

# Predicting Laboratory Testing in Intensive Care using Fuzzy and Neural Modeling

Federico Cismondi<sup>\*†‡</sup>, André S. Fialho<sup>\*†‡</sup>,  
Susana M. Vieira<sup>‡</sup>, João M.C. Sousa<sup>‡</sup>, Shane R. Reti<sup>§</sup>,  
Leo A. Celi<sup>||</sup>, Michael D. Howell<sup>¶</sup> and Stan N. Finkelstein<sup>\*</sup>

<sup>\*</sup>Massachusetts Institute of Technology, Engineering Systems Division,  
77 Massachusetts Avenue, 02139 Cambridge, MA, USA (email: cismondi@mit.edu)

<sup>†</sup>MIT Portugal Program, 77 Massachusetts Avenue, E40-221, 02139 Cambridge, MA, USA

<sup>‡</sup>Technical University of Lisbon, Instituto Superior Técnico, Dept. of Mechanical Engineering,  
CIS/IDMEC – LAETA, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

<sup>§</sup>Division of Clinical Informatics, Department of Medicine,  
Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

<sup>¶</sup>Silverman Institute for Healthcare Quality and Safety,  
Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

<sup>||</sup>Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

**Abstract**—Laboratory testing is a frequent activity for patients in intensive care units (ICU). Recent studies demonstrate that frequent laboratory testing does not necessarily relate to better outcomes. We hypothesize that unnecessary laboratory testing can be reduced by predicting which tests are unlikely to influence clinical management. Reducing unnecessary tests could reduce morbidity and hospitalization costs. We analyzed an ICU database containing 26,665 patient records at Beth Israel Deaconess Medical Center, Boston, and selected a subset of patients with gastrointestinal bleeding. Database knowledge discovery was applied involving data preprocessing, feature selection, and classification. Conventional soft computing tools such as fuzzy models and neural networks were utilized in this work, combined with statistical and mathematical tools. The input variables included bedside monitor trends, lab tests, arterial/central catheter information, urine collections, transfusions, indexes and scores calculated for the patients. The outcome variable was a binary classification based on falling levels of hematocrit. Feature selection was performed by a bottom-up strategy, maximizing the area under the ROC curve (AUC), the integrated discrimination improvement (IDI) and a multiobjective function primarily pondering the sensitivity of the models. A leave-one-out cross validation process was used to evaluate the overall models' performance, as well as the additional predictive value of the variables selected. Urine output was selected by all models as the best predictor of useful hematocrit testing. Our results show that it is possible to correctly classify the usefulness of a hematocrit lab test up to 81% of the time by using fuzzy models and neural networks.

**Index Terms**—Knowledge Discovery in Databases, Classification, Intensive Care Unit, Laboratory testing, Modeling, ICD9, Neural Networks, Fuzzy Models

## I. INTRODUCTION

Laboratory testing occurs frequently in hospitalized patients [1]. This is especially so for patients in intensive care where frequent blood draws may cause increased general phlebotomy complications such as blood loss [2], [3]. Although many lab tests are performed during the hyperacute first day of the

patient's ICU stay, about two-thirds of laboratory studies are ordered later in the stay. While part of this testing reflects the intrinsic critical status of ICU patients, other tests are run almost by default, following general guidelines and not patient specific characteristics [4].

Recent data demonstrate that patients who have received more frequent laboratory tests do not necessarily have better outcomes [5]. On a cost comparative basis, patients with the highest discretionary laboratory costs are around US\$273 higher per ICU stay than the lowest spenders, even when corrected for different ICU lengths of stay and hospital mortality [2]. The highest spenders also spend more on other discretionary costs, which could be driven by increased laboratory use, including pharmacy costs and blood banking costs (e.g., red blood cell transfusion in a patient with anemia attributable to laboratory testing).

A further laboratory risk is false positives associated with over-testing [6]. The probability of false positives (lab results out of the normal range, when in fact the real values are normal) depends on the laboratory equipment, employees training and correct blood draws. However, the incidence of false positives increases with the number of tests run [7]. For example, if a given lab test randomly misclassifies people as *diseased* at a 1% rate (i.e. the test is 99% *accurate*), then the probability of having a false positive in the healthy population after an arbitrary 50 lab tests is

$$P_{wrong(n)} = 1 - (P_{right})^n \quad (1)$$

$$P_{wrong(50)} = 1 - (0.99)^{50} = 0.40$$

where  $n$  is the number of lab tests,  $P_{wrong(n)}$  is the probability of obtaining one wrong result (in this case, a false positive) in  $n$  lab tests and  $P_{right}$  is the probability of obtaining an

accurate result in a given lab test. One strategy to reduce this increasing probability of obtaining false positives is to avoid testing when no additional information is expected or, in other words, to reduce  $n$ .

In [3], it was shown that the introduction of new laboratory testing guidelines in an ICU setting resulted in a significant decrease in the number of tests performed (37%), and a significant increase in the number of tests obtained with a proper physician order (38%). These results were sustained over time, and were associated with no detectable morbidity and possibly to a decrease in red blood cell transfusions.

To prove that it is possible to model and classify the usefulness of a given lab test using information from other variables, this work focuses specifically in classifying the need for a hematocrit lab test. This measurement was chosen as a starting point because its dynamics are well understood, and specifically related to other measurable phenomena, like bleeding, transfusions, urine output and red blood cell count lab tests [8]. Also the impact of its prediction may be high, since it would be one step forward into reducing the amount of complete blood count (CBC) tests run and therefore, the amount of blood draws. We hypothesize that it is possible to reduce hematocrit laboratory testing by predicting when further testing is unlikely to provide new information that would change treatment strategy. This reduction would be achieved by assessing a range of physiological variables, and predicting when significant variations in hematocrit are likely to occur.

