Mortality assessment in intensive care units via adverse events using artificial neural networks[1]

Álvaro Silva[a], Paulo Cortez[b⋆], Manuel Filipe Santos[b],
Lopes Gomes[c], José Neves[d]

[a] Serviço de Cuidados Intensivos, Hospital Geral de Santo António, Porto, Portugal
[b] Departamento de Sistemas de Informação, Universidade do Minho, Guimarães, PORTUGAL
[c] Clínica Médica I, Inst. de Ciências Biomédicas Abel Salazar, Porto, Portugal
[d] Departamento de Informática, Universidade do Minho, Braga, PORTUGAL

⋆ Corresponding author:

Paulo Cortez
Tel: +351-253-510313; fax: +351-253-510300.
E-mail: pcortez@dsi.uminho.pt
Departamento de Sistemas de Informação, Universidade do Minho, Campus de Azurém, 4800-058 Guimarães, PORTUGAL

Summary

Objective: This work presents a novel approach for the prediction of mortality in intensive care units (ICUs) based on the use of adverse events, which are defined from four bedside alarms, and artificial neural networks (ANNs). This approach is compared with two logistic regression (LR) models: the prognostic model used in most of the European ICUs, based on the simplified acute physiology score (SAPS II), and a LR that uses the same input variables of the ANN model.

Materials and Methods: A large dataset was considered, encompassing forty two ICUs of nine European countries. The recorded features of each patient include the final outcome, the case mix (e.g. age) and the intermediate outcomes, defined as the daily averages of the out of range values of four biometrics (e.g. heart rate). The SAPS II score requires seventeen static variables (e.g. serum sodium), which are collected within the first day of the patient's admission. A nonlinear least squares method was used to calibrate the LR models while the ANNs are made up of multilayer perceptrons trained by the RPROP algorithm. A total of 13164 adult patients were randomly divided into training (66%) and test (33%) sets. The two methods were evaluated in terms of receiver operator characteristic (ROC) curves.

Results: The event based models predicted the outcome more accurately than the currently used SAPS II model ($P < 0.05$), with ROC areas within the ranges $83.9 - 87.1\%$ (ANN) and $82.6 - 85.2\%$ (LR) vs 80% (LR SAPS II). When using the same inputs, the ANNs outperform the LR (improvement of $1.3 - 2\%$).

Conclusion: Better prognostic models can be achieved by adopting low cost and real-time intermediate outcomes rather than static data.

Keywords: Artificial neural networks; Classification; Data mining; Intensive care; Logistic regression;

[1] Á. Silva, P. Cortez, M. F. Santos, L. Gomes and J. Neves. Mortality assessment in intensive care units via adverse events using artificial neural networks. In *Artificial Intelligence in Medicine* Journal, Elsevier, In press, ISSN:0933-3657.

1 Introduction

In the last decades, there has been an increasing development in intensive care medicine, where the goal is to provide the best outcome for critically ill patients. Indeed, a worldwide expansion occurred in the number of intensive care units (ICUs) [1]. Moreover, scoring the severity of illness has become a daily practice, with several metrics available, such as the acute physiology and chronic health evaluation system (APACHE II), the simplified acute physiology score (SAPS II) or mortality probability model (MPM), just to name a few [2].

The intensive care improvement comes with a price, being ICUs responsible for an increasing percentage of the health care budget. Resource availability limitations force them to make sure that intensive care is applied only to those who are likely to benefit from it. Critical decisions include interrupting life-support treatments and writing do-not-resuscitate orders when intensive care is considered futile. Under this context, mortality assessment is a crucial task, being used not only to predict the final clinical outcome but also to evaluate the ICU effectiveness. The prevalent prognostic models are built using a logistic regression over a *static* score (e.g. SAPS II); i.e., computed with data collected only within the first twenty four hours of the patient's admission. This limits the impact of clinical decision making, since the scores are usually not updated during the patients' length of stay.

On the other hand, the use of data mining in medicine is a rapidly growing field, which aims at discovering some structure in large clinical heterogeneous data [3]. This interest arose due to the rapid emergence of electronic data management methods, holding valuable and complex information. Human experts are limited and may overlook important details, while automated discovery tools can analyze the raw data and extract high level information for the decision-maker [4].

The artificial neural networks (ANNs) are one of the most successful data mining techniques, denoting a set of connectionist models inspired by the behavior of the human brain and presenting useful capabilities for medicine such as nonlinear learning, multi-dimensional mapping and noise tolerance [5]. The interest in ANNs was stimulated by the advent of the backpropagation algorithm in 1986. Since then, the number of ANN publications in Medicine has spawned from two in 1990 to five hundred in 1998 [6] and the search term "neural network computer" in the MEDLINE database displays more than two thousand articles from 1999 to 2004.

In the past, there has been work comparing ANNs and logistic regression models for ICU mortality prediction, reporting either better [7][8] or similar [9][10][11] performances. Yet, in all these studies, the ANNs were trained with the static variables used by the APACHE II score. This work follows an alternative direction, the use of data collected after the first twenty four hours of a patients' admission. A similar approach has been proposed by Kayaalp and his collaborators [12][13] where they adopted a time series prediction point of view, using twenty three temporal fields, such as the daily sequential organ failure assessment (SOFA) score, which takes time and costs to be obtained. However, in previous work [14], it has been shown that the SOFA can be replaced by real-time and less costly outcomes, known as events, which are automatically measured as out of range values of four commonly bedside monitored physiological parameters. Hence, this article presents a novel approach for ICU mortality prediction, based on the use of daily intermediate events.

A final remark will be given to the relation between prognostic models and treatments. A prognostic scoring system should be independent of treatment, providing a means of measuring disease or health status of a patient, where usually a higher score

corresponds to greater severity. Different elements contribute to this total score, including physiological variables. Although therapeutical approaches may influence the final patient outcome, their effect will be reflected upon the average number of events per day, the alarms signs. By augmenting a prognostic model with utility assessments of potential outcomes and indicating particular variables for decision support, optimal decisions for a group or individual patients can be determined [15].

The paper is organized as follows: first, the ICU clinical data is presented and the prognostic models are introduced (Section 2); next, a description of the performed experiments is given, being the results analyzed and discussed (Section 3); finally, closing conclusions are drawn (Section 4).

