Patient Classification and Outcome Prediction in IgA Nephropathy

M. Diciolla [(a)], G. Binetti [(a)], T. Di Noia [(a)], F. Pesce [(b,c)], F. P. Schena [(c,d)], A.-M. Vågane [(e)], R. Bjørneklett [(e)], H. Suzuki [(f)], Y. Tomino [(f)], D. Naso [(a)]

[a] *Department of Electrical and Information Engineering, Polytechnic University of Bari, Bari, Italy.*

[b] *Cardiovascular Genetics and Genomics, National Heart & Lung Institute, Royal Brompton Hospital, Imperial College London, UK.*

[c] *Department of Emergency and Organ Transplantation, University of Bari, Bari, Italy.*

[d] *C.A.R.S.O. Consortium, Valenzano-Casamassima, Italy.*

[e] *Department of Clinical Medicine, Renal Research Group, University of Bergen, Bergen, Norway; Department of Medicine, Haukeland University Hospital, Bergen, Norway*

[f] *Division of Nephrology, Department of Internal Medicine, Juntendo University, Faculty of Medicine, Tokyo, Japan*

ARTICLE INFO

Keywords:

IgA Nephropathy

Artificial neural network

Neuro fuzzy system

Support vector machine

Decision tree

Decision support system

ABSTRACT

Objective: IgA Nephropathy (IgAN) is a common kidney disease which may entail renal failure, known as *End Stage Kidney Disease* (ESKD). One of the major difficulties dealing with this disease is to predict the time of the long-term prognosis for a patient at the time of diagnosis. In fact, the progression of IgAN to ESKD depends on an intricate interrelationship between clinical and laboratory findings. Therefore, the objective of this work has been the selection of the best data mining tool to build a model able to predict (I) if a patient with a biopsy proven IgAN will reach ESKD and (II) if a patient will reach the ESKD before or after 5 years.

Material and Methods: The largest available cohort study worldwide on IgAN has been used to design and compare several data-driven models. The complete dataset was composed of 1174 records collected from Italian, Norwegian, and Japanese IgAN patients, in the last 30 years. The data mining tools considered in this work were artificial neural networks (ANNs), neuro fuzzy systems (NFSs), support vector machines (SVMs), and decision trees

(DTs). A 10-fold cross validation was used to evaluate unbiased performances for all the models.

Results: An extensive model comparison based on accuracy, precision, recall, and f-measure was provided. Overall, the results indicate that ANNs can provide superior performance compared to the other models. The ANN for time-to-ESKD prediction is characterized by accuracy, precision, recall, and f-measure greater than 90%. The ANN for ESKD prediction has accuracy greater than 90% as well as precision, recall, and f-measure for the class of patients not reaching ESKD, while precision, recall, and f-measure for the class of patients reaching ESKD are slightly lower. The obtained model has been implemented in a Web-based decision support system (DSS).

Conclusions: The extraction of novel knowledge from clinical data and the definition of predictive models to support diagnosis, prognosis, and therapy is becoming an essential tool for researchers and clinical practitioners in medicine. The proposed comparative study of several data mining models for the outcome prediction in IgAN patients, using a large dataset of clinical records from three different countries, provides an insight into the relative prediction ability of the considered methods applied to such a disease.

1. Introduction

In the last years, the extraction of novel knowledge from clinical data has attracted the attention of many researchers [1], [2]. Several data mining techniques [3] have been investigated in the literature with the aim to support diagnosis, prognosis and therapy. Many works have been successfully applied to deal with different clinical conditions, such as breast cancer [4], cardiovascular diseases [5], cytomegalovirus disease after renal transplantation [6], celiac disease [7], epilepsy [8], liver transplantation [9], learning disabilities [10], Parkinson [11] and multiple sclerosis diseases [12]. Indeed, the prediction of a disease outcome and/or the patient risk stratification is one of the most interesting tasks in which to develop data mining applications.

This paper deals with *IgA Nephropathy* (IgAN) that is a common form of glomerulonephritis characterized by episodes of macroscopic hematuria (so called “coca-cola” urine), concomitant with upper respiratory tract infections,

or asymptomatic microscopic hematuria [13]. *End Stage Kidney Disease* (ESKD) occurs when the kidney function is very seriously damaged and requires renal replacement therapy, with dialysis or kidney transplantation. One of the major difficulties dealing with IgAN is to predict the long-term prognosis for a patient since the progression to ESKD depends on intricate interrelationships between demographic and clinical data, and medical interventions. Indeed, the ability to accurately predict the risk level is still limited, and the need for novel prediction tools for ESKD risk degree has been widely emphasized [14] to improve the care of patients with IgAN.

Several studies in the medical literature have been dedicated to the long-term outcome prediction of patients with biopsy proven IgAN [15], [16] [17], [18] [19], [20], [21]. These studies are mainly based on statistical techniques, such as univariate and multivariate Cox regression models. Hence, several absolute renal risk scores have been proposed by evaluating different set of risk factors present at the time of diagnosis. A first attempt of data mining application to IgAN has been proposed by Geddes et al. [22] who used an artificial neural network to classify a patient as "stable" or "non-stable" based on the serum creatinine value after 7 years from the time of renal biopsy. However, this is a surrogate endpoint which only partially represents the main outcome defined by the ESKD and, in addition, a limited set of 54 patients was included in the study. Later, Cannone et al. [23] investigated the interpretability of some classification tools (i.e., support vector machine, decision tree, artificial neural network, naïve Bayes classifier) to predict the time to ESKD, which was divided into "small" (less than 8 years) and "long" (more than 8 years) groups, on a dataset of only 98 entries. Finally, our previous work [24] proposed an ensemble of artificial neural networks to predict the risk of ESKD, using a large dataset of 587 records. The numerical value was associated to the prediction for the purpose of having a human-readable representation of the risk degree, but without relying on any information in the dataset about the timing of the IgAN progression.

Variable	Range	Mean	Median	Standard Deviation
Age (years)	16-76	34.65	32	13.43
Serum Creatinine (mg/dl)	0.45-5.83	1.35	1.1	0.97
Proteinuria (g/24h)	0-17	1.56	0.8	2.04

Table 1. Continuous independent variables.

