Detecting a proper patient with a help of medical data retrieval

Teresa Małecka-Massalska[1], Ryszard Maciejewski[2],
Piotr Wąsiewicz[3], Wojciech Załuska[4], Andrzej Książek[4]
[1]Physiology Department, Medical University of Lublin
[2] Human Anatomy Department, Medical University of Lublin; Institute of Biomedical Informatics, Univeristy of Information Technology and Management in Rzeszow
[3] Institute of Electronic Systems, Warsaw University of Technology
[4]Nephrology Department, Medical University of Lublin

ABSTRACT

Electric bioimpedance is one of methods to assess the hydrate status in hemodialyzed patients. It is also being used for assessing the hydration level among peritoneal dialysed patients, diagnosed with neoplastic diseases, patients after organ transplantations and the ones infected with HIV virus. During measurements sets were obtained from two groups, which were named a control (healthy volunteers) and test group (hemodialyzed patients). Its variables were the following: body mass index (BMI), intracellular water (ICW) - water volume inside you cells. (i.e., water in the “living” cells), extracellular water (ECW) - water volume outside the body cell mass (i.e., water in the “inactive” cells), total body water (TBW) - sum of ICW and ECW, ECW_TBW - ECW divided by TBW, ECW_mass - ECW divided by body mass, height, weight and age. Zscored, discretized data and data retrieval results were computed in R language environment in order to find a simple rule for recognizing health problems. The executed experiments affirm possibilities of creating good classifiers for detecting a proper patient with the help of medical data sets, but only with previous training.

Key words: bioimpedance technique, data retrieval, decision tree, hydrate status, extracellular compartment, intracellular compartment, quadratic discriminant analysis

INTRODUCTION

Monitoring of dialyzed patients’ hydration level is an important clinical aspect of the quality of their treatment. There are many tools to assess the hydrate status. One of them is electrical bioimpedance used for measuring the hydration level [1, 2]. Apart from measuring the hydration level, it is also being used for monitoring the adequacy of dialysis (it allows calculation of the total body water volume V, needed for calculating the Kt/V factor), as well as for determining the level of patients’ nourishing [3]. It is also being used for assessing the hydration level among peritoneal dialyzed patients, diagnosed with neoplastic diseases, patients after organ transplantations and the ones infected with HIV virus [4].

Precise evaluation of hydration level in patients with chronic renal failure requires gathering of important data concerning assessment of size of water compartments, such as: TBW (total body water), ECW (extracellular water), ICW (intracellular water) and interstitial compartment. Fluid is removed during dialysis with the use of ultrafiltration mainly from intravascular space. The introduction of electrical bioimpedance became useful in assessment of the size of TBW and ECW – the method is based on electrical resistance evaluation in body tissues with relation to an alternating current with various frequency amplitude [5]. Most of reports describe a method of whole body bioimpedance assessment (WBIA) based on placing the electrodes on a palm and foot. It is possible to choose an option with a usage of one current frequency or a multifrequency option with an amplitude from a few to a few hundred (500) kHz. Whole body bioimpedance assessment is dependent on changes in body position, therefore an introduction of independent body segments assessment, such as upper extremities, trunk, lower extremities, results in a more precise evaluation of state of hydration and dynamical changes during dialysis session.

Electrical bioimpedance is a method that uses electrical features of a tissue subjected to an alternating current action. Main rules of bioimpedance technique were first mentioned by Thomassett in 1963. However, big interest in this technique appeared in the early seventies of last century when Nyboer showed a correlation between bioimpedance value assessed with a use of an alternating current and changes in blood volume [6]. These measures are based on an elementary principle that electrical resistance of a cylinder is directly proportional to the length and inversely proportional to the cross section area of the cylinder multiplied by the density.

$$Z=qL/A \qquad (1)$$

where: Z – impedance (Ohm), g – tissue density (Ohm/cm), L – cylinder's length, A – cross-section area of cylinder

$$ZxL/L=qxL/A \text{ x } L/L \qquad (2)$$

Since A x L = V (volume, cm^3), we receive a following correlation:

$$V=qxL^2/R \qquad (3)$$

where: L (cm) – length of cylinder, R (cm) – electrical resistance, a q – cylinder density

The assumption, that a human body is a sum of homogenous cylinders and a current can run through extracellular and intracellular space, was an impulse for Hoffer in 1969 to apply this method in measuring the total body water (TBW). Presently, the most commonly used is the multifrequency bioimpedance method with the use of frequency spectrum between 5 and 500 kHz. When using a low frequency current (<10 kHz), cell membrane acts as an isolator and prevents permeating of electric current into a cell. A high frequency current penetrates cell membrane and reaches both extracellular and intracellular space.