2 Materials and methods

2.1 Clinical data

This work adopted part of data collected during the EURICUS II project [16], which involved forty two ICUs of nine European Union countries, from November/98 to August/99. The patient's data was manually collected and registered by the nursing staff. In every hour, the monitored bedside parameters were introduced into daily patient records. The whole data was gathered at the Health Services Research Unit of the Groningen University Hospital, the Netherlands. The final database presented one entry (or example) per patient. After a consult with ICU experts, the patients with age lower than eighteen, burned or with bypass surgery were discarded. It should be noticed that these last two classes of patients are often not treated in ICUs but in specialized areas (e.g. burn or coronary units). In addition, four entries were removed due to the presence of missing values [17], remaining a total of 13164 records.

The main features of the clinical data are described in Table 1. The case mix appears in the first four rows, an information that remains unchanged during the patient's admission. The frequency distributions (or histograms) related to these variables are plotted in Figure 1, which also includes the patient's length of stay and the final outcome. Next, there are twelve variables related to the intermediate outcomes, which are defined from four monitored biometrics: the systolic blood pressure (BP), the heart rate (HR), the oxygen saturation (O2) and the urine output (UR). Finally, the last attribute denotes the patients' final outcome.

In a UCI, there are several events that may affect the patient's condition (e.g. shock, extemporary extubation or hypoxia). For the selection of an event, it is important that its occurrence and duration can be registered by physiological changes (e.g. shock and not pneumonia). Moreover, such physiological variables should be commonly registered in the ICUs, at regular intervals. These are the main reasons for the choice of the four biometrics. A panel of seven EURICUS II experts elaborated a protocol that defines the normal ranges for these parameters (see Table 2). When an out of range value occurs for a given interval, an alarm is triggered, defining an event. A critical event is a more serious event and it is classified by a longer event or a more extreme out of range measurement. Thus, each event or critical event is defined as a binary variable, which will be set to 0 (false), if the physiologic value lies within the advised range; or 1 (true) else, according to the time criterion.

The first eight outcomes of Table 1 denote the number of events/critical events, while the latter ones denote the total time in minutes considering only the critical

Table 1: The attributes of the intensive care data.

Attribute	Description	Domain Values
SAPS II	SAPS II score	$\{0, 1, \ldots, 163\}$
age	Patients' age	$\{18, 19, \ldots, 100\}$
admtype	Admission type	$\{1, 2, 3\}^a$
admfrom	Admission origin	$\{1, 2, \ldots, 7\}^b$
NBP	Average number of blood pressure events	$[0.0, \ldots, 33.0]$
NCRBP	Average number of critical blood pressure events	$[0.0, \ldots, 6.0]$
NHR	Average number of heart rate events	$[0.0, \ldots, 42.0]$
NCRHR	Average number of critical heart rate events	$[0.0, \ldots, 6.0]$
NO2	Average number of oxygen events	$[0.0, \ldots, 28.0]$
NCRO2	Average number of critical oxygen events	$[0.0, \ldots, 6.0]$
NUR	Average number of urine events	$[0.0, \ldots, 38.0]$
NCRUR	Average number of critical urine events	$[0.0, \ldots, 8.0]$
TCRBP	Average time (min.) of blood pressure critical events	$[0.0, \ldots, 36.0]$
TCRHR	Average time (min.) of heart rate critical events	$[0.0, \ldots, 33.0]$
TCRO2	Average time (min.) of oxygen critical events	$[0.0, \ldots, 33.0]$
TCRUR	Average time (min.) of urine critical events	$[0.0, \ldots, 40.0]$
death	The occurrence of death	$\{0, 1\}^c$

[a] 1 - Non scheduled surgery, 2 - Scheduled surgery, 3 - Physician.
[b] 1 - Surgery block, 2 - Recovery room, 3 - Emergency room, 4 - Nursing room, 5 - Other ICU, 6 - Other hospital, 7 - Other sources.
[c] 0 - No death, 1 - Death.

events. The whole twelve outcomes were analyzed as daily averages using the patients' records, since this allows prediction at early stages of the patients' admission. During the patients' length of stay in the ICU, intermediate outcomes of the first day can feed into a real-time prognostic model; in next day, an accumulative knowledge is built by averaging current events with the ones obtained from the previous day; and so on.

Table 2: The protocol for the out of range measurements.

	BP	**O2**	**HR**	**UR**
Normal Range	$90 - 180$mmHg	$\geq 90\%$	$60 - 120$bpm	≥ 30ml/h
Event[a]	≥ 10min.	≥ 10min.	≥ 10min.	≥ 1h
Event[b]	≥ 10min. in 30min.	≥ 10min. in 30min.	≥ 10min. in 30min.	–
Critical Event[a]	≥ 1h	≥ 1h	≥ 1h	≥ 2h
Critical Event[b]	≥ 1h in 2h	≥ 1h in 2h	≥ 1h in 2h	–
Critical Event[c]	< 60mmHg	$< 80\%$	< 30bpm $\vee > 180$bpm	≤ 10ml/h

[a] Defined when continuously out of range.
[b] Defined when intermittently out of range.
[c] Defined anytime.