Variable	Possible values	Numeric value	Number of records	Percentage (%)
Gender	M	0	719	69.13
	F	1	321	30.87
Histological grade	G1	1	363	34.90
	G2	2	508	48.85
	G3	3	169	16.25
Hypertension	No	0	605	58.17
	Yes	1	435	41.83

Table 2. Categorical independent variables.

In this paper, the design and performance analysis of several data mining tools for a quantitative risk assessment of ESKD and its timing using available clinical data at the time of renal biopsy have been proposed. The main contributions of the proposed work are listed in the following.

- The largest available cohort study worldwide on IgAN has been exploited to design and compare several data-driven models. With respect to the data of Italian IgAN patients [24], the dataset in this work has been enriched with new clinical data from Norwegian and

	Possible values	Numeric values	Number of records	Percentage (%)
ESKD	No	0	799	76.83
	Yes	1	241	23.17
Time-to-ESKD	> 5 years	0	114	47.30
	≤ 5 years	1	127	52.70

Table 3. Categorical dependent *ESKD* and *Time-to-ESKD* variables.

Japanese patients. Moreover, the dataset has been also enriched with explicit temporal information about the IgAN progression to ESKD.

- Exploiting the new dataset, two different classification systems have been designed for each data mining technique considered in this work. The first classifier predicts the renal failure occurrence, i.e., if a patient will reach or not the ESKD. The second one has been designed to define if a patient will reach ESKD within 5 years or not, by exploiting the new temporal information available in the dataset.
- The performance analysis of different data-mining techniques has been proposed. In details, artificial neural networks (ANNs), neuro-fuzzy systems (NFSs), support vector machines (SVMs) and decision trees (DTs) were compared.
- Finally, an implementation of the best performing tool has been presented as a Web-based decision support system (DSS).

In Section 2 of our paper, the complete description of the available dataset and the classification system design has been illustrated. In Section 3, the discussion on the performance of each technique and an exhaustive comparison has been provided. Section 4 includes a brief description of the implemented Web-based decision support system. Final remarks close our paper.

2. Material and Methods

2.1. Dataset

The dataset includes clinical records of patients with biopsy-proven IgAN from three different countries: Italy, Norway and Japan. Such a heterogeneous data collection is composed of 1174 records distributed as follows:

- 677 records collected from 1972 to 2010 by the Renal, Dialysis and Transplant Unit at University of Bari, Bari, Italy;
- 441 records collected from 1988 to 2013 by the Renal Research Group of the Department of Clinical Medicine at University of Bergen, Bergen, Norway;
- 56 records collected from 1974 to 2013 by the Division of Nephrology of the Department of Internal Medicine at Juntendo University, Tokyo, Japan.

After an initial data cleaning to remove records with missing information, the dataset was composed of 1040 records of which 546 records from Italian, 441 records from Norwegian, and 53 from Japanese patients. Each record is composed of 8 attributes, where the first 6 are independent variables corresponding to the patient information collected at the time of the renal biopsy, while the remaining two attributes are the dependent variables of clinical interest.

Predictor	Odds Ratio (95% confidence interval)	P value
Gender		
Female	Referent	
Male	1.50 (0.96-2.35)	0.075
Age	0.98 (0.96-0.99)	0.004
Histological grade		
Mild	Referent	
Moderate	2.44 (1.42-4.40)	0.002
Severe	3.65 (1.89-7.21)	<0.001
Serum Creatinine	1.02 (1.02-1.02)	<0.001
Proteinuria	1.01 (1.01-1.01)	<0.001
Hypertension		
No	Referent	
Yes	1.57 (1.04-2.36)	0.030

Table 4. Multivariate logistic regression for ESKD status.

Predictor	Odds Ratio (95% confidence interval)	P value
Gender		
Female	Referent	
Male	1.56 (0.72-3.47)	0.264
Age	0.98 (0.96-1.01)	0.173
Histological grade		
Mild	Referent	
Moderate	4.28 (0.74-81.90)	0.182
Severe	10.64 (1.73-208.21)	0.033
Serum Creatinine	1.03 (1.02-1.05)	<0.001
Proteinuria	1.02 (1.01-1.03)	0.002
Hypertension		
No	Referent	
Yes	0.55 (0.25-1.15)	0.119

Table 5. Multivariate logistic regression for Time-to-ESKD.

The independent variables represent the inputs of the classification model:

1. *Gender* of the patient.
2. *Age* of the patient.
3. *Histological grade*: histological classification of the renal biopsy which indicates three levels of severity (G1: mild, G2: moderate, G3: severe), as reported in [25].
4. *Serum Creatinine* (mg/dl): biological marker whose normal values are 0.7-1.3 mg/dl for men and 0.6-1.1 mg/dl for women; higher quantity of this molecule in the blood may indicate that the kidney does not work properly.
5. *Proteinuria* (g/24h): biological marker, related to the presence of an excess of proteins in the urine, whose normal value is less than 0.3g/24h.
6. *Hypertension*: hypertensive state characterized by blood pressure greater than 140/90 mmHg or patient treated with anti-hypertensive drugs; this condition represents a risk factor for renal failure.

On the other hand, the dependent variables represent the outputs of the classification model:

1. ESKD: final stage of IgAN with complete renal failure, meaning that the patient requires dialysis or kidney transplantation.
2. Time-to-ESKD: time between biopsy-proven diagnosis and ESKD stage; each patient is defined as *high* or *low* risk with the meaning that he/she will reach ESKD before or after 5 years.

Tables 1, 2, and 3 synthetize the main characteristics of both the independent and dependent variables. In details, Table 1 shows range, mean, median, and standard deviation for the continuous independent variables. Table 2 shows the possible values for each categorical independent variable as well as the number of records and the relative percentage. For the design of the classification models, each categorical variable is represented by an integer value (e.g., the histological grades G1, G2, and G3 are represented by the numbers 1, 2, and 3, respectively). Table 3 shows statistical information about the dependent

variables. The ESKD variable is distributed in 799 and 241 records of value No and Yes, respectively, i.e., a larger number of patients who did not reach the final stage of the IgAN. About the time-to-ESKD variable, it indicates patients who reached the final renal failure. In details, among the patients reaching ESKD, the dataset is composed of about the same number of patients with ESKD before and after 5 years.