Taking into account the voltage quantity we calculate the electrical resistance (impedance), which is next converted in proportion to cylinder's volume (upper extremity, lower extremity, trunk) with the use of the formula (3).

The terms: electrical impedance and electrical resistance, are often used interchangeably. In fact impedance comprises function of inductive resistance (R) and capacitive resistance (Xc).

$$Z^2=R^2+Xc^2 \quad . \qquad (4)$$

The inductive resistance (R) refers to extracellular fluid impedance, whereas capacitive resistance (Xc) refers to intracellular fluid resistance. Concluding from the formula (1), impedance is a function of cylinder's length (upper extremity, lower extremity, rib cage) and cross-sectional area of the cylinder (A) with given current frequency. The inductive (R) and capacitive (Xc) resistance values depend on the frequency of alternating current. The correlation between capacitive and inductive resistance is shown in Fig. 1 on the left.

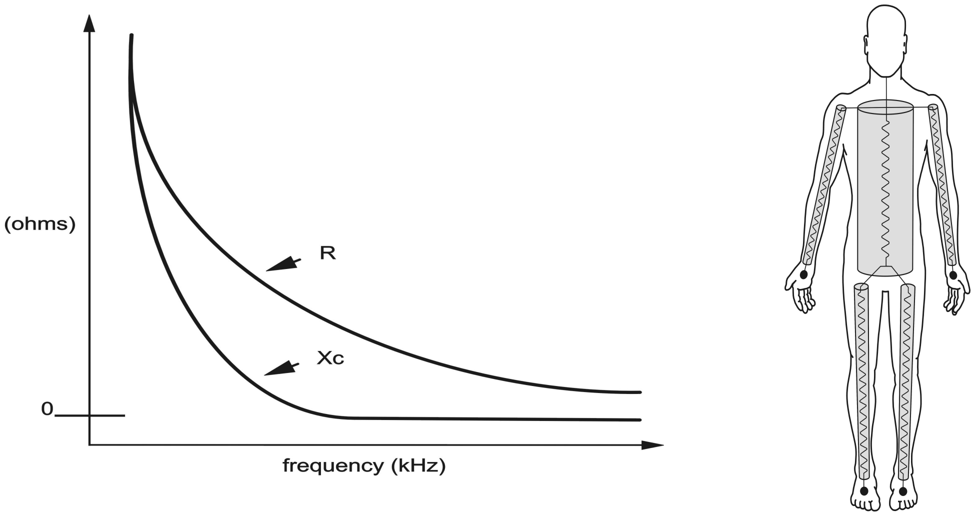

Fig. 1. On the left the relation between capacitive (R) and inductive (Xc) resistance with given frequency of an alternating current; R – inductive resistance, Xc – capacitive resistance [7]. On the right a human body as a conductor comprising five cylinders [7]

Impedance is a function of conductor geometry, which as a rule is assumed to be a cylinder. The bioimpedance assumption that a human body is a conductor and consists of five cylinders, is shown in the Fig. 1 on the right.

Based on the analysis of the formula, one may assume that extremities, because of their size, will have a bigger share in the total impedance. Therefore the total of 80 % of impedance consist of lower an upper extremity resistance, which after conversion into total body water gives 30 %. With the use of bioimpedance measurement it is difficult to evaluate total amount of water in a trunk (70% TBW).