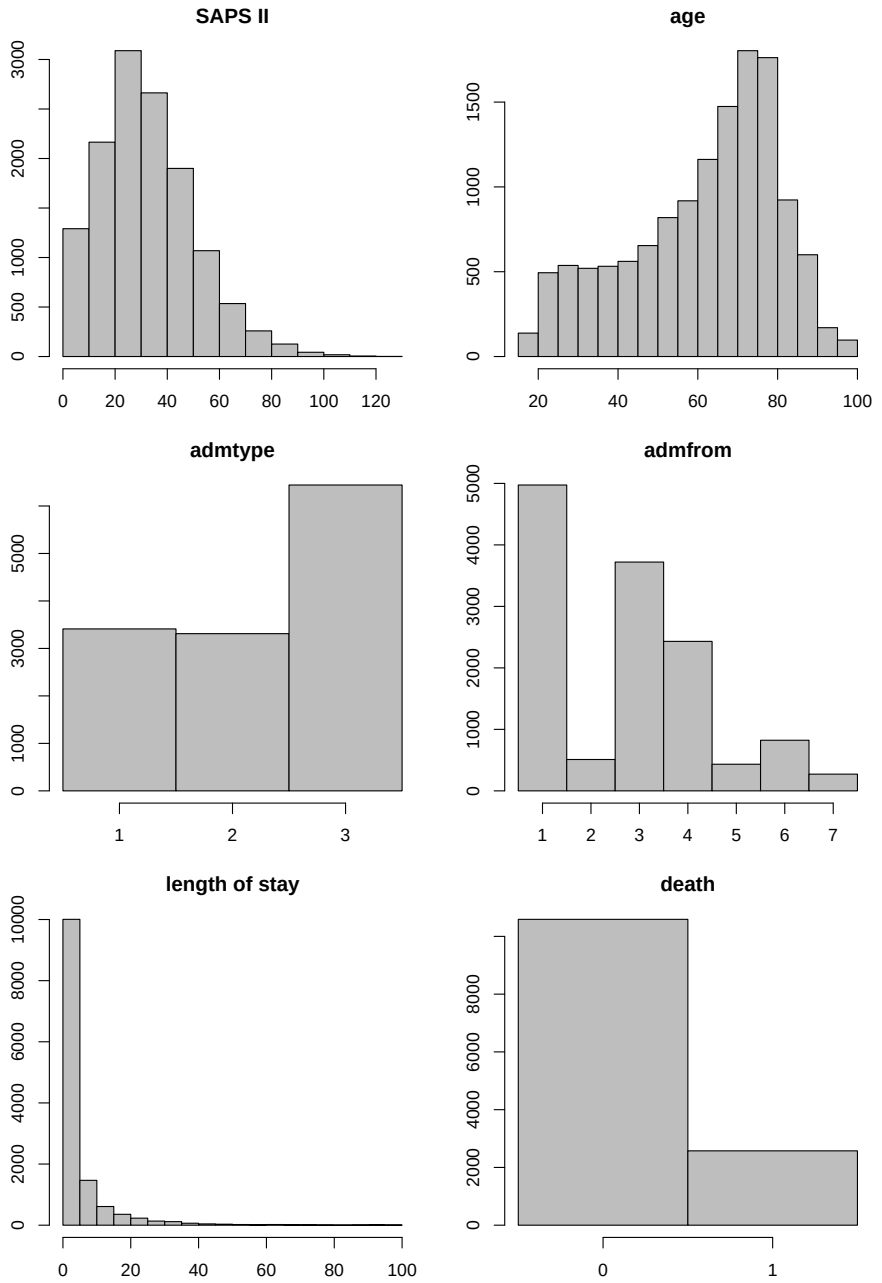

Figure 1: The histograms for the case mix (SAPS II, age, admtype and admfrom) variables, the length of stay and the final outcome (death).

2.2 Logistic regression

The logistic regression is a statistical method that is commonly used within the Medicine field to model the probability of binary events (e.g. the proportion of patients that respond to a given therapy). It operates a smooth non linear transformation, where the probability of the event k is given by [4]:

$$P_k = \frac{1}{1 + exp(\beta_0 + \sum_{i=1}^{N} \beta_i x_i)} \quad (1)$$

where $\beta_0, \ldots, \beta_N$ denote the parameters of the model and $x_1, \ldots, x_N$ the dependent variables.

On the other hand, the SAPS II is the most widely European score for mortality assessment, ranging from 0 to 163, where a high value indicates a high death probability [18]. The model encompasses a total of seventeen variables (Table 3), such as age, previous health status or diagnosis, which are collected within the first twenty four hours of the patient's admission. Some of these variables are associated to clinical tests (e.g. blood samples) that are expensive.

Table 3: The SAPS II variables (adapted from [18]).

Attribute	**Description**
age	Patients' age (in years)
heart rate	Worst value in 24 hours (either low or high)
systolic blood pressure	Worst value in 24 hours (either low or high)
body temperature	Highest temperature in degrees
ventilation	Lowest PaO_2/FiO_2 ratio if ventilated
urinary output	Total urinary output in 24 hours
serum urea level	Highest value in mmol/L, g/L or mg/dL
WBC count	Worst WBC count from the scoring sheet
serum potassium level	Worst value in mmol/L
serum sodium level	Worst value in mmol/L
serum bicarbonate level	Lowest value in mEq/L
bilirubin	Highest value in μmol/L or mg/dL
Glasgow coma score	Lowest value
admission	Unscheduled surgical, scheduled surgical or medical
AIDS	Yes, if HIV-positive with illness
hematologic malignancy	Yes, if lymphoma, acute leukemia or multiple myeloma
metastatic cancer	Yes, if proven by surgery or any other method

The prognostic model, known as predictive death rate (pdr_k), is given by the logistic regression:

$$\begin{array}{rcl} logit_k & = & B_0 + B_1 \times SAPSII_k + B_2 \times ln(SAPSII_k + 1) \\ pdr_k & = & exp(logit_k)/(1 + exp(logit_k)) \end{array} \quad (2)$$

where $SAPSII_k$ denotes the score for the patient k, being B_0, B_1 and B_2 internal parameters estimated by calibration procedures. The majority of the European ICUs use the values $B_0 = -7.7631$, $B_1 = 0.0737$ and $B_2 = 0.9971$, which were optimized in the study of Le Gall and his collaborators [18] by applying a Hosmer-Lemeshow goodness-of-fit test to an international database with a total of 13152 examples collected from September/91 to February/92. The predicted class (P_k) for the k patient is given by the nearest class value to a decision threshold D within the range $[0.0, 1.0]$:

$$P_k = \begin{cases} 0 \ , \ if \ pdr_k < D \\ 1 \ , \ else \end{cases} \quad (3)$$

where the class values $P_k = 0$ denotes *no death* and $P_k = 1$ *death*.

2.3 Artificial neural networks

The multilayer perceptron is one of the most popular ANN architectures. It consists of a feedforward network, where processing neurons are grouped into layers and connected by weighted links [5]. Supervised learning is achieved by an iterative adjustment of the network connection weights, called the training procedure, in order to minimize an error function, which is computed over a set of training examples (E). Each $p \in E$ example maps an input vector $(x_1^p, \ldots, x_N^p)$ with a vector of target values $(t_1^p, \ldots, t_M^p)$, where N and M denote the number of the input and output neurons. Typically, the sum squared error (SSE) is used as the cost function [19]:

$$SSE = \sum_{p \in E} \sum_{i \in M} (t_i^p - s_i^p)^2 \tag{4}$$

where s_i^p is the ANN output value.

The network is activated by feeding the input layer with the input vector and then propagating the activations in a feedforward fashion, via the weighted connections, through the entire network. For a given input $(x_1, \ldots, x_N)$, the state of a neuron (s_i) is computed by:

$$s_i = f(w_{i,0} + \sum_{j \in I} w_{i,j} \times s_j) \tag{5}$$

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; and $s_1 = x_1, \ldots, s_N = x_N$. The $w_{i,0}$ connections, called bias, work as a constant term, since their input is always 1.0. Bias connections are normally added to multilayer perceptrons, since they increase the network learning flexibility; i.e., the hyperplanes defined by the neurons are not constrained to pass through the origin [20].