In order to identify significant predictors of ESKD and time-to-ESKD to be used as inputs for the data mining methods, we first assessed their importance by logistic regression. Multivariate analysis, illustrated in Tables 4 and 5, shows that all the 6 independent variables have a significant impact on ESKD and time-to-ESKD in our data sets. Moreover, the largest effect was observed in patients with a severe histological grade both for ESKD and time-to-ESKD predictions. Thus, we used these independent variables resulted significant at the multivariate analysis as inputs for the data mining techniques in the following sections.

2.2. Models

In this section, artificial neural networks (ANNs), neuro-fuzzy systems (NFSs), support vector machines (SVMs) and decision trees (DTs) have been first introduced. Then, these techniques have been used to design two different models for ESKD and time-to-ESKD predictions. Each model takes in inputs the six independent variables listed in Tables 1 and 2, after a zero-mean unit-variance normalization, to predict one of the independent variable in Table 3.

In this study, a *10-fold cross validation* was performed, that is, 10 folds were randomly created with roughly the same class proportion as in the initial dataset. Moreover, since the dataset collects information of three different populations, each of them was uniformly spread over all the folds. The *K-folds cross validation* was used to limit the bias of a model performance associated with a random sampling of training data and, as a consequence, to guarantee a good generalization property [30]. This approach guarantees higher reliability since the impact of specific dataset on the system performance is minimized.

2.2.1. Artificial Neural Networks

The artificial neural networks [26], [27], are commonly used in medical applications, e.g., [9], [22]. An ANN is a universal function approximator which can solve non-linearly separable problems and learn any arbitrarily complex linear function with arbitrary accuracy level [28], [29]. As in the biological neural systems, an ANN consists of several highly interconnected processing elements (the neurons) whose outputs depend on the linear combination of weighted inputs. The Multi-Layer Perceptron (MLP) is a feedforward network, successfully used also for pattern recognition [30], which is composed by an input layer, one or more hidden layers, and an output layer, as shown in Figure 1. Formally, assuming p inputs, n hidden layers and 1 neuron in the output layer, the output of the j-th hidden neuron, is computed as follows:

$$h_j = f^h\left(\sum_{i=1}^{p} w_{ij}^h x_i + b_j^h\right), \quad (1)$$

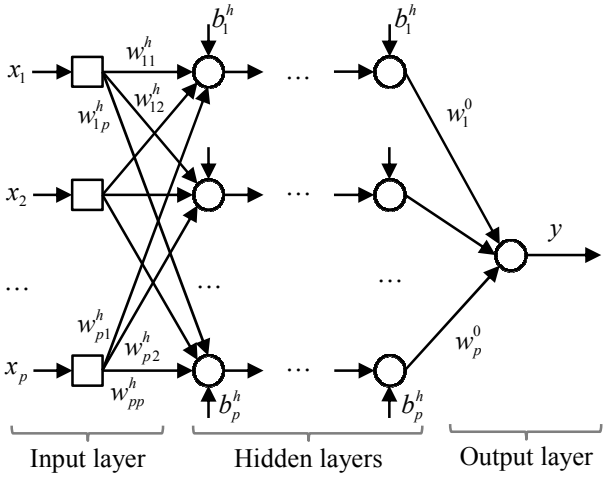

Figure 1. Multi-Layer Perceptron architecture for an Artificial Neural Network.

where f^h is the activation function, w_{ij}^h is the weight between input x_i and hidden neuron j, and b_j^h is a bias term. Similarly, the output layer computes

$$y = f^o\left(\sum_{j=1}^{p} w_j^o x_i + b_j^o\right), \tag{2}$$

where f^o is the activation function of the output neuron.

A classic MLP architecture, with sigmoidal and linear activation functions for hidden and output neurons, respectively, was used to build both ESKD and time-to-ESKD classification models. The dataset was divided into training, validation, and test sets. The validation set was used during the learning process to avoid overfitting by adopting the early stop strategy [30], while the test set is used to properly evaluate the classifier performance. The training process was performed using the well-known *Levenberg-Marquardt algorithm* to minimize the mean square error [31]. The best ANN was chosen among many trials varying the number of hidden layers from 1 to 3, the number of hidden neurons from 3 to 30, and re-training each network 1000 times with different random initial weights. The criterion used to choose the best ANN is detailed in Section 2.3.1.

2.2.2. Neuro Fuzzy Systems

The fuzzy set and logic theory is a prominent tool to handle uncertainty in decision-making [12], [32]. This approach allows to express and process relationships by a set of *if-then rules*. The Tagaki-Sugeno-Kang (TSK) model is a class of models where the conclusion of the rule *i* for the output *j* is computed as a linear function of the inputs (antecedent variables), as

$$y_j^i = b_{j0}^i + b_{j1}^i x_1 + \ldots + b_{jp}^i x_p, \tag{3}$$

with *p*-dimensional input $\mathbf{x}$ and *q*-dimensional output $\mathbf{y}$. The adaptive neuro fuzzy inference system (ANFIS) is an adaptive network that optimizes the TSK fuzzy inference system by exploiting the ANNs learning capability [36]. Figure 2 shows an example of a zero-order TSK model with two rules represented by a neuro-fuzzy network. Each layer of the neuro fuzzy network represents a part of the inference system: the first layer nodes compute the input *fuzzification* (membership degree evaluation in premise fuzzy sets); the second layer nodes compute the *rule inference* and computation (by the product node Π); finally, the normalization nodes *N* and the summation node ∑ perform the output *defuzzification* [36]. Clustering methods are known to be fast and robust algorithms to identify fuzzy models from data [39].

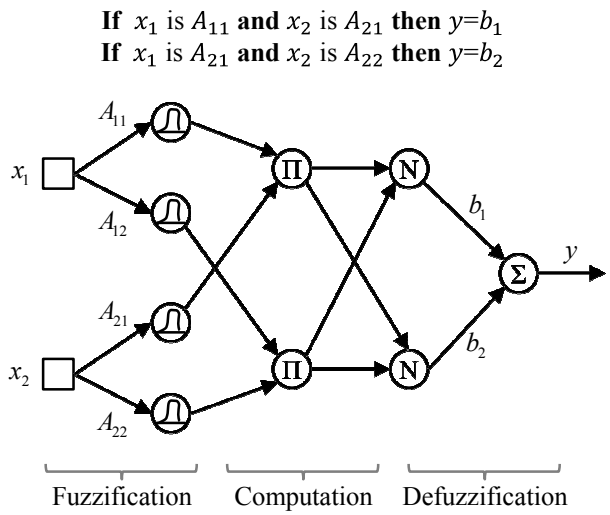

Figure 2. Tagaki-Sugeno-Kang model with two rules represented as neuro-fuzzy network.