We apply a Knowledge Discovery in Databases (KDD) process to find the classification models on an ICU database named MIMIC II. Conventional soft computing techniques, namely fuzzy modeling and neural networks, combined with bottom-up feature selection were used to classify the usefulness of hematocrit lab tests. Other mathematical and statistical tools were also applied during data preprocessing.

This work is organized as follows. Section 2 formally describes the MIMIC II database and the methods used to preprocess, reduce and classify the outcome proposed. Section 3 describes and discusses the results obtained from these methods. Conclusions are presented in Section 4.

## II. METHODS

Knowledge Discovery in Databases (KDD) is an interactive and iterative process [9], involving numerous steps, which aims to discover hidden patterns and/or useful information in large datasets that do not express those patterns easily.

According to the steps described by Figure 1 for the KDD process, the following subsections define the approaches used in this work to tackle each step.

### A. Application domain and target dataset

The application domain identified for this work is in the field of critical care medicine, specifically, a medical ICU. Laboratory testing guidelines are generally created according to evidence based medicine informed by different population studies. However, lab tests are frequently run by default,

without considering an individual patient's condition. The main role of the KDD process in this work is to classify the relevance of hematocrit lab testing for ICU patients based on existing data specific to the patient, and whether a given test would then add information that could change management.

For this study, an ICU database named MIMIC II was used. MIMIC II was created as part of a Bioengineering Research Partnership (BRP) grant from the National Institute of Biomedical Imaging and Bioengineering entitled Integrating Data, Models and Reasoning in Intensive Care (RO1-EB001659). MIMIC II has been collected since 2001 at Beth Israel Deaconess Medical Center in Boston, MA, USA, including high frequency sampled data of bedside monitors, clinical data (laboratory tests, physicians' and nurses' notes, imaging reports, medications and other input/output events related to the patient) and demographic data [10]. All data is appropriately de-identified. As this is being written, MIMIC II continues to evolve with new versions being posted on the PhysioNet web site (<http://www.physionet.org/>). The version 2.5, used in this work, contains 26,665 ICU patients.

The target dataset used in this work was created using the following inclusion criteria:

- patients with Gastrointestinal (GI) bleeding were selected from MIMIC II, according to GI bleeding-related ICD9 (International Classification of Diseases) codes registered for them during their whole ICU stay [11];
- patients were required to be in the ICU for 7 days or more in order to have useful time series for analysis;
- data collection was considered to start when the first heart rate value was stored in the database, and ended with the last recorded heart rate, or a gap of 24 hours or more;
- only first ICU stays were considered, mainly to avoid patients that could have had several ICU admissions during the same hospitalization, reflecting a more complex state;
- only patients 18 years or older were included as MIMIC II also includes neonates and infants [10];
- bedside monitor trends, lab tests, ventilator variables, indexes, scores, arterial/central/catheter line monitoring were considered, while medications and demographic information were discarded (Table I).

### B. Data Preprocessing

Data preprocessing is a complex and crucial step as it defines the information that will be used later in the classification process, and it has to be carefully carried out in order to obtain clean data without altering the real information contained therein. Noise and outliers removal, misalignment correction and missing data treatment are considered as substeps of the preprocessing, accounting for up to 60% of the effort in a knowledge discovery process [12], [13].

The nature and occurrence of noise in MIMIC II is difficult to manage mainly due to the low sampling frequency of most variables, and to the broad band of frequencies that originally compose them. Consequently, it was not considered during the preprocessing.

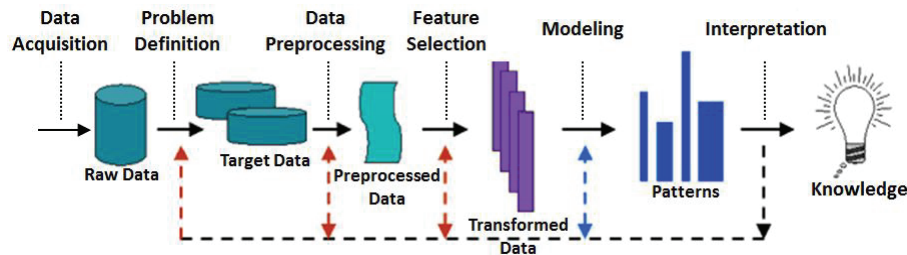


Fig. 1. Overview of the steps that compose the KDD process.

TABLE I  
VARIABLES USED FOR THE MODELING PURPOSES.