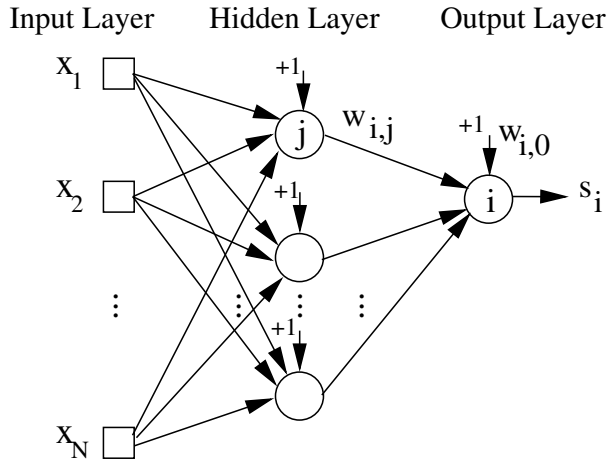

Figure 2: A fully connected network one hidden layer, 1 output node and bias connections.

Multilayer perceptrons with bias connections, one hidden layer with a fixed number of hidden nodes, one output node and logistic activation functions $f(x) = 1/(1+exp(-x))$ were adopted (Figure 2). The predicted class (P_k) for the k example is given by:

$$P_k = \begin{cases} 0 \ , \ if \ ANN_k < D \\ 1 \ , \ else \end{cases} \tag{6}$$

where ANN_k denotes the neural output value for the k patient.

Before training the ANNs, the data was preprocessed: the input values were standardized into the range $[-1.0, 1.0]$ and an *1-of-C* encoding, which uses one binary variable per each class, was applied to the nominal (non ordered) attributes with few categories (**admtype** and **admfrom**). For example, the **admtype** variable is fed to 3 input nodes, according to the scheme: $1 \rightarrow (-1.0\ -1.0\ 1.0)$; $2 \rightarrow (-1.0\ 1.0\ -1.0)$; and $3 \rightarrow (-1.0\ -1.0\ 1.0)$.

At the beginning of the training process, the network weights are randomly set within the range $[-1.0, 1.0]$. Then, the resilient propagation (RPROP) algorithm is applied during the training, with the aim of minimizing the SSE. The RPROP is an enhanced version of the backpropagation algorithm, performing a local adaptation of the weight-updates based on behavior of the error function, given by the local gradient information. A full description of the RPROP details can be found in [19]. This algorithm contains two parameters, which were set to the RPROP advised values: $\Delta_0 = 0.1$ and $\Delta_{max} = 50.0$. Nevertheless, the choice of these parameters is rather uncritical, since the error convergence is usually insensitive to the Δ_0 and Δ_{max} values. Furthermore, when compared with other algorithms, such as the standard backpropagation, the RPROP presents a faster training, requiring less computational effort [19][21]. Finally, the training is stopped when the error slope is approaching zero or after a maximum of 100 epochs [22].

When compared to other data mining techniques, such as decision trees, ANNs often present a higher predictive accuracy [23]. However, in data mining applications, besides obtaining a high predictive performance, it is also important to provide explanatory knowledge; i.e., what has the model learned. With ANNs this can be given after the training process, by measuring the importance of the inputs and extracting a set of rules.

In the past, several approaches have been proposed to measure the importance of inputs, such as the weight-elimination algorithm [24], which is based in the network smallest weights. However, a huge (tiny) input-to-hidden weight does not necessarily mean that the input has a huge (tiny) effect on the output, since the weight effect is limited by the squashing functions of the hidden neurons [25]. A better alternative is to use sensitivity analysis [26], which has outperformed other input selection techniques, such as forward selection and genetic algorithms. It can be measured as the variance (V_a) produced in the output (y) when the input attribute (a) is moved through its entire range:

$$\begin{array}{rcl} V_a & = & \sum_{i=1}^{L} (y_i - \overline{y})/(L-1) \\ R_a & = & V_a / \sum_{j=1}^{A} V_j \end{array} \quad (7)$$

where A denotes the number of input attributes and R_a the relative importance of the a attribute. The y_i output is obtained by holding all input variables at their average values. The exception is x_a, which varies through its range with L levels. In this work, L was set to two for the binary attributes ($x_a \in \{-1.0, 1.0\}$) and eleven for the continuous inputs ($x_a \in \{-1.0, -0.8, -0.6, \ldots, 0.8, 1.0\}$).

The extraction of knowledge from ANNs is still an active research area [23]. Currently, there are two main approaches [27]: the use of decompositional and pedagogical techniques. The former models the ANN at the minimum level of granularity: first, rules are extracted from each individual neuron (hidden and output); then, the subsets of rules are aggregated to form a global relationship. The latter algorithms extract the direct relationships between the inputs and outputs of the

ANN. By using a black-box point of view, less computation effort is required and a simpler set of rules may be achieved.

2.4 Statistics

Thirty runs were applied in all experiments. The results are shown in terms of the mean and t-student 95% confidence intervals. The accuracy estimates are achieved using the Holdout method [28]: in each simulation, the available data is divided into two mutually exclusive partitions, using stratified sampling. The training set is used during the modeling phase, while the test set is adopted after training, in order to compute the accuracy estimates.

A common tool for classification analysis is the confusion matrix which matches the test results (predicted) and patients real condition (actual) values [29]. From the matrix, three measures of performance can be defined [30]: the sensitivity (Sen), also known as recall and type II error; the specificity (Spe), also known as precision and type I error; and the accuracy (Acc), also called correct classification rate, which gives an overall evaluation. The receiver operating characteristic (ROC) curve shows the performance of a two class classifier across the range of possible threshold (D) values, plotting one minus the specificity (x-axis) versus the sensitivity (y-axis) [31]. The global accuracy is given by the area under the curve ($AUC = \int_0^1 ROC dD$). A random classifier will have an AUC of 0.5, while the ideal value should be close to 1.0.

3 Results

3.1 Training setup

All experiments reported in this work related to the ANNs, including the RPROP algorithm, were conducted using an object oriented programming environment developed in *JAVA* [32]. On the other hand, the logistic regression models and the statistics, including the Acc, Spe, Sen, AUC and t-student tests; were computed using the R statistical environment [33].

The commonly used 2/3 and 1/3 partitions were adopted for the definition of the training and test set sizes [28]. A model is said to overfit when it correctly handles the training data but fails to generalize. Usually, overfitting is a critical issue in ANN modeling. As pointed out by Sarle [34], the best way to avoid overfitting is to use a large and diverse dataset. Since this is such a case, generalization loss is unlikely to occur, specially if small ANNs are adopted. To accomplish this and also to reduce the computational searching space, the number of hidden nodes was set to $round(N/2)$, where N denotes the number of input nodes. It should be stressed that this rule does not constitute the optimal solution. However, in data mining applications, it is often more important to select the training data carefully than decisions regarding learning algorithms [35]. In effect, previous experiments [14] have shown that: the number of hidden nodes has a minor impact, when compared with filtering or feature selection setups; and with the $round(N/2)$ setup overfitting is not an important issue.