An adaptive neuro fuzzy inference system, based on the zero-order TSK model (with Gaussian membership functions), was used to define both the ESKD and time-to-ESKD classification models. The rule induction definition was performed trying the *subtractive* [38] and *Fuzzy C-Means* (FCM) [39] clustering algorithms. The former calculates a measure of the likelihood that each data point would define a cluster center, considering the density of neighboring data points. The latter associates to each data point a certain degree of membership for each cluster, and performs an iterative procedure to move the initial guess of each cluster center to the best location within the dataset. In particular, the consequent and premise parameters of each rule are updated by the ANFIS's hybrid learning algorithm [36], [37], which is a combination of least square and back-propagation techniques. This method compares the generated predictions with the target values to update the membership function shapes, while using a validation set to avoid overfitting. The cluster center range of influence in each data dimension varies from 0.1 to 0.9 for the subtractive clustering, while the number of clusters varies from 2 to 30 with 100 induction trials for each initial guess of the FCM clustering. The best classification model among all the trials was evaluated on the test set by maximizing the criterion described in section 2.3.1.

2.2.3. Support Vector Machines

The support vector machines are powerful algorithms for data classification and regression [33] that have been also used for prognosis and diagnosis problems [4], [10]. An interesting feature of SVM is that it is quite insensitive to the relative number of records in each class, since they do not try to minimize the error rate [34]. Formally, denoting the feature vector by $\mathbf{x_i} \in \mathcal{R}^p$ and the class label by $y_i \in \{-1,1\}$, if the two classes are linearly separable, the SVM approach sets the convex optimization problem

$$\begin{aligned} &\min_{w,b} \quad \frac{1}{2}\|w\|_2 \\ &\text{s.t.} \quad y_i\left[\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right] \geq 1, \quad i = 1,2,...,N \end{aligned}, \tag{4}$$

to find the parameters $\mathbf{w}$ and b that maximizes the separation margin $2/\|\mathbf{w}\|_2$ defined by the hyperplane $y = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b = 0$ discriminating the two classes, as shown in Figure 2. If the problem is not linearly separable, using a Lagrangian formulation and a Kernel function $K(\mathbf{x},\mathbf{x}_i)$, the SVM transforms the data points into another high dimensional linearly separable feature space, where the classification function is given by

$$y(\mathbf{x}) = \operatorname{sign}\left[\sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x},\mathbf{x}_i) + b\right]. \tag{5}$$

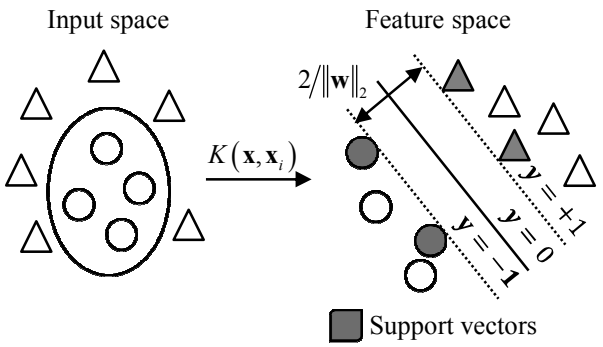

Figure 3. Example of mapping data points into a linearly separable space for a Support Vector Machine.

To build the ESKD and time-to-ESKD classification models, the SVM was applied using different kernel functions to find the data transformation into the kernel space that maximizes the performance criterion described in section 2.3.1. In details, the kernels used in the different trails are linear, quadratic, polynomial (with orders varying from 3 to 9), multi-layer perceptron and Gaussian radial basis function (with scaling factor varying from 1 to 9). The method to find the separating hyperplane during the training procedure is the *Sequential Minimal Optimization* (SMO) [35]. Finally, the best SVM model was chosen by evaluating the performance of different SVMs only on the test set, according to the criterion described in Section 2.3.1.

2.2.4. Decision Trees

The decision trees are data mining tools characterized by a great human interpretability [40], [41]. Indeed, a decision tree is a flowchart-like structure in which each internal node represents a test on an attribute, each branch represents the outcome of the test and each leaf node represents a class label (i.e., the decision taken after computing all attributes). The paths from the root to each leaf represent all the classification rules. An example of decision tree is illustrated in Figure 4. A decision tree is built by recursively calling a tree construction algorithm in each generated node. The build procedure is heuristically driven by choosing the "most informative" attribute at each step. Then, the training instances are splitted into subsets according to their value of the current test attribute. Finally, the tree generation stops if all tuples in a subset belong to the same class or if it is not worth to proceed with additional separation into further subsets.

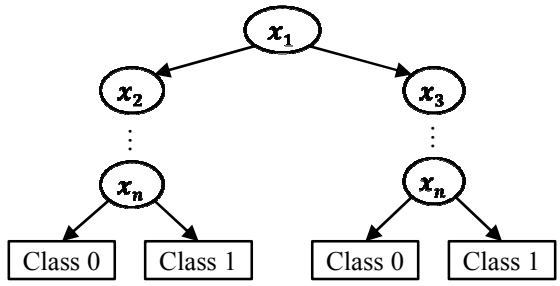

Figure 4. Example of binary decision tree.

The so-called classification and regression tree (CART) algorithm [42] was used to build both the ESKD and time-to-ESKD classification models. A top-down recursive algorithm was used for binary trees generation that performs a greedy search through the space of data: at each stage, it uses an evaluation function to select the attribute that has the best ability to discriminate among the classes. In this work, the Gini index was used as splitting criterion during the learning procedure [42]. Moreover, to avoid the overfitting problem, the post-pruning based on the impurity heuristic [42] was used during the tree construction process. To determine the best tree that maximizes the criterion presented in Section 2.3.1, many trials were performed varying the minimum number of clinical records

per leaf between 10 and 100. Finally, since the training data were imbalanced for the ESKD classification, a penalty cost was incorporated into the decision tree algorithm, so that when the model misclassified an instance belonging to the class "ESKD=1", the cost was doubled with respect to the other class.