MEDICAL MEASUREMENTS

The study was performed among 50 hemodialyzed patients and 46 healthy volunteers. Inclusion criteria were the following: patients with diagnosed terminal renal insufficiency were included in the study, age between 18 and 80 years, clinically stable. Exclusion criteria were the following: mental problems that could terminate the study in any way, pregnancy or lactation, amputation of a lower limb, implanted pacemaker, severe hemostatic circulatory insufficiency. Following parameters were measured in each healthy volunteer: body mass (in kg), height (in cm), blood pressure, TBW, ECW, ICW. Following parameters were measured in each patient: body mass before and after hemodialysis (in kg), height of patient (in cm), blood pressure before hemodialysis, TBW, ECW, ICW. Body mass of a patient was measured with the use of scale with an acceptable deviation of 0.1 kg. Height of a patient (in cm without shoes) was measured with the use of a standard measure. For the purpose of bioimpedance measure a bioimpedance analyzer was used (a Xitron Hydra 4200 Bioimpedance spectroscopy device measuring at 50 frequencies between 5 kHz and 1 MHz) with electrodes (7.7 x 1.9 cm^2). All parameters were measured at the beginning of hemodialysis (not during it so that errors in evaluation were avoided, as the greatest fluid distribution occurs within first hour of hemodialysis).

Measures were performed at the beginning of hemodialysis. Measures before hemodialysis were performed in 10 minutes after the moment a patient was lain down. Bioimpedance was measured in a logarithmic spectrum of 10 frequencies starting from 5 to 500 kHz (analyzer, Xitron Hydra 4200 Bioimpedance spectroscopy device) with electrodes (7.7 x 1.9 cm^2). Two electrodes inducing an alternating current were placed dorsally on hands (I1) and ankle (I2) of the same body side. Measuring electrodes were placed on a wrist (S1) and ankle (S2). A computer was used to collect and store the data.

DATA RETRIEVAL METHODOLOGY

Decision tree, a typical data mining (retrieval) method is described in a internet medical dictionary as „a graphic construct showing available choices at each decision node of managing a clinical problem along with probabilities (if known) of possible outcomes for patient's freedom from disability, life expectancy, and mortality. In computer science its properties are the following: each internal node tests an attribute (a column in data collection), each branch corresponds to attribute value, each leaf node assigns a classification.

In more general, a classifier is a mapping from a (discrete or continuous) variable space *X* to a discrete set of classes denoted by labels *Y*. Learning classifiers are divided into unsupervised learning and supervised classifiers. The first ones need training sets labeled by experts in order to obtain knowledge about classes e.g. about a patient class. The second ones may be able to make the proper classification only with the help of raw data and special distance measures between examined patients – points in the multivariable space[9]. This space can be visualized in two dimensions after scaling based e.g. on principle component analysis (PCA), which transform real dimensions to artificial ones, where first ones have the most correlations within and the rest of dimensions may be omitted with small resulting errors.

DATA RETRIEVAL RESULTS

During measurements sets were obtained from two groups, which were named a control (healthy volunteers) and test group (hemodialyzed patients). The test group was consisted of 50 patients and the test one of healthy young medical staff (46 people). Raw data were collected by medical equipment. Its variables were the following: body mass index (BMI) - a ratio of weight to height used as a quick measure of health status, BMI values from 18-24.9 are desirable [kg/m^2], intracellular water (ICW) - water volume inside the body cells. (i.e., water in the “living” cells), extracellular water (ECW) - water volume outside the body cell mass (i.e., water in the “inactive” cells), total body water (TBW) - sum of ICW and ECW, ECW_TBW - ECW divided by TBW, ECW_mass - ECW divided by body mass, height, weight and at the end age.

Zscored, discretized data (three levels) and data retrieval results were computed in R language environment[8] in order to find a simple rule for recognizing health problems. Z-score, so called standard score is equal to raw score minus mean of this raw score population, the subtraction result divided by its standard deviation, used in statistics. All data except age were zscored separately for women and men and for age five intervals from infinity to mean plus standard deviation, from mean plus standard deviation to mean and so on up to minus infinity.