In the preliminary experiments, all inputs were considered. The exception is the age attribute, which was not included since it is already used by the SAPS II metric. The decision threshold was set to the middle of its range ($D = 0.5$). The table also presents the AUC values, which where computed using a 0.005 grid search for the

D value, with a total of 201 decision points. To prevent *tuning*, the adjustment of the models' parameters to the test performances, the training configuration will be improved using only training errors (Table 4).

Table 4: The training set performances (in percentage).

Setup	*Acc*	*Sen*	*Spe*	AUC
Constant Predictor	80.45±0.00	0.00±0.00	100.00±0.00	50.00±0.00
Normal	85.75±0.10	41.78±0.34	96.43±0.08	84.02±0.14
Balanced	78.22±0.20	76.45±0.37	79.99±0.36	86.42±0.19
Log Balanced	79.87±0.19	78.99±0.22	80.76±0.26	88.15±0.19

The constant predictor is defined as the most prevalent outcome. This naive classifier provides an accuracy of 80.45%, corresponding to the proportion of *no death* cases. The first neural simulation (Normal) presents a better accuracy (85.75%) when compared with the former setup. Yet, from the sensitivity point of view, the first two configurations are poor classifiers. Since there is a higher number of false conditions in the training set, the learning process will tend to favor such cases. Although popular within the data mining community, the accuracy measure is not sufficient in Medicine, where a test should report both high sensitivity and specificity values [30]. To solve this handicap, a common solution is to balance the training data [17], by using a filtering method that selects an equal number of true and false examples. Therefore, a balanced setup was devised by adopting a under sampling procedure where false (*no death*) examples were randomly deleted from the training set. This setup obtained better results, with a double increase in the sensitivity (34.7%) when compared to the decrease in the specificity (16.4%). Finally, since all outcomes presented high positive skewed distributions, with mean values close to zero, a third configuration was tested (Log Balanced), where the intermediate outcomes suffered the logarithm transform $ln(x + 1)$. Since this last setup gave the best results, it will be the selected training technique. It should be noted that paired t-tests confirmed statistical differences ($P < 0.05$) between the setups for all measures.

3.2 Test set performances

The results reported on this section were computed over the test sets, which contain the 1/3 samples that were not used in any part of the training process. Each test set also keeps the original proportions of death cases, which is 19.55%. The comparison will be performed between artificial neural network (ANN) and logistic regression (LR) models. In total, eight distinct models will be evaluated (Table 5):

ANN/LR ALL - similar to the Log Balanced setup, where all inputs are used except the age;

ANN/LR Case Outcomes - with all inputs except the SAPS II;

ANN/LR Outcomes - which only uses the intermediate outcomes;

LR SAPS II - the most used European prognostic model [18], defined by Equation 2, with the fixed parameters $B_0 = -7.7631$, $B_1 = 0.0737$ and $B_2 = 0.9971$; and

LR SAPS II B - equivalent to the previous model, except that in each run, the parameters (B_0, B_1 and B_2) are calibrated to a balanced training.

The LR models (except LR SAPS II) were optimized by a nonlinear least squares algorithm that minimizes the SSE, under the R environment [33].

Table 5: The test set performances (in percentage).

Setup	*Acc*	*Sen*	*Spe*	AUC
ANN All	79.21±0.24	78.11±0.51	79.48±0.35	87.12±0.21
ANN Case Outcomes	78.22±0.26	75.78±0.66	78.82±0.36	85.52±0.20
ANN Outcomes	77.60±0.31	70.00±0.59	79.45±0.48	83.88±0.23
LR All	75.97±0.29	77.06±0.68	75.71±0.40	85.17±0.27
LR Case Outcomes	77.15±0.26	70.63±0.67	78.73±0.41	83.78±0.24
LR Outcomes	77.57±0.15	65.91±0.62	80.41±0.21	82.59±0.23
LR SAPS II	82.60±0.14	42.57±0.50	92.33±0.12	79.84±0.26
LR SAPS II B	69.37±0.32	77.57±0.64	67.37±0.48	80.04±0.25

The results are shown in Table 5, while the correspondent ROCs are plotted in Figure 3. Paired t-tests were also performed, confirming statistical differences ($P < 0.05$) for the AUC metric between all models, except the pairs: LR SAPS II/LR SAPS II B ($P = 0.27$); ANN Outcomes/LR Case Outcomes ($P = 0.40$); and ANN Case Outcomes/LR All ($P = 0.13$).

First, the last two setups will be analyzed. Due to the balanced training, the second logistic setup presents higher sensitivity values than the first one. Yet, in terms of the *ROC* analysis, and despite of using quite different learning procedures, the two logistic models are statistically equivalent. Furthermore, both setups are outperformed by all models that use the intermediate events, with a difference ranging from 2.19% to 6.72%. This is an important result, since it backs the main claim of this work: the selected daily events do provide useful information for predicting ICU mortality.

When using the same inputs, the ANNs outperform the LR models, with improvements of 1.95% (**All**), 1.74% (**Case Outcomes**) and 1.29% (**Outcomes**). Regarding the feature selection, the best option, in terms of predictive performance, is to use all attributes. When replacing the SAPS II attribute by the age (setup **Case Outcomes**), there is a decay of 1.60% (ANN) and 1.39% (LR). A similar effect occurs when only the intermediate outcomes are used, with a 1.64% (ANN) and 1.9% (LR) decrease in terms of the AUC values. It is also important to stress that if the AUC is above 80% then the discriminator is considered excellent [18]. Since the first six setups are above this limit, this means that high quality results were achieved.