2.3. Classification Performance

The quality of a data-driven model for a classification problem can be defined according to several performance measures based on the *confusion matrix*. Given m classes, with $m \geq 2$, the confusion matrix is an $m \times m$ table that compares the classifier outputs with the actual values in the dataset. Thus, the classification result for a specific record can be described as a *true positive* (TP), *false positive* (FP), *true negative* (TN), or *false negative* (FN).

The *accuracy* (a) is a global criterion defined as the percentage of correct prediction in the dataset:

$$a = \frac{TP + TN}{(TP + TN + FP + FN)}. \quad (6)$$

For a specific class, it is possible to consider different performance criteria. The *precision* (p), also called *positive predictive value*, is a measure of exactness which is defined as the percentage of correct prediction of class i with respect to the total number of class i prediction:

$$p = \frac{TP}{(TP + FP)}. \quad (7)$$

The *recall* (r), also called *sensitivity* or *true positive rate*, is a measure of completeness which is defined as the percentage of correct prediction of class i with respect to the total number of class i records in the dataset:

$$r = \frac{TP}{(TP + FN)}. \quad (8)$$

The *f-measure* (f) is defined as the harmonic mean of precision and recall, i.e.,

$$f = \frac{2 \cdot p \cdot r}{p + r}, \quad (9)$$

and it is a useful criterion to evaluate a trade-off between the two metrics.

In our application, there are two classes and it is possible to define precision, recall, and f-measure for each of them, i.e., p_0, r_0, and f_0 for class 0 ("ESKD=0" or "time-to-ESKD=0"), and p_1, r_1, and f_1 for class 1 ("ESKD=1" or "time-to-ESKD =1"), respectively.

2.3.1. Performance Criterion

In medical problems, the accuracy is not necessarily the best criterion to evaluate the performance of a classifier [43]. For instance, the dataset used in this work for the ESKD prediction was composed by 799 records of class "ESKD=0" and 241 records of class "ESKD=1", i.e., a number of critical IgAN patients that is three time less than the number of non-critical ones. A high value of accuracy may mean that the classifier can correctly classify many non-critical patients to the detriment of the critical ones. Thus, for a medical application, a performance criterion is often defined in terms of sensitivity and specificity [5], [34], to better represent the performance of the classification tool on each class.

In this work, the FP and FN errors were considered to choose the best classifier among all the trials (obtained varying peculiar parameters of each data mining technique)

and to compare the different classification systems. For the ESKD prediction, a FP error means that a patient will need dialysis while he/she actually will not need it, and a FN error means that a patient will not need dialysis while he/she actually need it. Thus, a FP error entails wrong treatments, while a FN error implies that a patient will remain untreated. For the time-to-ESKD prediction, a FP error means that the model predicts that the patient will reach ESKD within 5 years while he/she actually will reach it later, and a FN error means that a patient will reach ESKD after 5 years while such a condition will occur earlier. In this case, a FP error involves the unnecessary application of aggressive treatment, while a FN error implies that a patient is not properly cured with stronger treatment in order to delay the end stage as more as possible. Thus, in order to consider the peculiar meaning of each of these errors, the average f-measure was maximized in this work to balance precision and recall of the different classes.

3. Results

3.1. ESKD Prediction

The best models for each data mining technique to predict the ESKD condition have been chosen among all the trials as follows:

- ANN with MLP architecture with 2 hidden layers and 9 hidden neurons per layer;
- NFS with TSK architecture, FCM clustering and 17 rules;

	Artificial Neural Network						
	a	p_0	p_1	r_0	r_1	f_0	f_1
mean	**90.1**	**92.1**	**83.4**	**94.9**	**73.5**	**93.4**	**77.9**
st. dev.	5.4	4.1	9.5	3.8	10.6	3.8	8.9

	Neuro Fuzzy System						
	a	p_0	p_1	r_0	r_1	f_0	f_1
mean	**87.2**	**92.1**	**73.7**	**92.1**	**72.8**	**91.7**	**72.9**
st. dev.	6.4	4.1	11.3	4.7	13.1	4.4	11.4

	Support Vector Machine						
	a	p_0	p_1	r_0	r_1	f_0	f_1
mean	**82.3**	**91.2**	**65.0**	**84.2**	**74.7**	**87.3**	**67.5**
st. dev.	10.3	5.8	17.7	12.6	11.6	8.9	11.1

	Decision Tree						
	a	p_0	p_1	r_0	r_1	f_0	f_1
mean	**85.6**	**91.6**	**72.6**	**89.1**	**73.8**	**82.2**	**71.9**
st. dev.	8.8	4.8	19.9	10.3	11.3	6.8	12.7

Table 6. Results for ESKD prediction.

- SVM with RBF kernel and scaling factor equal to 4.
- DT with 13 rules.

Table 6 shows all the performance criteria of such models for ESKD prediction, i.e., if a patient reaches the final stage of the disease or not. Overall, the ANN provides superior performance compared to the other techniques. The accuracy for the ANN is equal to 90.1%, while NFS, SVM, and DT have accuracy equal to 87.2%, 82.3%, and 85.6%, respectively. The ANN also provides a small standard deviation, equal to 5.4%, while the other techniques exhibit higher standard deviations (6.4%, 10.3%, and 8.8% for NFS, SVM, and DT, respectively). The precision, recall, and f-measure for the two classes show that all the models are able to classify better the patients belonging to the class 0. This behavior seems to be

	Artificial Neural Network						
	a	p_0	p_1	r_0	r_1	f_0	f_1
mean	**91.0**	**92.9**	**90.1**	**90.4**	**91.8**	**91.3**	**90.6**
st. dev.	5.1	6.8	7.9	7.9	7.9	4.9	5.3

	Neuro Fuzzy System						
	a	p_0	p_1	r_0	r_1	f_0	f_1
mean	**83.0**	**83.1**	**84.5**	**85.7**	**80.2**	**84.0**	**81.8**
st. dev.	4.3	5.5	9.4	9.0	7.6	4.4	4.4

	Support Vector Machine						
	a	p_0	p_1	r_0	r_1	f_0	f_1
mean	**79.8**	**81.3**	**80.1**	**81.5**	**77.7**	**80.7**	**78.1**
st. dev.	6.3	9.8	8.7	9.2	13.3	5.7	7.6

	Decision Tree						
	a	p_0	p_1	r_0	r_1	f_0	f_1
mean	**80.5**	**80.9**	**81.5**	**83.1**	**77.6**	**81.7**	**79.0**
st. dev.	5.0	6.2	8.8	9.4	9.4	5.6	6.0

