be tested, using different *feature selection*, *pre-processing* and *modeling* configurations. A particular focus will be given to the use of daily intermediate adverse events, obtained from four hourly bedside measurements.

The paper is organized as follows: first, the clinical data is described; then, the *ANN* models are presented; next, a description of the different experiments performed is given, being the results analyzed; finally, closing conclusions are drawn.

2 Clinical Data

In this work, a part of the *EURICUS II* database (www.frice.nl) was adopted, which encompasses 5355 patients from 42 *ICUs* and 9 EU countries, during a period of 10 months. The database has one *entry* (or *example*) per each day (with a total of 30570), being its main features described in Table 1.

The first six rows denote the *SOFA* values of the pa-

tient's condition in the previous day. In terms of notation, these will be denoted by $SOFA_{d-1}$, where d represents the current day. The *case mix* appears in the next four rows, an information that remains unchanged during the patient's internment. Finally, the last four rows denote the intermediate outcomes, which are triggered from four monitored parameters: the *systolic Blood Pressure (BP)*, the *Heart Rate (HR)*, the *Oxygen saturation (O2)* and the *URine Output (UR)*. A panel of *EURICUS II* experts defined the normal ranges for these variables (Tables 2 and 3), being considered an *Out of Range Measurement (ORM)* when an *Event* or *Critical Event* occurs.

Table 1. The clinical data characteristics.

Attribute	Description	Domain
respirat	Respiratory	$\{0,1,2,3,4\}$
coagulat	Coagulation	$\{0,1,2,3,4\}$
liver	Liver	$\{0,1,2,3,4\}$
cardiova	Cardiovascular	$\{0,1,2,3,4\}$
cns	Central nervous system	$\{0,1,2,3,4\}$
renal	Renal	$\{0,1,2,3,4\}$
admfrom	Admission origin	$\{1,\ldots,7\}$
admtype	Admission type	$\{1,2,3\}$
sapsII	SAPSII score	$\{0,\ldots,160\}$
age	Patients' age	$\{18,\ldots,100\}$
nbporms	Number of *BP ORMs*	$\{0,\ldots,28\}$
nhrorms	Number of *HR ORMs*	$\{0,\ldots,26\}$
no2orms	Number of *O2 ORMs*	$\{0,\ldots,30\}$
nurorms	Number of *UR ORMs*	$\{0,\ldots,29\}$

Before attempting *ANN* modeling, the data was preprocessed, in order to set the desired classification outputs. First, six new attributes were created, by sliding the $SOFA_{d-1}$ values into each previous example, since the intention is to predict the patient's condition ($SOFA_d$) with the available data at day d ($SOFA_{d-1}$, *case mix* and *ORMs*). Then, the last day of the patient's admission entries were discarded (remaining a total of 25309), since in this cases, no $SOFA_d$ information is available. Finally, the new attributes were transformed into binary variables, according to the expression:

$$\begin{array}{ll} 0\ ,\ if\ SOFA_d < 3 & (\textit{false}, \text{no organ failure}) \\ 1\ ,\ else & (\textit{true}, \text{organ dysfunction}) \end{array} \quad (1)$$

3 Artificial Neural Networks

In *MultiLayer Perceptrons (MLPs)*, one of the most popular *ANN* architectures, *neurons* are grouped into *layers* and only *forward connections* exist [4]. Supervised learning is achieved by an iterative adjustment of the network *connection weights* (the *training* procedure), in order to minimize an error function, computed over the training examples (*cases*).

The state of a neuron (s_i) is given by:

$$s_i = f(w_{i,0} + \sum_{j \in I} w_{i,j} \times s_j) \quad (2)$$

where I represents the set of nodes reaching node i, f the activation function (possibly of nonlinear nature), $w_{i,j}$ the weight of the connection between nodes j and i (when $j = 0$, it is called *bias*); and $s_1 = x_1, \ldots, s_n = x_n$, being $x_1, \ldots, x_n$ the input vector values for a network with n inputs.