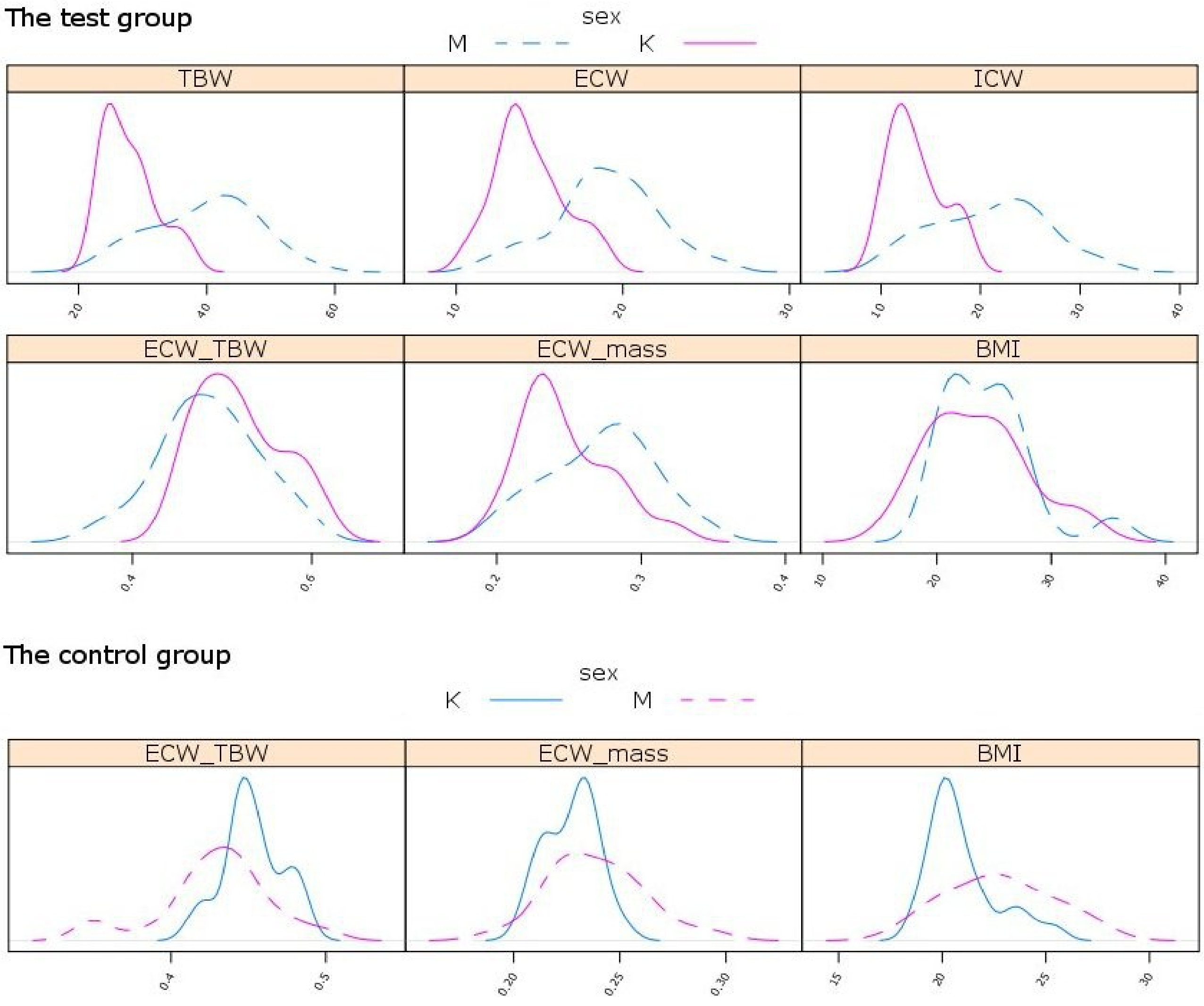

Fig. 3 Marginal plots of our dataset variable raw data

Raw data marginal plots from Fig. 3 are similar in control and test groups except ECW_TBW and BMI, which distributions are different in both groups. Thin, continuous lines are associated with women (K). The dashed lines are connected with men (M).

In Fig. 4 the all zscored variable correlation cluster dendogram is depicted. The shorter arms the stronger correlation between joint variables. The separated analyses for control and test group show the similar correlations, so in the dendogram data obtained from both groups were used. The strongest correlation is between ICW and TBW. ECW_TBW has the smallest correlation, the next one is ECW_mass together with height.