| Bedside monitor trends                          |           |                              |                    |                               |               |                                |                         |
|-------------------------------------------------|-----------|------------------------------|--------------------|-------------------------------|---------------|--------------------------------|-------------------------|
| Heart Rate                                      |           | Temperature                  |                    | Respiratory Rate              |               | Oxygen Saturation (SpO2)       |                         |
| Arterial/Central/Catheter Line                  |           |                              |                    |                               |               |                                |                         |
| Arterial Blood Pressure (ABP)                   | ABP Mean  | Arterial pH                  | Arterial CO2       | Arterial PaCO2                | Arterial PaO2 | Arterial Base Excess           | Central Venous Pressure |
| Laboratory results                              |           |                              |                    |                               |               |                                |                         |
| Hematocrit                                      | Potassium | Glucose                      | Creatinine         | Sodium                        | Chloride      | Lactic Acid                    | Uric Acid               |
| Carbon Dioxide                                  | Magnesium | Ionized Calcium              | Phosphorous        | Calcium                       | SaO2          | CaO2                           | Albumin                 |
| Ventilator monitored variables                  |           |                              |                    |                               |               |                                |                         |
| PEEP Set                                        |           | Respiratory Rate             |                    | Minute Volume (Observed)      |               | Low Exhaled Minute Volume      |                         |
| High Inspiratory Pressure                       |           | Plateau Pressure             |                    | Mean Airway Pressure          |               | Respiratory Rate Set           |                         |
| Resp Rate (Spontaneous)                         |           | Peak Inspiratory Pressure    |                    | Tidal Volume (Observed)       |               | Tidal Volume (Set)             |                         |
| Indexes calculated for combination of variables |           |                              |                    |                               |               |                                |                         |
| Cardiac Output                                  |           | Systemic Vascular Resistance |                    | Left Ventricular Cardiac Work |               | Right Ventricular Cardiac Work |                         |
| Scores                                          |           |                              |                    |                               |               |                                |                         |
| SAPS-I                                          | SAPS-II   | APACHE                       |                    |                               |               |                                |                         |
| Other variables collected                       |           |                              |                    |                               |               |                                |                         |
| Non-invasive Blood pressure (NBP)               |           | NBP Mean                     | Blood transfusions |                               |               |                                |                         |

An outlier is defined as an observation that is markedly different in value from the others of the sample or, in other words, a value out of the normal range of occurrence. The standard approach to identify and eliminate outliers is to filter the signal by amplitude, which means that specific upper and lower thresholds are defined for the signal. Those values above or below the threshold are considered outliers and therefore, removed from the signal. Another approach is to analyze every value in its context, comparing it with its neighbors to assess the likelihood of that point of being an outlier. Since it is sometimes possible for hematocrit to change quite drastically in a short period of time (e.g. hemorrhage) and then very rapidly rebound closer to their prior value (e.g. with massive transfusion), the second approach cannot be used in this work and, because of that, a filter with upper and lower amplitude thresholds was designed considering as possible extreme values those clinically accepted as possible for humans. This was reviewed and validated under medical assessment. Those values identified as outliers were assigned as missing data for further preprocessing.

According to [14], misalignments can be corrected using one variable in the dataset as a template, and shifting the data points of other variables to align sampling times. In this work, the outcome variable (*Hematocrit*) was used as the template to unshift the values of all the other variables, as it defines the

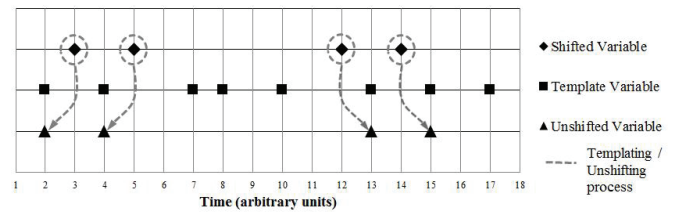


Fig. 2. Simplified representation of a *template variable*, together with the *shifted* and *unshifted* version of a second variable.

points for which the model is meant to give a classification result. Although [14] proposes to use the variable with the highest sampling rate (in this dataset it corresponds to *Heart Rate*, with a mean sampling time of  $0.76 \pm 0.2$  hours), in this work it would create an excessive amount of points for which the outcome is not expected to be measured (*Hematocrit* has a mean sampling time of  $9.15 \pm 1.24$  hours). As shown in Figure 2, all the existing entries are shifted to the closest template alignment location, and values are then obtained through an interpolation strategy using the template variable as a time reference. No new points were created in this step to fill otherwise blank values. The values for the new time positions were obtained through cubic interpolation, as suggested by [14].

To differentiate true missing values from lack of sampling [14], a 95% confidence interval was defined for the sampling time of every variable, and absent segments inside those intervals were recovered by cubic interpolation.

In order to assess whether true missing values were recoverable or not in each specific variable, a method based in fuzzy models, which classifies every missing point according to the ability to find information about the missingness contained in other variables [15], was used. For each point of a specific variable, a binary outcome is created containing a 1 if the variable is missing or a 0 if it has a value. Using the other variables of the dataset as inputs and the binary outcome as the output, if the fuzzy model is able to correctly classify a missing point as missing it is because it found information in the other variables that explain the missingness. In [15] it was pointed out that when there is an underlying medical reason for the missingness, the value should not be imputed; information in other variables may be considered as medical reasons. After classifying missing data, those missing segments identified as *recoverable* were recovered using expectation maximization [16], [17]; those classified as *not-recoverable* caused the deletion of all the variables for that sampling time.

### C. Data Reduction. Bottom-Up Feature Selection.

Data reduction involves finding those variables with useful information to model and predict the pursued outcome. A bottom-up feature selection process to reduce data dimensionality was used in this work.

A detailed description of the bottom-up approach used here may be encountered in [18]. However, it is important to note that a more recent algorithm that minimizes the computational time with similar performance has been proposed in [19]. The bottom-up approach looks for single inputs that may influence the output, and combines them in order to achieve the model with the best performance criterion. Two subsets of data are used in this stage, train and test. Using the train set, a model is built for each of the  $n$  features in consideration, and evaluated using the performance criterion upon the test set. The feature that returns the best value of the performance criterion is the one selected. Next, other feature candidates are added to the previous best model, one at a time, and evaluated. Again, the combination of features that maximizes the performance criterion is selected. When this second stage finishes, the model has two features. This procedure is repeated until the value of the performance criterion stops increasing. In the end, all the relevant features for the considered process should be obtained.