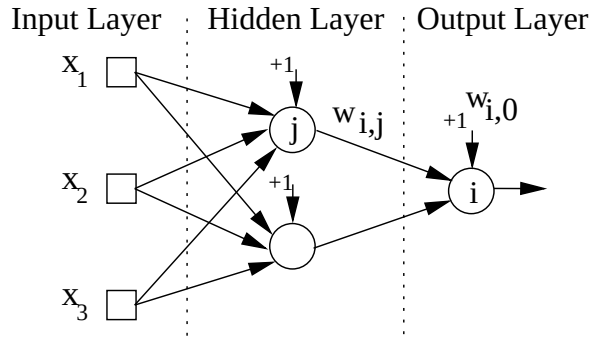

Figure 1. A fully connected *MLP* with 3 inputs, 2 hidden nodes, 1 output and *bias* connections.

In this study, fully connected *MLPs* with *bias* connections, one hidden layer (with a fixed number of hidden nodes) and logistic activation functions ($f(x) = \frac{1}{1+e^{-x}}$) were adopted (Figure 1). Only one output node is used, since each organ system will be modeled by a different *MLP*. This splitting is expected to facilitate the *ANN* learning process. Therefore, the predicted class (P_k) for the k example is given the nearest class value:

$$P_k = \begin{cases} 0\ ,\ if\ s_{k,o} < 0.50 \\ 1\ ,\ else \end{cases} \quad (3)$$

where $s_{k,o}$ denotes the output value for the o output node and the k input example.

The *MLP* input values were rescaled into the range $[-1, 1]$ and the *MLP* initial weights were randomly set within the same range. Then, the *RPROP* algorithm [6] is selected for the *MLP* learning, due to its faster convergence and stability, being stopped when the training error slope is approaching zero or after a maximum of E epochs [7].

To insure statistical significance, 30 runs were applied in all tests, being the accuracy estimates achieved using the *holdout* method [8]. In each simulation, the available data is divided into two mutually exclusive partitions, using stratified resampling: the *training set*, used during the *ANN* learning phase; and the *test set*, being used after training, in order to compute the accuracy estimates.

A common tool for classification analysis is the *confusion matrix* [9], a matrix of size $L \times L$, where L denotes the number of possible classes (*domain*). This matrix is

Table 2. The *Event* parameters.

Event	Suggested Range	Continuously Out of Range	Intermittently Out of Range
BP (mmHg)	$90 - 180$	$\geq 10'$	$\geq 10'$ in $30'$
$O2$ (%)	≥ 90	$\geq 10'$	$\geq 10'$ in $30'$
HR (bpm)	$60 - 120$	$\geq 10'$	$\geq 10'$ in $30'$
UR (ml/hour)	≥ 30	≥ 1 hour	

Table 3. The *Critical Event* parameters.

Critical Event	Suggested Range	Continuously Out of Range	Intermittently Out of Range	Event Anytime
BP (mmHg)	$90 - 180$	$\geq 60'$	$\geq 60'$ in $120'$	$BP < 60$
$O2$ (%)	≥ 90	$\geq 60'$	$\geq 60'$ in $120'$	$O2 < 80$
HR (bpm)	$60 - 120$	$\geq 60'$	$\geq 60'$ in $120'$	$HR < 30 \vee HR > 180$
UR (ml/hour)	≥ 30	≥ 2 hours		≤ 10

Table 4. The 2×2 *confusion matrix*.

$\downarrow$ **actual** $\setminus$ **predicted** $\rightarrow$	**negative**	**positive**
negative	TN	FP
positive	FN	TP

created by matching the *predicted* (*test result*) and *actual* (*patients real condition*) values. When $L = 2$ and there are four possibilities (Table 4): the number of correct positive - *True Positive (TP)*, correct negative - *True Negative (TN)*, incorrect positive - *False Positive (FP)*; and incorrect negative - *False Negative (FN)* classifications.

From this table, three accuracy measures can be defined [10]: the *true Positive Rate (PR)*, also known as *sensitivity*, *recall* and *Type II Error*; the *true Negative Rate (NR)*, also known as *specificity*, *precision* and *Type I Error*; and the *Predictive Accuracy (PA)*, which gives an overall evaluation. These metrics can be computed using the following equations:

$$\begin{aligned} PR &= \frac{TP}{FN+TP} \times 100\,(\%) \\ NR &= \frac{TN}{TN+FP} \times 100\,(\%) \\ PA &= \frac{TN+TP}{TN+FP+FN+TP} \times 100\,(\%) \end{aligned} \tag{4}$$

4 Experiments

4.1 Feature Selection

Four different feature selection configurations will be tested, in order to evaluate the input attribute importance:

SOFA - which uses only the $SOFA_{d-1}$ values (1 variable).