Table 7. Results for time-to-ESKD prediction.

mainly determined by the different cardinality of the two classes, that is, 241 patients reached the ESKD condition and 799 patients who did not reach the final stage. Indeed, the precision p_0 is almost the same for all the models, while the precision p_1 ranges from 65% for the SVM to the 83.4% for the ANN. Thus, the SVM is the model whose precision is mostly affected by the limited dataset of patients for class 1. The largest standard deviation of 19.9% for the precision is shown by the DT, while the SVM is characterized by a slight lower value of 17.7%. Again, the ANN has lowest standard deviation, i.e., 9.5%. The recall r_0 ranges from 84.2% for the SVM to 94.9% for the ANN, while the recall r_1 is in a smaller range between 72.8% and 74.7% for all the models. Moreover, the standard deviation of the recall r_1 is almost the same for all the models, about 10-13%, while the standard deviation of the recall r_0 has minimum value of 3.8% for the ANN and maximum value of 12.6% for the SVM. Finally, the f-measure provides certain information which weights precision and recall. The ANN provides the highest f-measure on both the two classes, with smallest standard deviations, compared to the other models. The NFS shows slight inferior f-measure compared to the ANN, while the SVM and the DT are characterized by the lowest average f-measure, around 77%.

3.2. Time-to-ESKD Prediction

The best models for each data mining technique to predict the time-to-ESKD have been chosen among all the trials as follows:

- ANN with MLP architecture with 2 hidden layers and 11 hidden neurons per layer;
- NFS with TSK architecture, FCM clustering and 14 rules;
- SVM with RBF kernel and scaling factor equal to 4.
- DT with 7 rules.

Table 7 shows all the performance criteria of such models. As in the previous case, also for the time-to-ESKD prediction the ANN provides superior performance compared to the other techniques. Indeed, all the performance criteria for the ANN are a few percentage points greater than the performance of other models. In details, all criteria for ANN are higher than 90% and are characterized by a small standard deviation, about 5-8% on

average. One can also note that precision, recall, and f-measure performance on the two classes are similar. Unlike the previous case, the datasets for the two classes for the time-to-ESKD prediction contain a similar number of patients of which 114 and 127 patients reached ESKD after or before 5 years, respectively. Precision, recall, and f-measure for SVM and DT are very similar: precision for the two classes is about 80-81%, recall r_0 is about 81-83% and recall r_1 is 77%, while f-measure for the two classes is about 78-81%. The standard deviation for the SVM and DT models is also comparable, ranging from a minimum of 5% for accuracy of DT to a maximum of 13.3% for recall r_1 of SVM. Such values are smaller than those in the previous case of ESKD prediction. Finally, the NFS exhibits performance that are intermediate between the ANN and other models, with precision of 83-84%, recall of 82-83% on average, and f-measure of 82-83% on average. Also in this case, the NFS are the models which have closer performance to the ANN, since they have some similarities.

3.3. Discussion

The results illustrated in Tables 6 and 7 demonstrate that the ANN model can provide superior performance compared to other data mining methods for the medical application considered in this work. All the considered performance criteria (accuracy, precision, recall, and f-measure) are greater than 90% for the time-to-ESKD prediction as well as the accuracy for the ESKD prediction. Instead, precision, recall, and f-measure for the ESKD prediction are greater than 90% only for class 0 (patients not reaching the final stage of the disease), while being equal to 83.4%, 73.5%, and 77.9% for class 1 (patients reaching the final stage of the disease), respectively. This behavior is shown by all models for the ESKD prediction suggesting that it is related to the peculiarities of the specific clinical dataset considered in this work. Moreover, the ANN shows superior performance also compared to the previous work of the authors in [24], where an ensemble of ANNs was used to predict only the ESKD exploiting a smaller dataset, without any temporal information. Thus, the new available information about the Norwegian and Japanese patients in this work (that is, 441 and 53 new patients in the dataset, respectively) allows us to improve the model performance.

4. System Implementation

The ANN showing the best classification values has been implemented in an online decision support system. The core of the system is a Java RESTful [44] service which interacts with Matlab code. We initially implemented the ANN in Java in a native way but we observed unexpected results due to the different floating point precision between Matlab and Java. In the current implementation of the system, a PHP client communicates with the Java service and formats the information shown to the user.

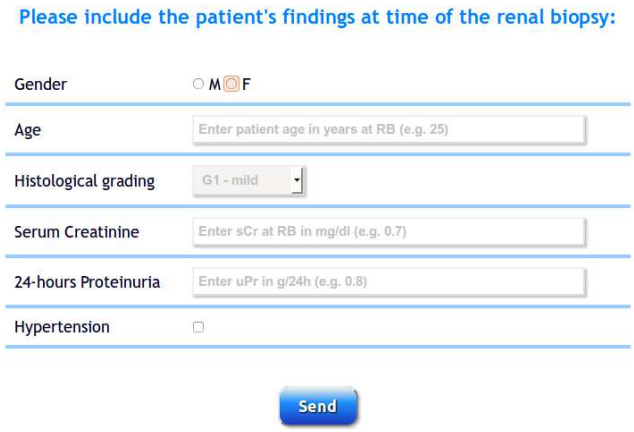

Figure 5. A snapshot of the initial Web page of the implemented DSS.

A snapshot of the initial Web page is shown in Figure 5. After the user opens it, it is possible to insert data corresponding to the features described in Section 2.1. Based on the inserted data, the system returns if there is no risk for ESKD (see Figure 6(a)) or if the patient is expected to reach ESKD after 5 years (see Figure 6(b)) or within 5 years (see Figure 6(c)). Moreover, the Web-based decision support system also shows a value associated to the outcome prediction that can be useful to alert clinicians about some border line patients. A patient is classified as class 1 if the ANN output is greater than 0.5. Thus, the farther is the ANN output from 0.5, the higher should be the value representing the risk prediction of the patient being of class 1. In order to have a numerical representation of this concept, we associated the value 60% to the case of ANN output equal to 0.5 and the value 95% to the case of ANN output equal to 1. Then, the remaining values are computed by using a linear function. Note that the values have no specific meaning and they have been chosen arbitrarily only for the purpose of having a human-readable representation of the computed risk prediction.