Three different performance criteria were separately tried in this work:

- the area under the ROC curve (AUC), as used by [20]
- the integrated discrimination improvement (IDI) [21], calculated in a simplified form as follows

$$IDI = (Sensitivity_{new} - Sensitivity_{old}) - (Specificity_{new} - Specificity_{old}) \quad (2)$$

where *Sensitivity* and *Specificity* refer to the true positive (hematocrit correctly classified as relevant) and true negative (hematocrit correctly classified as not relevant) classification rate of a model<sup>1</sup>, subscript *new* refers to the model with the new feature selected and subscript *old* to the model without it. The IDI can be viewed as a difference between improvement in average sensitivity and any potential increase in average specificity.

- the sensitivity, as part of a multi-objective function in which the specified criterion is considered together with specificity and percentage of correct classifications<sup>2</sup> (PCC), but with a higher degree of consideration through different membership functions of a fuzzy model [22].

### D. Classification strategy

The data mining strategy used in this work was defined for a binary classification of every hematocrit lab test as:

- Relevant, with a value of 1, when there is a drop  $d$  (varied between 1 and 5) in the value of hematocrit, or when those values are under a certain critical threshold  $T$  (varied between 15 and 40). Both  $d$  and  $T$  were varied to find the best classification point, and then reviewed with clinicians to check their practical significance and management influence on events such as triggering a decision of start or stop a blood transfusions or medications.
- Not relevant, with a value of 0, if the variations were below the previously defined thresholds.

The preprocessed data subset was randomly divided into two equal parts, one for the feature selection process (FS dataset) and the other for the model selection (MS dataset). This was done to assess select the relevant features and to assess the model's performances over independent datasets.

The subsets of features were defined over the FS dataset by randomly selecting the train (60%), test (30%) and validation (10%) sets from the FS dataset; the subset of features resulting in the highest PCC for the validation set was selected.

A leave-one-out cross-validation (LOOCV) process was used to select the best model, by using a subset 10% of the MS dataset, defined as the validation set, and the remaining 90% as the training set. To reduce variability, 10 rounds of cross-validation were performed using different partitions, and the validation results were averaged over the rounds. The model with the best performance criteria was selected.

### E. Modeling and Data Mining

The classification previously defined was performed using two different soft computing techniques and their results were compared. Namely, fuzzy models and neural networks were independently used, due to their universal approximators nature and previous results showing their performance in medical

<sup>1</sup>Sensitivity = TP / (TP + FN); true positives over true positives and false negatives, or total positives.

Specificity = TN / (TN + FP); true negatives over true negatives and false positives, or total negatives.

<sup>2</sup>The percentage of correct classifications is equal to the number of true positives (TP) together with the number of true negatives (TN) divided by the total number of patients, and then multiplied by 100.

databases [20]. For the specific case of fuzzy models, an additional motivation for their use was the easy-understandable rules they generate after classifying data, which could be useful for medical interpretation.

1) *Fuzzy Models*: Fuzzy modeling is a tool that allows an approximation of nonlinear systems when there is few or no previous knowledge of the system to be modeled. The fuzzy modeling approach has several advantages compared to other nonlinear modeling techniques ([23],[24]). In general, fuzzy models provide not only a more transparent model, but also a linguistic interpretation in the form of rules. A fuzzy classifier contains a rule base consisting of a set of fuzzy if-then rules together with a fuzzy inference mechanism.

In this work, Takagi-Sugeno (TS) fuzzy models [24] were used, which consist of fuzzy rules where each rule describes a local input-output relation. When TS fuzzy systems are used, each discriminant function consists of rules of the type

$$\text{Rule } R_i^c : \text{ If } x_1 \text{ is } A_{i1}^c \text{ and } \dots \text{ and } x_M \text{ is } A_{iM}^c \\ \text{ then } d_i^c(\mathbf{x}) = f_i^c(\mathbf{x}), i = 1, 2, \dots, K,$$

where  $f_i^c$  is the consequent function for rule  $R_i^c$ . In these rules, the index  $c$  indicates that the rule is associated with the output class  $c$ . Note that the antecedent parts of the rules can be for different discriminants, as well as the consequents. Therefore, the output of each discriminant function  $d_c(\mathbf{x})$  can be interpreted as a score (or evidence) for the associated class  $c$  given the input feature vector  $\mathbf{x}$ . The degree of activation of the  $i$ th rule for class  $c$  is given by:

$$\beta_i = \prod_{j=1}^M \mu_{A_{ij}^c}(\mathbf{x}) \quad (3)$$

where  $\mu_{A_{ij}^c}(\mathbf{x}) : \mathbb{R} \rightarrow [0, 1]$ . The discriminant output for each class  $c$ , with  $c = 1, \dots, C$ , is computed by aggregating the individual rules contribution:

$$d_c(\mathbf{x}) = \frac{\sum_{i=1}^K \beta_i f_i^c(\mathbf{x})}{\sum_{i=1}^K \beta_i} \quad (4)$$

The classifier assigns the class label corresponding to the maximum value of the discriminant functions, i.e

$$\max_c d_c(\mathbf{x}) \quad (5)$$

2) *Neural Networks*: Artificial Neural Networks (ANN) have been largely used as input-output mapping for different applications including modeling and classification. There are typically three parts in a neural network: an input layer with neurons representing the input variables, one or more hidden layers with a definable number of neurons, and an output layer with one or more units representing the output variables(s). The neuron output signal is given by the following relationship:

$$\sigma = f(\mathbf{w}^T x) = f\left(\sum_{j=1}^n (w_j x_j)\right) \quad (6)$$

where  $\mathbf{w} = (w_1, \dots, w_n)^T \in \mathbb{R}^n$  is the weight vector, and  $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$  is the vector of neuron inputs. The function  $f(\mathbf{w}^T x)$  is often referred to as the activation (or transfer) function. Its domain is the set of activation values,  $net$ , of the neuron model, and is often represented by  $f(net)$ . The variable  $net$  is defined as a scalar product of the weight and input vectors:

$$net = \sum_{j=1}^n (w_j x_j) = w_1 x_1 + \dots + w_n x_n \quad (7)$$

In this work, a feed forward multilayer perceptron (FF MLP) neural network was used, with 10 neurons and logarithmic activation functions in the hidden layer, and tangential activation function in the output layer, following the structure proposed by [20].