ALL - where all available input information is used ($SOFA_{d-1}$ of the corresponding organ system, the *case mix* and the *ORMs*, in a total of 9 attributes);

NO SOFA - in this case, the $SOFA_{d-1}$ is omitted (8 variables); and

ORM - which uses only the four *ORMs*.

Since the *SOFA* score takes costs and time to obtain, in this study, a special attention will be given to the last two settings.

To boost the *MLP* learning efficiency, a *1-of-C* encoding (one binary variable per class) was applied to the nominal attributes with few different values ($SOFA_{d-1}$, *admfrom* and *admtype*). For example, the *admtype* variable is fed into 3 input nodes, according to the scheme: $1 \rightarrow 001$, $2 \rightarrow 010$ and $3 \rightarrow 001$.

For the initial experiments, it was considered more important to approach feature selection than model selection. Due to time constrains, the number of hidden nodes was set to $round(N/2)$, where N denotes the number of input nodes ($N = 21$, $N = 5$, $N = 16$ and $N = 4$, for the **ALL**, **SOFA**, **NO SOFA** and **ORM** setups); and $round(x)$ gives nearest integer to the x value.

The commonly used $2/3$ and $1/3$ partitions were adopted for the *training* and *test* sets [8], while the maximum number of training epochs was set to $E = 100$. Each input configuration was tested for all organ systems, being the accuracy measures given in terms of the mean of thirty runs (Tables 5 and 6).

The **SOFA** selection manages to achieve a high performance, with a PA ranging from 86% to 97%, even surpassing the **ALL** configuration. This is not surprising, since it is a established fact that the $SOFA$ is a adequate score for organ dysfunction. Thus, the results sug-

Table 5. The **SOFA** and **ALL** performances (in percentage).

Organ	SOFA			ALL		
	PA	PR	NR	PA	PR	NR
respirat	86.3	72.4	90.2	86.2	70.0	90.8
coagulat	97.4	68.8	98.7	97.3	59.6	99.0
liver	98.3	68.6	99.1	98.3	60.2	99.4
cardiova	94.2	84.1	96.3	94.2	84.0	96.3
cns	95.7	92.7	96.4	95.7	92.3	96.4
renal	95.5	71.3	97.8	95.3	66.6	98.1

Table 6. The **NO SOFA** and **ORM** performances (in percentage).

Organ	NO SOFA			ORM		
	PA	PR	NR	PA	PR	NR
respirat	77.9	4.4	98.8	77.6	1.8	99.4
coagulat	95.8	4.6	99.9	95.7	0.0	100
liver	97.3	7.6	99.9	97.3	0.0	100
cardiova	82.8	7.5	99.0	82.2	0.5	99.8
cns	83.5	23.4	97.1	81.6	0.4	99.9
renal	91.4	5.7	99.7	91.1	0.3	100

gest that there is a high correlation between $SOFA_{d-1}$ and $SOFA_d$.

When the $SOFA$ index is omitted (**NO SOFA** and **ORM**), the PA values only decay slightly. However, the PA measure (which is popular within *Data Mining* community) is not sufficient in *Medicine*. Ideally, a test should report both high PR and NR values, which suggest a high level of confidence [10]. In fact, there seems to be a trade-off between these two characteristics, since when the $SOFA$ values are not present (Table 6), the *PR* values suffer a huge decrease (sensitivity loss), while the NR values increase (specificity gain).

4.2 Balanced Training

Why do the **NO SOFA/ORM** selections produce high PA /NR values and low *PR* ones? The answer may be due to the biased nature of the organ dysfunction distributions; i.e., there is a much higher number of *false* (0) than *true* (1) conditions (Figure 2).

One solution to solve this handicap, is to *balance* the training data; i.e., to use an equal number of true and false learning examples. Therefore, another set of experiments was devised (Table 7), using random sampling training sets, which contained $2/3$ of the true examples, plus an equal number of false examples. The test set was composed of the other $1/3$ positive entries. In order to achieve a fair comparison with the previous results, the negative test examples were randomly selected from the remaining ones, with a distribution identical to the one found in the original dataset (as given by Figure 2).