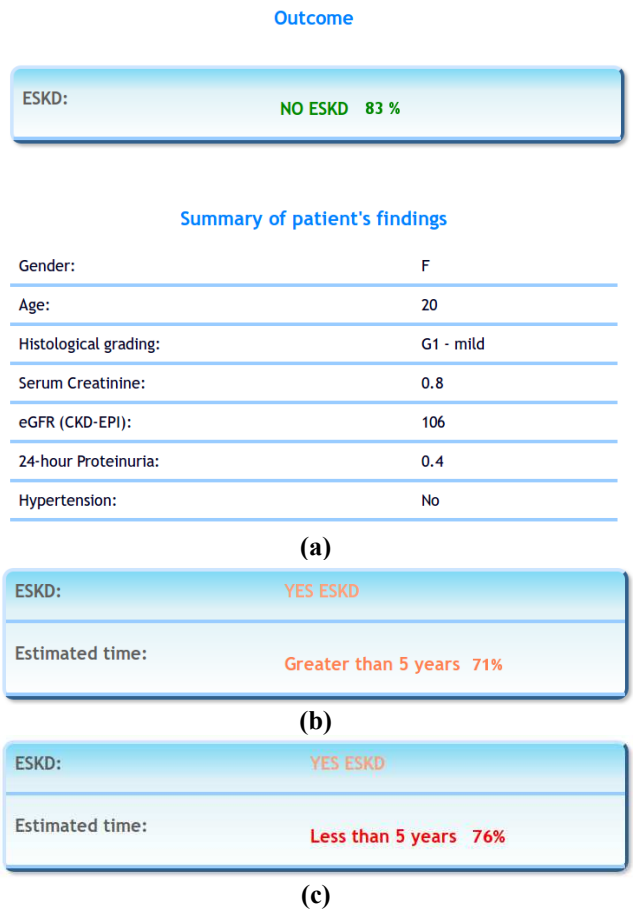

Figure 6. Outcome of the ANN-based DSS.

5. Conclusions

In this paper, several data mining techniques have been investigated with the aim to support diagnosis, prognosis, and therapy for patients affected by a common kidney disease, the IgA Nephropathy. Four different data-driven models (artificial neural networks, neuro fuzzy systems, support vector machines, and decision trees) have been trained using the largest available cohort study worldwide on IgAN. Two binary classifiers have been designed to predict if a patient will reach the renal failure, *End Stage Kidney Disease* (ESKD), and the time between the diagnosis and the ESKD occurrence. Each patient is described by six independent variables which represents a

set of clinical information collected at the time of renal biopsy: gender, age, histological grade, serum creatinine, proteinuria, and hypertension. The ESKD prediction model has an output binary variable meaning that a patient will reach or not the final renal failure. Moreover, the time-to-ESKD prediction model has a binary output variable meaning that a patient reaching the ESKD will be subject of such a critical condition after or before 5 years.

A stratified 10-folds cross validation has been used during the training of each model to evaluate unbiased performance measures. Among all the models generated varying several parameters (e.g., the number of hidden neurons for the ANNs or the kernel functions for the SVMs), the selection of the best model in a given family has been done considering the average f-measure on the different classes with the aim to obtain a classifier with a good compromise between precision and recall for each class. A comparative analysis of the data mining techniques has been presented in this work. Overall, the ANN model can provide superior performance compared to all the other models.

In conclusion, the ANN for time-to-ESKD prediction is characterized by accuracy, precision, recall, and f-measure greater than 90%.

The ANN for ESKD prediction is characterized by accuracy greater than 90% as well as precision, recall, and f-measure for class of patients not reaching the ESKD, while precision, recall, and f-measure for class of patients reaching the ESKD are slight lower. This difference is due to the peculiarities of the dataset which includes a lower number of clinical records.

References

[1] R. Bellazzi and B. Zupan, Predictive data mining in clinical medicine: current issues and guidelines, Journal of Medical Informatics 77 (2008) 81-97.

[2] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artificial Intelligence in Medicine 23 (2001) 89-109.

[3] J. Han, M. Kamber, and J. Pei, Data Mining, Concepts and techniques (M. Kaufmann, 2012).

[4] I. Maglogiannis, E. Zafiropoulos, and I. Anagnostopoulos, An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers, Applied Intelligence 30 (2009) 24-36.

[5] E. Avci, A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier, Expert Systems with Applications 36 (2009) 10618–10626.

[6] D. Sheppard, D. McPhee, C. Darke, B. Shrethra, R. Moore, A. Jurewitz, A. Gray, Predicting cytomegalovirus disease after renal transplantation: an artificial neural network approach, Journal of Medical Informatics 54 (1999) 55-76.

[7] Artificial intelligence techniques applied to the development of a decision-support system for diagnosing celiac disease, Journal of Medical Informatics 80 (2011) 793-802.

[8] Y. Kassahun, R. Perrone, E. De Momi, E. Berghöfer, L. Tassi, M.P. Canevini, R. Spreafico, G. Ferrigno, and F. Kirchner, Automatic classification of epilepsy types using ontology-based and genetics-based machine learning, Artificial Intelligence in Medicine 61 (2014) 79-88.

[9] M. Cruz-Ramírez, C. Hervás-Martínez, J.C. Fernández, J. Briceno and M. de la Mata, Predicting patient survival after liver transplantation using evolutionary multi-objective artificial neural networks, Artificial Intelligence in Medicine 58 (2013) 37-49.

[10] T.K. Wu, S.C. Huang, and Y.R. Meng, Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities, Expert Systems with Applications 34 (2008) 1846-1856.

[11] I. Mandal and N. Sairam, Accurate telemonitoring on Parkinson's disease diagnosis using robust inference system, International Journal of Medical Informatics 82 (2013) 359-377.

[12] M. Esposito, I. De Falco, and G. De Pietro, An evolutionary-fuzzy DSS for assessing health status in multiple sclerosis disease, International Journal of Medical Informatics 80 (2011) e245-e254.

[13] J. Berger and N. Hinglais, Intercapillary deposits of IgA-IgG, Journal of Urology Nephrology 74 (1968), 694-695.

[14] S. Barbour and H. Reich, Risk stratification of patients with IgA nephropathy, American Journal of Kidney Diseases 59 (2012) 865-873.