### III. RESULTS AND DISCUSSION

A subset of GI bleeding patients was selected from MIMIC II database according to ICD9 (International Classification of Diseases) codes. After the outliers were removed and misalignment correction and missing data recovery/deletion, the resulting dataset consisted of 235 patients with 50 variables present in all of them, summing up a total of 9382 records. The variables considered for posterior modeling within this dataset were bedside monitor trends, ventilator acquired/set values, indexes, scores, urine output and blood transfusions. All lab tests were included, excluding red and white blood cell counts, hemoglobin and platelets, since they are considered to be tested together with hematocrit in a complete blood count (CBC) [25]. Hematocrit was excluded from the inputs, since it was used as the outcome to classify. After the preprocessing, a complete dataset was obtained for modeling purposes.

The outcome was created as a binary variable, taking the value 1 when the hematocrit was below a threshold  $T$  or when a drop of  $d$  points occurred on it.  $T$  was varied between 15 and 40, and  $d$  was varied between 1 and 5. Tables II and III show the values of AUC obtained when using neural networks and fuzzy modeling in the classification the resulting outcome. The maximum AUC obtained when classifying the outcomes generated with fuzzy models and neural networks agreed, and corresponded to  $T = 35$  and  $d = 3$ . These values were validated by clinical co-authors as significant for influencing clinical management. In other words, both a drop of 3 points in the hematocrit value or a sustained hematocrit value below 35 would be triggers for urgent diagnostic activities exploring causes of blood loss (e.g. endoscopy) and urgent therapeutic activities such as insertion of a central line, cross matching blood and increased laboratory testing. The resulting outcome can be seen in Figure 3 for an arbitrarily chosen patient.

After defining the outcome, the complete dataset was randomly divided into two equal parts, one for the feature selection process (FS dataset) and the other for the model selection (MS dataset). This was done to assess model performances and to select the best model over an independent dataset not used during any step of model construction.

TABLE II

AUC FOR FUZZY AND NEURAL MODELS WHEN CLASSIFYING DIFFERENT HEMATOCRIT LOWER THRESHOLDS  $T$ , VARIED FROM 15 TO 40. FM - FUZZY MODELING; NN - NEURAL NETWORKS; HCT - HEMATOCRIT

| HCT threshold $T$ | AUC             |                 |
|-------------------|-----------------|-----------------|
|                   | NN              | FM              |
| 15                | $0.52 \pm 0.09$ | $0.53 \pm 0.03$ |
| 20                | $0.54 \pm 0.04$ | $0.60 \pm 0.04$ |
| 25                | $0.61 \pm 0.05$ | $0.67 \pm 0.02$ |
| 30                | $0.65 \pm 0.02$ | $0.72 \pm 0.04$ |
| 35                | $0.82 \pm 0.03$ | $0.84 \pm 0.02$ |
| 40                | $0.60 \pm 0.06$ | $0.61 \pm 0.04$ |

TABLE III

AUC FOR FUZZY AND NEURAL MODELS WHEN CLASSIFYING DIFFERENT HEMATOCRIT DROPS  $d$ , VARIED FROM 1 TO 5. FM - FUZZY MODELING; NN - NEURAL NETWORKS; HCT - HEMATOCRIT

| HCT Drop $d$ | AUC             |                 |
|--------------|-----------------|-----------------|
|              | NN              | FM              |
| 1            | $0.55 \pm 0.02$ | $0.55 \pm 0.01$ |
| 2            | $0.64 \pm 0.01$ | $0.66 \pm 0.03$ |
| 3            | $0.71 \pm 0.03$ | $0.75 \pm 0.04$ |
| 4            | $0.65 \pm 0.02$ | $0.63 \pm 0.01$ |
| 5            | $0.60 \pm 0.01$ | $0.59 \pm 0.02$ |

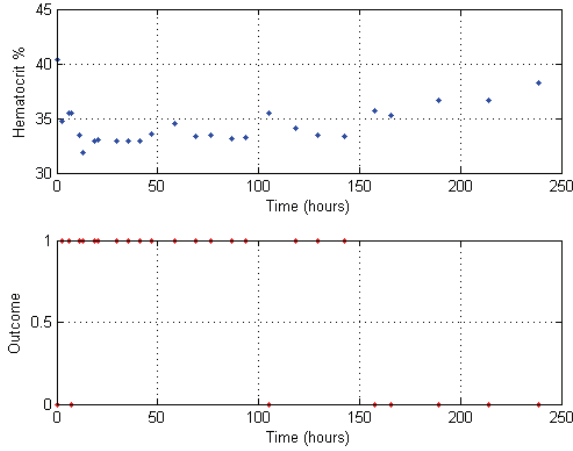


Fig. 3. Outcome generation from the original hematocrit time series for an arbitrarily chosen patient.

The subsets of features were defined over the FS dataset by maximizing AUC, IDI or sensitivity (considered in a multiobjective function). This process was performed 10 times for each maximization criterion, randomly selecting the train (60%), test (30%) and validation (10%) sets from the FS dataset, using neural networks and fuzzy models to classify the outcome; the subset of features resulting in the highest percentage of correct classifications (PCC) for the validation set was selected.