Table 7. The balanced **NO SOFA** and **ORM** performances (in percentage).

Organ	NO SOFA			ORM		
	PA	PR	NR	PA	PR	NR
respirat	61.3	66.4	59.8	67.1	41.1	74.5
coagulat	67.6	66.8	67.7	73.7	41.5	75.1
liver	70.0	71.6	70.0	66.9	36.5	67.8
cardiova	65.9	62.5	66.7	68.2	37.9	74.8
cns	73.6	63.9	75.7	66.8	36.3	73.7
renal	67.8	65.6	68.0	73.2	37.6	76.6

Table 8. The **NO SOFA** performances for a *MLP* with 16 hidden nodes (in percentage).

Organ	NO SOFA		
	PA	PR	NR
respirat	63.3	70.4	61.3
coagulat	70.0	72.0	69.9
liver	72.5	77.3	72.4
cardiova	69.1	66.3	69.8
cns	75.2	72.2	75.8
renal	71.9	70.5	72.0

The obtained results show a clear improvement in the PR values, specially for the **NO SOFA** configuration. Yet, the overall results are still far from the ones given by the **SOFA** selection.

4.3 Improving Learning

Until now, the main focus was over selecting the correct training data. Since the obtained results are still not satisfactory, the attention will move towards better *MLP* modeling. This will be achieved by changing two parameters: the *number of hidden nodes* and the *maximum number of training epochs*. Due to computational power restrictions, these factors were kept fixed in the previous experiments. However, the adoption of *balanced training* leads to a considerable reduction of the number of training cases, thus reducing the required training time.

After some experimental trials, the number of hidden nodes was increased to 16, being the maximum number of epochs set to $E = 1000$. Table 8 shows the results for the **NO SOFA** selection. These settings lead to better results, for all organ systems and accuracy measures.

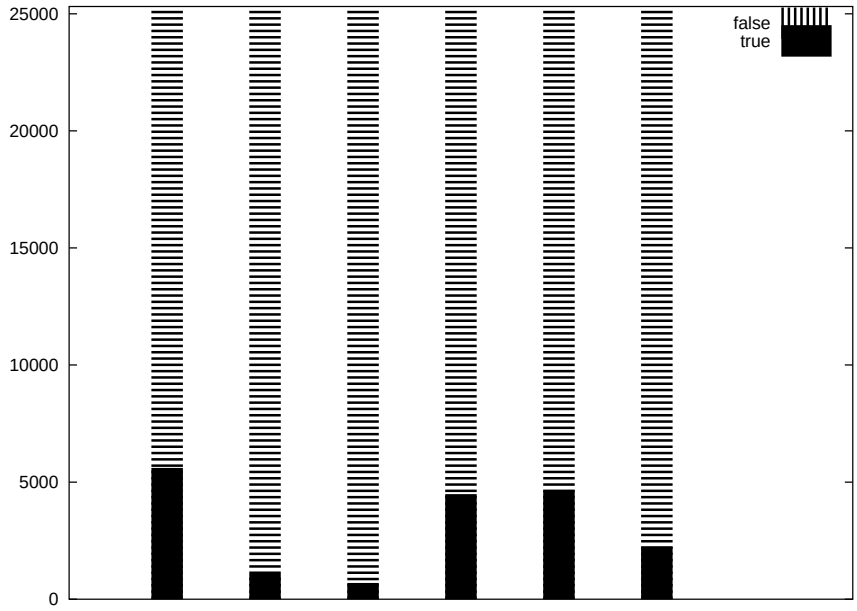

Figure 2. The organ failure true/false proportions (**respirat**, **coagulat**, **liver**, **cardiova**, **cns** and **renal**).

5 Conclusions

The surge of novel bio-inspired techniques, such as *ANNs*, has created new exciting possibilities for the field of *clinical data mining*. In this work, *ANNs* were applied for *Organ Failure Diagnosis* of six organ systems.

Preliminary experiments were drawn to test several feature selection configurations, being the best results obtained by the solely use of the *SOFA* value, measured in the previous day.