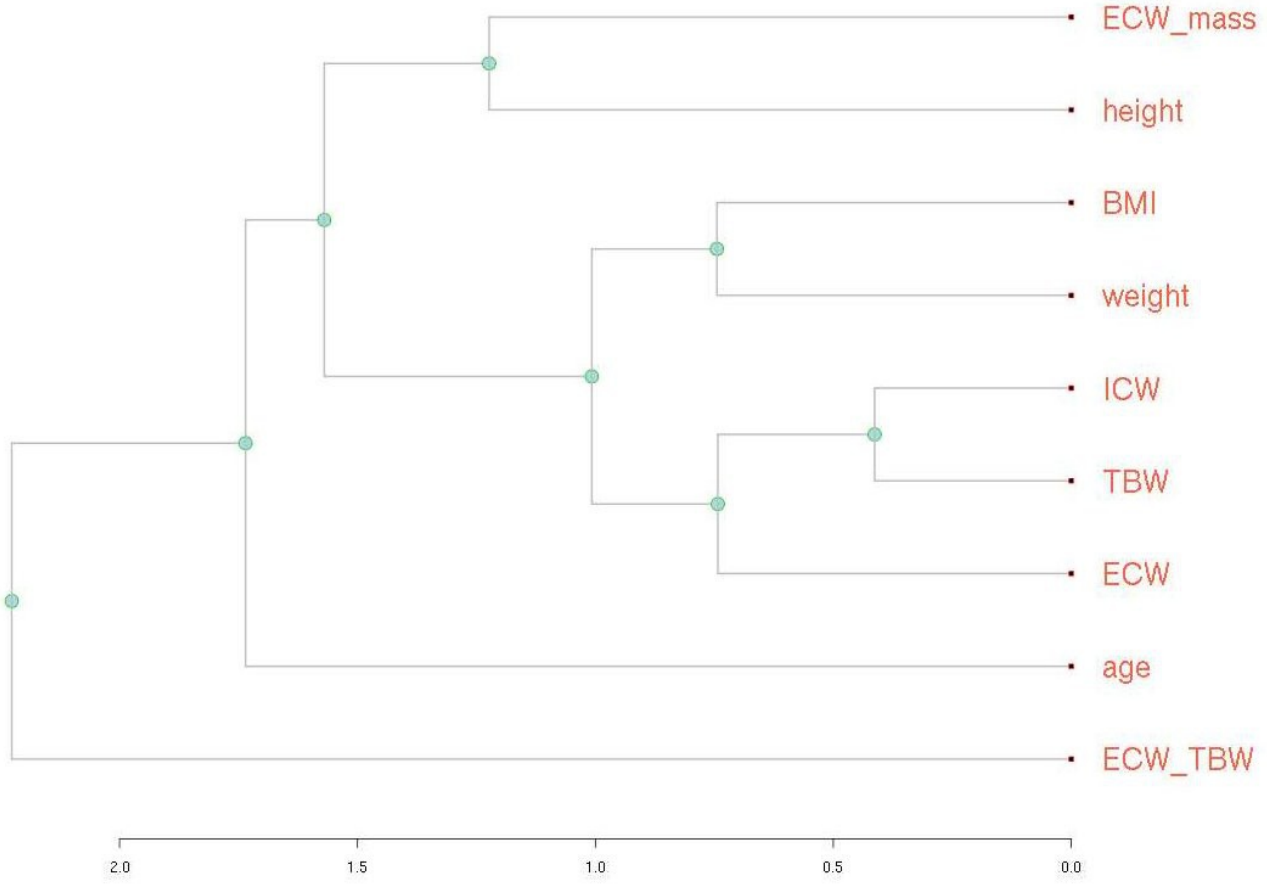

Fig.4 Correlation clusters dendogram of zscored data.

In next experiments the group marker variable was also added: P – for patients, and K – for control group. In the search for effective classifiers the several ones were tested on the training set sampled randomly from original data without replacements, usually 50%. The rest of our data was used for testing purposes.