Variables selected with AUC, IDI and sensitivity maximization criteria and their respective PCC values, obtained with neural networks and fuzzy models are shown in Table IV. For both neural networks and fuzzy models, it is possible to see that when AUC was used as the maximization criterion, variables selected result in the highest PCC values. Although

TABLE IV

PCC AND VARIABLES SELECTED WHEN USING AUC, IDI AND SENSITIVITY AS THE MAXIMIZATION CRITERIA DURING A BOTTOM-UP FEATURE SELECTION, PERFORMED WITH FUZZY AND NEURAL MODELS. FM - FUZZY MODELING; NN - NEURAL NETWORKS; IDI - INTEGRATED DISCRIMINATION IMPROVEMENT

| Model | Maximization | PCC  | Variables Selected                                                    |
|-------|--------------|------|-----------------------------------------------------------------------|
| NN    | AUC          | 81.3 | Urine Output<br>Creatinine<br>NBP<br>Sodium                           |
|       | IDI          | 60.1 | Urine Output<br>Creatinine<br>Chloride                                |
|       | Sensitivity  | 77.5 | Urine Output<br>NBP                                                   |
| FM    | AUC          | 85.2 | Urine Output<br>SpO2<br>Creatinine                                    |
|       | IDI          | 59.7 | Urine Output<br>Chloride<br>Blood Transfusion<br>SpO2<br>Arterial CO2 |
|       | Sensitivity  | 78.8 | Urine Output<br>NBP<br>Blood Transfusion<br>Heart Rate                |

there was always a coincidence in the first variable selected by all models, the other features seem to respond to minor variations of the subsets used to create the models, since they varied a lot even for different runs of the same modeling technique.

Although it was expected that the use of sensitivity, as the maximization criterion, could increase the final sensitivity and maybe the AUC too, the results obtained showed no improvements (Table V). Moreover, the computational time (not shown in the results) increased for this approach. In the case of IDI, it was expected for it to result in values similar to AUC, since it integrates both sensitivity and specificity in one measurement in a similar way than AUC. However, its use resulted in the lowest PCC values.

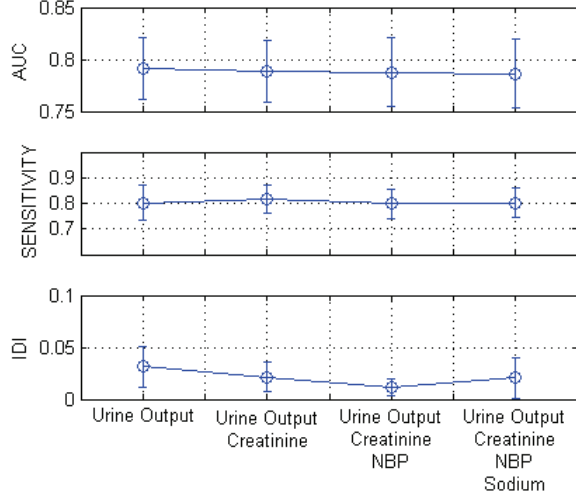
The three resulting features subsets, one for each maximization criterion, were separately used in a leave-one-out cross-validation (LOOCV) process. A subset 10% of the MS dataset was defined as the validation set, and the remaining 90% as the training set.

Figure 4a and 4b show an accumulative representation of the results for each additional variable used as input, when AUC was previously used as the maximization criterion to obtain the subset of variables, averaged over the ten runs of cross validation. It can be seen that no improvement is obtain by the use of any variable after *Urine Output*. This, together with the variability reflected in the variables selected under the same maximization criterion (not shown in the results), suggests that only *Urine Output* has predictive value for the outcome defined, while the other variables are selected responding to the variability of the subsets used to create the models. Some other frequently selected variables such as *Non-invasive Blood Pressure* (NBP) and *Creatinine*, could relate to

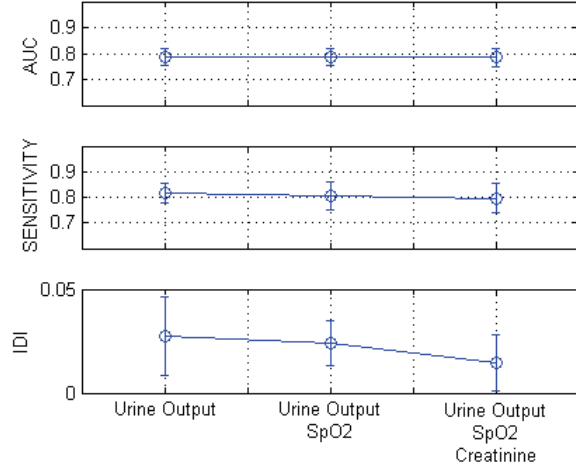
TABLE V

AUC, SENSITIVITY, SPECIFICITY AND PCC OF FUZZY AND NEURAL MODELS OBTAINED DURING THE LEAVE-ONE-OUT CROSS VALIDATION PROCESS, WHEN USING THE FEATURES SELECTED BY AUC, IDI AND SENSITIVITY MAXIMIZATION CRITERIA. FM - FUZZY MODELING; NN – NEURAL NETWORKS

| Criterion   | NN              |                 |                 |            | FM              |                 |                 |            |
|-------------|-----------------|-----------------|-----------------|------------|-----------------|-----------------|-----------------|------------|
|             | AUC             | Sensitivity     | Specificity     | PCC        | AUC             | Sensitivity     | Specificity     | PCC        |
| AUC         | $0.78 \pm 0.03$ | $0.77 \pm 0.03$ | $0.75 \pm 0.01$ | $79 \pm 3$ | $0.79 \pm 0.05$ | $0.79 \pm 0.04$ | $0.80 \pm 0.03$ | $81 \pm 3$ |
| IDI         | $0.59 \pm 0.07$ | $0.62 \pm 0.05$ | $0.55 \pm 0.04$ | $59 \pm 6$ | $0.60 \pm 0.03$ | $0.63 \pm 0.01$ | $0.58 \pm 0.07$ | $58 \pm 5$ |
| Sensitivity | $0.72 \pm 0.04$ | $0.71 \pm 0.01$ | $0.70 \pm 0.02$ | $67 \pm 2$ | $0.73 \pm 0.04$ | $0.70 \pm 0.03$ | $0.74 \pm 0.02$ | $73 \pm 2$ |