When compared with the physiologic intermediate outcomes, the *SOFA* score takes much more time and costs to be obtained. Therefore, another set of experiments were conducted, in order to improve the use of *ORMs*. First, the training sets were balanced, in order to contain similar proportions of positive and negative examples. Then, the number of hidden nodes and training epochs was increased. As the result of these changes, an improved performance was gained, specially in terms of *sensitivity*.

A final comparison is given in Table 9, which compares the diagnostic accuracy of the **SOFA** and the best **ORM** configurations. The former still manages to outperform the latter, although the sensitivity (PR) values are close (being even higher for the **ORM** in the *coagulation* and liver systems).

It is important to stress the main goal of this work: to show that is it possible to diagnose organ failure by using cheap and fast intermediate outcomes. The results so far obtained (an overall accuracy of 70%), although not authoritative, back this claim. In addiction, the proposed approach opens room for the development of automatic tools for clinical decision support.

Table 9. Comparison among the **SOFA** and **NO SOFA** performances (in percentage).

Organ	**SOFA**			**NO SOFA**		
	PA	PR	NR	PA	PR	NR
respirat	86.3	72.4	90.2	63.3	70.4	61.3
coagulat	97.4	68.8	98.7	70.0	72.0	69.9
liver	98.3	68.6	99.1	72.5	77.3	72.4
cardiova	94.2	84.1	96.3	69.1	66.3	69.8
cns	95.7	92.7	96.4	75.2	72.2	75.8
renal	95.5	71.3	97.8	71.9	70.5	72.0
Mean	**94.6**	**76.3**	**96.4**	**70.3**	**71.5**	**70.2**

In future research it is intend to improve the *ORMs* performances, by exploring different *ANNs* topologies (e.g., *Radial Basis Functions*). Another interesting direction is based in the use of training algorithms that can optimize other learning functions (e.g., *Evolutionary Algorithms* [11] or *Particle Swarms* [12]), since the gradient-based methods (such as *RPROP* [6]) work by minimizing the *Sum Squared Error*, a target which does not necessarily correspond to maximizing the sensitivity and specificity rates. Finally, it is intended to enlarge the *ANN* experiments to other *ICU* applications (e.g., predicting *life expectancy*).

References

[1] D. Teres and P. Pekow. Assessment data elements in a severity scoring system (Editorial). *Intensive Care Med*, 26:263–264, 2000.

[2] K. Cios and G. Moore. Uniqueness of Medical Data Mining. *Artificial Intelligence in Medicine*, 2002.

[3] R. Dybowski. Neural Computation in Medicine: Perspectives and Prospects. In H. Malmgreen et al., editor, *Proceedings of the ANNIMAB-1 Conference (Artificial Neural Networks in Medicine and Biology)*, pages 26–36. Springer, 2000.

[4] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[5] J. Vincent et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction / failure. *Intensive Care Med*, 22:707–710, 1996.

[6] M. Riedmiller. Supervised Learning in Multilayer Perceptrons - from Backpropagation to Adaptive Learning Techniques. *Computer Standards and Interfaces*, 16, 1994.

[7] L. Prechelt. *Early Stopping – but when?* In: *Neural Networks: Tricks of the trade*, Springer Verlag, Heidelberg, 1998.

[8] A. Flexer. Statistical evaluation of neural networks experiments: Minimum requirements and current practice. In *Proceedings of the 13th European Meeting on Cybernetics and Systems Research*, volume 2, pages 1005–1008, Vienna, Austria, 1996.

[9] R. Kohavi and F. Provost. Glossary of Terms. *Machine Learning*, 30(2/3):271–274, 1998.

[10] D. Essex-Sorlie. *Medical Biostatistics & Epidemiology: Examination & Board Review*. McGraw-Hill, 1995.

[11] P. Cortez, M. Rocha, and J. Neves. A Lamarckian Approach for Neural Network Training. *Neural Processing Letters*, 15(2):105–116, April 2002.

[12] R. Mendes, P. Cortez, M. Rocha, and J. Neves. Particle Swarms for Feedforward Neural Network Training. In *Proceedings of The 2002 International Joint Conference on Neural Networks (IJCNN 2002)*, pages 1895–1899, Honolulu, Havai, USA, IEEE Computer Society, May 2002.