3.3 Explanatory knowledge

Table 6 shows the relative importance of the inputs for the event based models when applying the sensitivity analysis procedure (Equation 7). To improve clarity, the importance of the attributes related to each biometric were summoned into a single value (last four columns). The obtained results are consistent for both ANN and LR models. For the **All** selection, the event based variables denote a strong

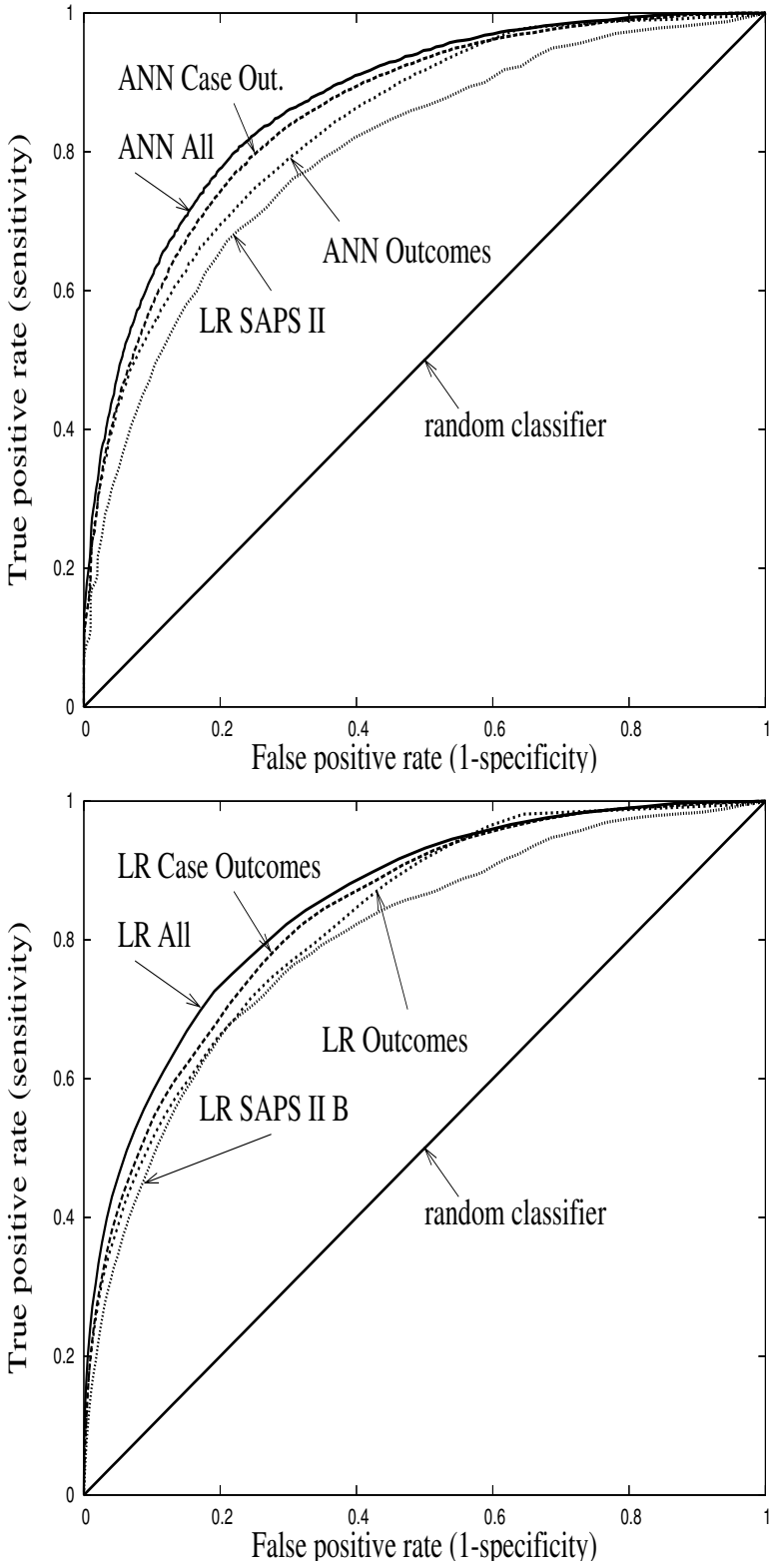

Figure 3: The receiver operating characteristic curves.

influence, summing a total of 80.2% (ANN) and 76.5% (LR). Furthermore, the table suggests that the case mix is not needed in this configuration, a scenario that

Table 6: The relative importance of the input variables (in percentage).

Setup	SAPS II	age	admtype	admfrom	BP⋆	HR⋆	O2⋆	UR⋆
ANN All	16.8	–	1.0	2.0	15.9	14.4	30.7	19.2
ANN Case Outcomes	–	14.3	5.9	11.7	13.1	16.7	22.5	15.8
ANN Outcomes	–	–	–	–	16.9	15.8	21.8	45.5
LR All	19.3	–	2.8	1.4	17.9	16.2	23.1	19.3
LR Case Outcomes	–	19.2	3.6	13.4	15.3	14.0	18.6	15.9
LR Outcomes	–	–	–	–	24.1	18.6	26.5	30.8

⋆ All attributes related to the variable where included: number of events/critical events and the time.

changes when the SAPS II is excluded (setup **Case Outcomes**). When comparing the bedside parameters, the importance order changes from model to model. The oxygen saturation is the most relevant biometric in the first two configurations. The urine output comes in second place for **ANN/LR All** setups and presents the the highest influence when only event based data is used.