(a) Neural networks



(b) Fuzzy models

Fig. 4. Accumulative AUC, Sensitivity and IDI when each variable selected, using AUC as the maximization criterion, is added during the LOOCV.

specific conditions of some patients that are not shared by the whole dataset.

Table V summarizes the results obtained for fuzzy and neural networks modeling, when the three maximization criteria were separately used to select the features, averaged during the LOOCV process. It can be seen that fuzzy models perform slightly better than neural networks, expressed in higher values for almost all performance measurements.

The fuzzy models gave best results when 2 clusters were

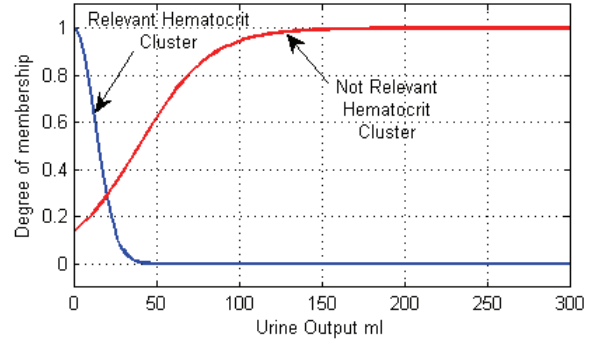


Fig. 5. Membership function for Urine Output obtained with a fuzzy model, maximizing the AUC, during the validation.

used during their training, testing and validation. Neural networks presented their best performance for one hidden layer composed by 10 neurons with logarithmic activation functions, and a linear summation in the output layer.

The selection of *Urine Output* as the primary variable in all models was reviewed and validated under medical assessment. Due to the selection of GI bleeding patients during the creation of the dataset used in this work, the presence of hemorrhage was somehow proved for all patients during their ICU stay. Blood volume can be significantly affected by hemorrhage, and so the urine output. If the hematocrit is low enough or presents sudden drops, poor urine collection will reflect those states with some accuracy. This can be seen in one membership function obtained from the fuzzy models, illustrated in Figure 5. This is a positive finding since, by using a simple variable that is collected by the nurse in an hourly assessment of the patient, phlebotomy associated risks and complete blood count (CBC) testing can potentially be avoided.

#### IV. CONCLUSIONS

Increasing health trends towards improved efficiency seek to align evidence based medicine with cost efficient medicine, both underpinned by a drive to quality. Unnecessary phlebotomy increases clinical and fiscal risks at many levels. In this work, hematocrit lab testing usefulness was evaluated by modeling and classifying the relevance of the results when other collected variables are assessed. A properly de-identified ICU database collected at Beth Israel Deaconess Medical Center, named MIMIC II, was used to develop the models. A subset of gastrointestinal bleeding patients were selected along

with their associated physiological ICU variables, including bedside monitor trends, ventilator records, scores, indexes, lab tests, urine output and blood transfusion information.

Informed by clinical decision making, hematocrit was defined as the positive binary outcome variable set to threshold levels of hematocrit less than 35% or a drop of 3 points. All other measures were represented as the negative binary outcome.

Three maximization criteria, namely the area under the ROC curve (AUC), integrated discrimination improvement (IDI) and sensitivity, were used in a bottom-up feature selection process, combined with neural networks and fuzzy models.

A leave-one-out cross validation process was used to evaluate the overall models' performance, as well as the additional predictive value of the variables selected. Urine output was selected by all models as the most predictive variable for hematocrit levels, and consequently for potential diagnostic and therapeutic interventions. This is consistent with known physiological processes whereby deteriorating renal function occurs as a result of falling blood volume due to factors such as hemorrhage. Results showed that it was possible to correctly classify the usefulness of a hematocrit lab test up to 81% of the time by using fuzzy models and neural networks.

This work is offered as a proof of concept for the use of artificial intelligence models in predicting the usefulness of diagnostic testing. Although the dynamics of hematocrit values are used here on the basis that they are supported by a deeper body of knowledge, the results obtained showed that other routinely variables collected could also be used for predicting diagnostic testing requirements in critical care patients. Future work will focus on:

- expanding the models to a broader population of patients, under different medical conditions
- predicting/classifying different and more complex lab tests currently run on ICU patients
- creating dynamical models to perform future predictions of lab tests usefulness based on present and previous data acquired.

#### ACKNOWLEDGMENT

The author would like to acknowledge the assistance of the Division of Clinical Informatics at the Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology. Both human and technical resources were made available, both of which were critical for the development of this work. This work is supported by the Portuguese Government under the programs: project PTDC/SEM-ENR/100063/2008, Fundação para a Ciência e Tecnologia (FCT), and by the MIT-Portugal Program and FCT grants SFRH/BPD/65215/2009, SFRH/43043/2008 and SFRH/43081/2008, Fundação para a Ciência e a Tecnologia, Ministério do Ensino Superior, da Ciência e da Tecnologia, Portugal.