[15] F. Berthoux, H. Mohey, B. Laurent, C. Mariat, A. Afiani, and L. Thibaudin, Predicting the risk for dialysis or death in IgA Nephropathy, Journal of the American Society of Nephrology 22 (2011) 752-761.

[16] E. Alamartine, J.C. Sabatier, C. Guerin, J.M. Berliet, and F. Berthoux, Prognotic factors in mesangial IgA glomerulonephritis: an extensive study with univariate and multivariate analyses, American Journal of Kidney Diseases 18 (1991) 12-19.

[17] M.G. Radford, J.V. Donadio, E.J. Bergstralh, and J.P. Grande, Predicting renal outcome in IgA nephropathy, Journal of the American Society of Nephrology 8 (1997) 199-207.

[18] L.P. Bartosik, G. Lajoie, L. Sugar, and D.C. Cattran, Predicting progression in IgA Nephropathy, American Journal of Kidney Diseases 38 (2001) 728-735.

[19] B. MacKinnon, E.P. Fraser, D. Cattran, J.G. Fox, and C.C. Geddes, Validation of the Toronto formula to predict progression in IgA nephropathy, Nephron Clinical Practice 109 (2008) c148-c153.

Summary points

What was already known before this study:

- The difficulties of predicting the progression of IgA Nephropathy to End Stage Kidney Disease are already known, since such a condition depends on intricate interrelationships between clinical and laboratory findings.
- Several studies mainly based on statistical techniques have described the outcome prediction of patients with biopsy proven IgA Nephropathy.

What this study has added to our knowledge:

- An overview of four data mining techniques to fit the structure of the medical knowledge and the peculiarities of the medical decision-making has been carried out together with experimental results.
- Artificial neural networks can provide superior performance (accuracy, precision, recall, and f-measure) compared to neuro fuzzy systems, support vector machine, and decision trees. This suggests that ANNs can be a decision support system for patient classification and renal outcome prediction in IgA Nephropathy.
- A Web-based decision support system has been developed to support clinical decisions to forecast the progression of IgA Nephropathy to End Stage Kidney Disease.

[20] V. Rauta, P. Finne, J. Fagerudd, K. Rosenlof, T. Tornroth, and C. Gronhagen-Riska, Factors associated with progression of IgA nephropathy are related to renal function: a model estimating risk of progression in mild disease, Clinical Nephrology 58 (2002) 85-94.

[21] M. Goto, K. Wakai, T. Kawamura, M. Ando, M. Endoh, and Y. Tomino, A scoring system to predict renal outcome in IgA nephropathy: a nationwide 10-year perspective cohort study, Nephrology Dialysis Transplantation 24 (2009) 3068-3074.

[22] C. Geddes, J. Fox, M. Allison, J. Boulton-Jones, K. Simpson, An artificial neural network can select patients at high risk of developing progressive IgA nephropathy more accurately than experienced nephrologists. Nephrology Dialysis Transplantation 1 (1998) 67-71.

[23] R. Cannone, C. Castiello, A. M. Fanelli, and C. Mencar, Assessment of semantic cointension of fuzzy rule-based classifiers in a medical context, in: Proc. of 11th International Conference on Intelligent Systems Design and Applications (Cordoba, Spain, 2011) 1353-1358.

[24] T. Di Noia, V.C. Ostuni, F. Pesce, G. Binetti, D. Naso, F.P. Schena, and E. Di Sciascio, An end stage kidney disease predictor based on an artificial neural network ensemble, Expert Systems with Applications 40 (2013) 4438-4445.

[25] C. Manno, G. Strippoli, C. D'Altri, D. Torres, M. Rossini and F.P. Schena, A novel simpler histological classification for renal survival in IgA nephropathy: a retrospective study, American Journal of Kidney Disease 49 (2007) 763-775.

[26] C.M. Bishop, Neural Networks for Pattern Recognition (Clarendon Press, Oxford, 1995).

[27] G.P. Zhang, Neural Networks for Classification: A Survey, IEEE Transactions on Systems, Man, and Cybernetics 30 (2000) 451-460.

[28] K. Hornik, M. Stinchcombe, and H. White, Universal approximation of an unknown mapping and its derivates using multilayer feedforward network, Neural Networks 3 (1990) 551-560.

[29] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Network 4 (1991) 251-257.

[30] C.M. Bishop, Pattern Recognition and Machine Learning (Springer, 2006).

[31] J.J. Moré, The Levenberg-Marquardt algorithm: implementation and theory, Lecture Notes in Mathematics, volume 630 (1978) 105-116.

[32] B. Rezaee and M.H. Fazel Zarandi, Data-driven fuzzy modeling for Tagaki-Sugeno-Kang fuzzy system, Information Sciences 180 (2010) 241-255.

[33] C. Cortes and V. Vapnik, Support-Vector Networks, Machine Leaming, 20 (1995) 273-297.

[34] H. Drucker, D. Wu, and V.N. Vapnik, Support Vector Machines for Spam Categorization, IEEE Transactions on Neural Networks 10 (1999) 1048-1054.

[35] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, Advances in kernel methods (MIT Press Cambridge USA, 1999) 185-208.

[36] J.-S.R. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Transactions on Systems, Man, and Cybernetics 23 (1993) 665-685.

[37] S. Guillaume, Designing fuzzy inference systems from data: an interpretability-oriented review, IEEE Transactions on Fuzzy Systems 9 (2001) 426-443.

[38] A.H. Tawafan, M. Bin Sulaiman, and Z. Bin Ibrahim, Adaptive neural subtractive fuzzy inference system for the detection of high impedance fault on distribution power system, International Journal of Artificial Intelligence 1 (2012) 63-72.

[39] S.L. Chiu, Fuzzy model identification based on cluster estimation, Journal of Intelligent and Fuzzy Systems 2 (1994) 267-278.

[40] S.R. Safavian and D. Landgrebe, A survey of decision tree classifier methodology, IEEE Transactions on Systems, Man Cybernetics, 21 (1991) 660-674.

[41] C. Cardie, Using decision trees to improve case-based learning, in: Proc. of 10th International Conference on Machine Learning (1993) 25-32.

[42] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and regression trees (Belmont, 1984).

[43] N. Lavrac, Selected techniques for data mining in medicine, Artificial Intelligence in Medicine 16 (1999) 3-23.

[44] R.T. Fielding, Architectural styles and the design of network-based software architectures, Ph.D. thesis (2000).