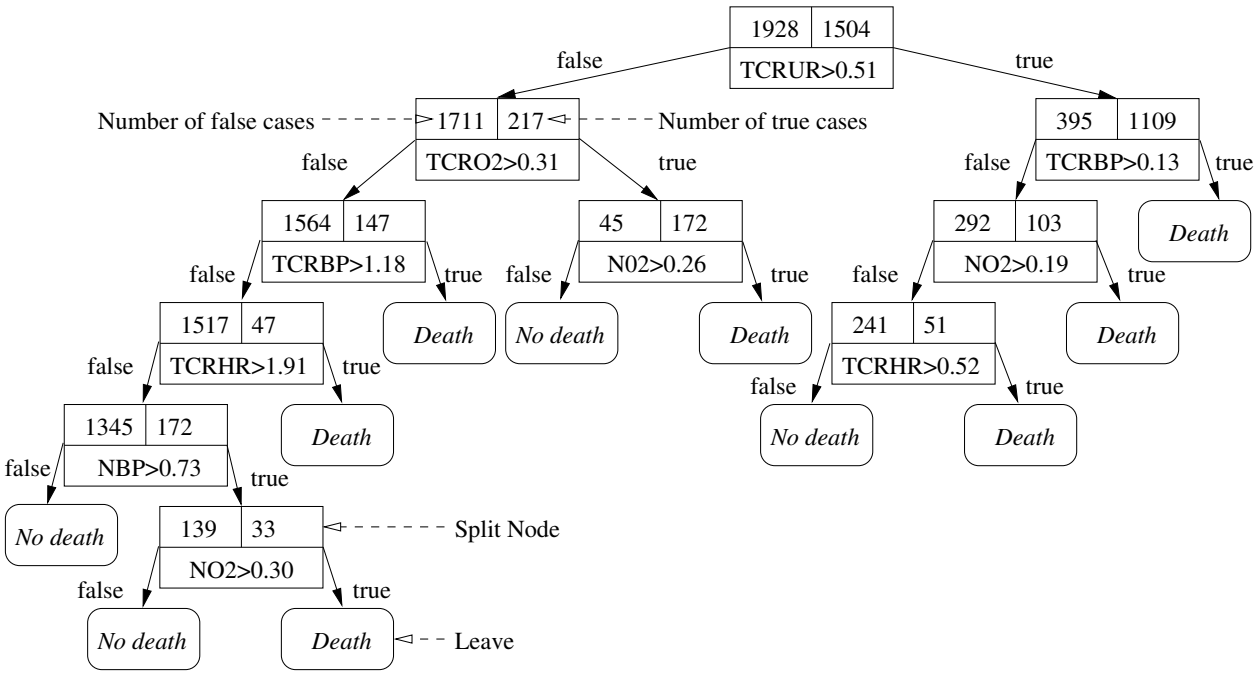

Figure 4: Example of a decision tree extracted from a trained neural network. Variables denote the logarithm transform $ln(x+1)$ of the intermediate outcomes from Table 1.

For demonstrative purposes, a simple pedagogical rule extraction technique will be applied to the ANN setup that only includes the intermediate outcomes. The intention is to provide a descriptive model of what a trained ANN has learned. The rules will be presented in terms of a decision tree [36], a branching structure based on

split nodes, that test a given feature, and leaves, which assign a class label. First, a new dataset was constructed, by considering the inputs of a given balanced training set, in a total of 3432 cases; and the corresponding ANN output class labels. Then, a decision tree algorithm was applied to this dataset by using the rpart library of the R environment [33]. The obtained decision tree managed to explain the ANN behavior with an accuracy of 90%. To simplify the visualization, the tree was pruned to a maximum of ten splits (Figure 4). As an example, the following rule corresponds to the top right path of the tree, explaining a total of 1109 cases:

$$\textbf{if } ln(TCRUR + 1) > 0.51 \wedge ln(TCRBP + 1) > 0.13 \textbf{ then } Death$$

It is interesting to notice that the most important attribute is the time of urine critical events (TCRUR), which appears at the root node. At the next levels, the relevant features are the oxygen saturation (TCRO2) and blood pressure (TCRBP), while the heart rate outcome only appears at the fourth level. Hence, the decision tree corroborates the sensitivity analysis procedure for input relevance (third row of Table 6).

4 Conclusion

The surge of data mining techniques, such as artificial neural networks (ANN), has created new exciting possibilities in medicine. In this work, these techniques were applied for the prediction of death in intensive care units (ICUs) by using daily intermediate outcomes, which are defined from four commonly bedside monitored variables. In contrast, the current ICU prognostic models, based on a logistic regression (LR), use data collected only in the first day of the patient's admission. Both approaches were tested in a large database with a total of 13164 records taken from forty two ICUs of nine European countries.

Experiments were drawn to test several training configurations, being the best performances obtained by the use of a balanced training procedure and a logarithm transform on the intermediate outcomes. The test set results clearly favor the event based models, when compared with current ICU models. Furthermore, when using the same input variables, the ANNs presented better performances than the LR methods, with improvements within the range 1.29-1.95%. Moreover, several input selection combinations were also considered, being the best result obtained by the use of all input variables except the patients' age. However, since the SAPS II score requires time and costs to be obtained, it can be replaced by the age, provided that a decrease of 1.39-1.60% in performance is acceptable.

The presented approach aims at improving clinical decision-making by providing the best estimate of the patient's condition. It is not intended for evaluating ICUs, since a bad performing ICU will produce more critical events and consequently the poor performance will not be identified. The results have been analyzed under the receiver operator characteristic (ROC) curves, which plots one minus the specificity versus the sensitivity, when varying a decision threshold ($D \in [0.0, 1.0]$). Globally, high quality results were achieved, with area under the curve (AUC) values higher than 80%. In a real scenario, the decision threshold could be optimized for distinct purposes. High sensitivity models would allow an early identification of deteriorating patients, while high specificity ones would be essential for interrupting life-support treatments.

Nowadays, the majority of the European ICUs adopted the SAPS II prognostic model, which is based on a logistic regression. Yet, this model was outperformed by all event based models (ANN and LR). Furthermore, the proposed approach (setup **Case Outcomes**) requires less variables. The SAPS II index was developed in the last decade and currently there is an intensive research for the development of its successor, the SAPS III [38]. This work shows that a different direction should be pursued: the use of intermediate information rather than static data.

Regarding the comparison between the logistic regression and ANNs, although the latter present a higher accuracy, the former method is more easy to interpret. Yet, it is possible to extract human friendly rules from trained ANNs, as shown in the previous section. Nevertheless, if the physicians are not comfortable with the ANNs, then the **LR Case Outcomes** setup is advised.

This study was based on an *off-line* learning, since the data mining techniques were applied after the conclusion of the EURICUS II project [16]. However, the proposed approach opens room for the development of tools for clinical decision support, which are expected to enhance the physician response and proactiveness. Although the data was manually collected, it could be acquired with a low cost and in real-time, since the events can be automatically fired from the four monitored variables. However, in a real environment there are several phenomena which may cause bad readings (e.g. spurious heart rates may be caused by loose leads). In such scenarios, data quality and cleansing procedures should be addressed [37] and/or the data could be validated by the ICU staff. It is also interesting to notice that such system could potentially provide more updated predictions (e.g. every hour).

Another important aspect is related to the concept of *lead time*: how long before the patient died was the system able to predict the death? Since the full daily records were not available for this study, this question will be addressed in future work. Indeed, it is intended to extend this approach to an *on-line* learning environment, by attaching computer systems with friendly human interfaces into ICUs, with the capability to learn and respond in real-time. This will allow us to obtain, after some time, a valuable feedback from the physicians, in terms of trust and acceptance of this alternative solution. Currently, a pilot project, called *INTCare*, is being developed under this perspective at the ICU of the Hospital Geral de Santo António, Oporto, Portugal.

Acknowledgments

We thank FRICE and the BIOMED project BMH4-CT96-0817 for the provision of part of the EURICUS II data and support for this study, which is integrated in a PhD program, developed at Instituto de Ciências Biomédicas Abel-Salazar from University of Porto and the Departments of Computer Science/Information Systems from the University of Minho. We also would like to thanks the anonymous reviewers for their helpful comments.