#### REFERENCES

- [1] D. E. Roberts, D. D. Bell, T. Ostryzniuk, K. Dobson, L. Oppenheimer, D. Martens, N. Honcharik, H. Cramp, E. Loewen, and S. Bodnar, "Eliminating needless testing in intensive care: An information-based management approach," *Crit Care Med*, vol. 10, pp. 1452–1458, 1993.
- [2] M. E. Ezzie, S. K. Abernethy, and J. M. O'Brien, "Laboratory testing in the intensive care unit," *Crit Care Clin*, no. 23, pp. 435–465, 2007.
- [3] K. Kumwilaisak, A. Noto, U. H. Schmidt, C. I. Beck, C. Crimi, K. Lewandowski, and L. M. Bigatello, "Effect of laboratory testing guidelines on the utilization of tests and order entries in a surgical intensive care unit," *Crit Care Med*, vol. 36, no. 11, 2008.
- [4] S. M. Mehari and J. H. Havill, "Written guidelines for laboratory testing in intensive care," *Critical Care Resuscitation*, vol. 3, pp. 158–162, 2001.
- [5] A. Garland, Z. Shaman, J. Baron, and J. A. F. Connors, "Physician-attributable differences in intensive care unit costs," *American Journal of Resp and Crit Care Medicine*, vol. 174, pp. 1206–1210, 2006.
- [6] R. J. Woolley, "Drug testing of physicians: the danger of false positives," *JAMA*, vol. 264, no. 24, p. 3148, 1990.
- [7] D. W. Bates, L. Goldman, and T. H. Lee, "Contaminant blood cultures and resource utilization: The true consequences of false-positive results," *JAMA*, vol. 265, pp. 365–369, 1991.
- [8] W. L. Lee and G. P. Downey, "Icu cornerstone: Changing our view of blood transfusions," *Crit. Care*, vol. 6, no. 4, pp. 291–292, 2002.
- [9] U. Fayyad, G. P. Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17(3), pp. 37–54, 1996.
- [10] G. D. Clifford, D. J. Scott, and M. Villarroel, "User guide and documentation for the mimic ii database," 2009, version 2.1.
- [11] N. C. for Health Statistics (NCHS), the Centers for Medicare, and M. S. (CMS), *The Int Class Diseases, 9th Rev, Clinical Modification*. NCHS, 2008.
- [12] K. Cios, L. Kurgan, K. J. Cios, and L. A. Kurgan, "Trends in data mining and knowledge discovery," in *In: Pal N.R., Jain, L.C. and Teoderesku, N. (Eds.), Knowledge Discovery in Advanced Information Systems*. Springer, 2005, pp. 200–202.
- [13] D. Pyle, *Data Preparation for Data Mining*. San Francisco, CA: Morgan Kaufmann, 1999.
- [14] F. Cisondi, A. S. Fialho, S. M. Vieira, J. M. C. Sousa, S. R. Reti, M. D. Howell, R. Welch, and S. N. Finkelstein, "New methods to processing misaligned, unevenly sampled time series containing missing data," *IEEE Symposium Series on Computational Intelligence - SSCI 2011*, 2011.
- [15] A. S. Fialho, F. Cisondi, S. M. Vieira, J. M. C. Sousa, S. R. Reti, M. D. Howell, R. Welch, and S. N. Finkelstein, "Missing data in large intensive care units databases," in *Soc of Crit Care Med (SCCM) 40th Crit Care Con*, vol. 38, San Diego, California, USA, 2011, pp. A1–A285.
- [16] N. M. L. A. P. Dempster and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the em algorithm," *J. Roy. Stat. Soc.*, vol. 39, pp. 1–38, 1977.
- [17] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data. Series in Probability and Mathematical Statistics*. Wiley, 1987.
- [18] M. Sugeno and T. Yasukawa, "A fuzzy-logic-based approach to qualitative modeling," *IEEE Trans. on Fuzzy Systems*, vol. 1, pp. 7–31, 1993.
- [19] L. F. Mendonça, S. M. Vieira, and J. M. C. Sousa, "Decision tree search methods in fuzzy modeling and classification," *International Journal of Approximate Reasoning*, vol. 44, pp. 106–123, 2007.
- [20] A. S. Fialho, F. Cisondi, S. M. Vieira, J. M. C. Sousa, S. R. Reti, M. D. Howell, and S. N. Finkelstein, "Predicting outcomes of septic shock patients using feature selection based on soft computing techniques," in *Applications 13th International Conference, IPMU 2010, Proceedings, Part II*, ser. Communications in Computer and Information Science (CCIS), R. K. E. Huellermeier and F. Hoffmann, Eds., vol. 81. Springer-Verlag Berlin Heidelberg, 2010, pp. 65–74.
- [21] M. Pencina, R. D. Sr., R. D. Jr., and R. Vasan, "Evaluating the added predictive ability of a new marker: From AUC to reclassification and beyond," *Statistics in Medicine*, vol. 27, pp. 157–172, 2008.
- [22] A. L. Horn, F. Cisondi, A. S. Fialho, S. M. Vieira, J. M. C. Sousa, S. R. Reti, M. D. Howell, and S. N. Finkelstein, "Multi-objective performance evaluation using fuzzy criteria: Increasing sensitivity prediction for outcome of septic shock patients," in *18th World Congress of the International Federation of Automatic Control (IFAC)*, Italy, 2011.
- [23] J. M. C. Sousa and U. Kaymak, *Fuzzy Decision Making in Modeling and Control*. Singapore: World Scientific Pub. Co., 2002.
- [24] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modelling and control," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 15, pp. 116–132, 1985.
- [25] W. K. Purves, D. Sadava, G. H. Orians, and H. C. Heller, *Life: The Science of Biology*, 7th ed. Sunderland, Mass, USA: Sinauer Associates, 2004.