References

[1] C. Hanson and B. Marshall, Artificial intelligence applications in the intensive care unit, *Crit. Care Med.*, 29(2) (2001) 1–9.

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

[3] K. Cios and G. Moore, Uniqueness of Medical Data Mining, *Artificial Intelligence in Med.* 26 (2002) 1–24.

[4] D. Hand, H. Mannila and P. Smyth, *Principles of Data Mining* (MIT Press, Cambridge MA, USA, 2001).

[5] S. Haykin, *Neural Networks - A Compreensive Foundation* (Prentice-Hall, New Jersey, USA, 2nd edition, 1999).

[6] R. Dybowski, Neural Computation in Medicine: Perspectives and Prospects, in: H. Malmgreen, M. Borga and L. Niklasson, eds, *Artificial Neural Networks in Medicine and Biology* (Springer, Heidelberg, Germany, 2000) 26–36.

[7] R. Dybowski, P. Weller, R. Chang and V. Gant, Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *Lancet*, 347(9009) (1996) 1146–1150.

[8] A. Nimgaonkar and S. Sudarshan, Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Intensive Care Med.*, 30 (2004) 248–253.

[9] G. Doig, K Inman, W. Sibbald, M. Martin and J. Robertson, Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. in: C. Safran, ed, *proceedings of Annual Symposium on Computer Applications in Medical Care* (McGraw-Hill, New York, USA, 1993) 361–365.

[10] L. Wong and J. Young, A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks. *Anaesthesia*, 54 (1999) 1048–54.

[11] G. Clermount, D. Angus, S. DiRusso, M Griffin and W. Linde-Zwirble, Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit. Care Med.*, 29(2) (2001) 291–296.

[12] M. Kayaalp, G. Cooper and G. Clermont, Predicting ICU mortality: A comparison of stationary and nonstationary temporal models. in: J. Overhage, ed, *Converging Information, Technology, and Health Care*, *proceedings of AMIA Symposium* (AMIA, Los Angeles CA, USA, 2000) 418–422.

[13] M. Kayaalp, G. Cooper and G. Clermont, Predicting with Variables Constructed from Temporal Sequences. in: *proceedings of the Eight International Workshop on Artificial Intelligence and Statistics*, (Key West FL, USA, 2001) 220–225.

[14] Á. Silva, P. Cortez, M. Santos, L. Gomes and J. Neves, Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks, in: I. Seruca, J. Cordeiro, S. Hammoudi and J. Filipe, eds, *Enterprise Information Systems VI* (Springer, Germany, 2005).

[15] A. Abu-Hanna and P. Lucas, Prognostic Models in Medicine: AI and Statistical approaches. *Methods Inf. Med.*, 40 (2001) 1–5.

[16] R. Miranda, The effect of harmonising and standardising the nursing tasks on Intensive Care Units (ICU) of the European Community, Second European Intensive Care Unit Study (EURICUS-II). Technical Report BMH4-CT96-0817, Health Services Research Unit - Groningen University Hospital, The Netherlands, 1999.

[17] D. Pyle, *Data Preparation for Data Mining* (Morgan Kaufmann, S. Francisco CA, USA, 1999).

[18] J. Le Gall, S. Lemeshow and F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European / North American multicenter study, *JAMA* 270 (1993) 2957–2963.

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

[20] K. Hornik, Some new results on neural network approximation. *Neural Networks*, 6 (1993) 1069–1072.

[21] R. Mendes, P. Cortez, M. Rocha and J. Neves, Particle Swarms for Feedforward Neural Network Training, in: *proceedings of The 2002 Int. Joint Conference on Neural Networks*, 2 (IEEE, Honolulu, Havai, USA, 2002) 1895–1899.

[22] L. Prechelt, Early Stopping – but when?, in: G. Orr and K. Müller, eds, *Neural Networks: Tricks of the trade* (Springer, Heidelberg, Germany, 1998).

[23] R. Setiono, Techniques for Extracting Classification and Regression Rules from Artificial Neural Networks, in: *Computational Intelligence: The Experts Speak*, D. Fogel and C. Robinson, eds, (IEEE, Piscataway NJ, USA, 2003) 99–114.

[24] C. Ennett and M. Frize, Weight-elimination neural networks applied to coronary surgery mortality prediction, *IEEE Trans. Inf. Technol. Biomed.* 7(2) (2003) 86–92.

[25] N. Cardell, W. Joerding and Y. Li, Why Some Feedforward Networks Cannot Learn Some Polynomials, *Neural Computation* 6 (1994) 761–766.

[26] R. Kewley, M. Embrechts and C. Breneman, Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks, *IEEE Trans. on Neural Networks* 11(2000) 668–679.

[27] A. Tickle, R. Andrews, M. Golea and J. Diederich, The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks, *IEEE Trans. on Neural Networks* 9(6) (1998) 1057–1068.

[28] A. Flexer, Statistical evaluation of neural networks experiments: Minimum requirements and current practice, in: R. Trappl, ed, *proceedings of the 13th European Meeting on Cybernetics and Systems Research* 2 (Vienna, Austria, 1996) 1005–1008.

[29] R. Kohavi and F. Provost, Glossary of Terms, *Machine Learning* 30 (1998) 271–274.

[30] D. Essex-Sorlie, *Medical Biostatistics & Epidemiology: Examination & Board Review* (McGraw-Hill/Appleton & Lange, Norwalk CT, USA, Int edition, 1995).

[31] J. Hanley J. and B. McNeil, The meaning and use of the area under the Receiver Operating Characteristics (ROC) curve, *Radiology* 143 (1982) 29–36.

[32] M. Rocha, P. Cortez and J. Neves, Simultaneous Evolution of Neural Network Topologies and Weights for Classification and Regression, *Lecture Notes in Computer Science* 2902 (2005) 59–66.

[33] R Development Core Team, R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2004).

[34] W. Sarle, Stopped Training and Other Remedies for Overfitting. in: *proceedings of the 27th Symposium on the Interface of Computer Science and Statistics*, (Pittsburgh PA, USA, 1995) 352–360.

[35] N. Lavrac, H. Motoda, T. Fawcett, P. Langley and P. Adriaans, Introduction: Lessons learned from data mining applications and collaborative problem solving. *Machine Learning*, 57(1-2) (2004) 13–32.

[36] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees* (Wadsworth, Belmont CA, USA, 1984).

[37] E. Turban, J. Aronson and T. Liang, *Decision Support Systems and Intelligent Systems* (Prentice-Hall, UK, 2004).

[38] The SAPS III Outcomes Research Group, From the Evaluation of the Individual Patient to the Evaluation of the ICU. Available at http://www.saps3.org (2005).