The decision tree classifier from Fig.5 (on the left) was calculated from 6 discretized variables TBW, ECW, ICW, ECW_TBW, ECW_mass, BMI with grouping mentioned marker values in the leaves. Only ECW_TBW and TBW were chosen for nodes in this tree by J48 algorithm from RWeka library. After some next trials TBW was sometimes replaced by BMI and ECW_TBW rarely by ECW_mass or all four variables were very rarely placed in nodes or only ECW_TBW was chosen for one node. Despite this diversity the tree from Fig.5 is one of the most efficient tree classifiers and node ECW_TBW was the „must have” node in the majority of decision trees despite being not correlated with other variables in the control group and the examined group.

Based on PCA k-means clustering without previous training performed on zscored data except sex, group, age parameters revealed that artificial clusters are connected with TBW discretized levels: lower, mean, higher (on the right in Fig. 5), but not with the ECW_TBW distribution. Therefore, further experiments with different distance measures and cluster methods (e.g. fuzzy, hierarchical, divisive, agglomerative, with dissimilarities or raw data) prove that supervised (without training) classification method of calculating patient and control group split does not exist and such clusters are associated with TBW levels, even after TBW variable removal from the given data set.

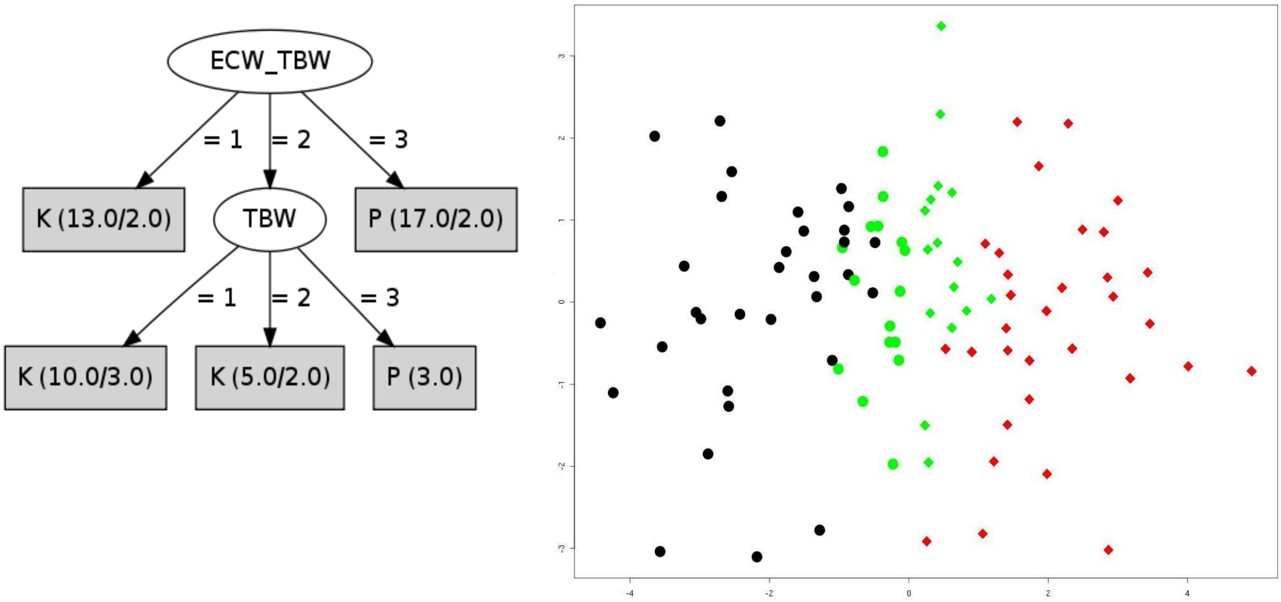

Fig. 5 On the left the decision tree generated by J48 Rweka procedure and on the right multidimensional scaling of given population generated two k-means clusters (rhombs and circles denotes separated clusters, and red, green, black colors – TBW levels: lower, mean, higher)

So after obtaining a good decision tree the next learning based classifiers were tested. The k-nearest neighbor and naive bayes classifiers were not appropriate for the given data set, though the naive bayes for raw data were sometimes better than the previous ones.

At the end linear discriminant analysis and quadratic discriminant analysis classifiers were tested. The first one has better results on raw data, and the more precise second one – on the zscored data. The second one was also better than J48 tree as is proved in depicted in Fig. 6 ROC curves: on the left for the J48 decision tree, on the right for the quadratic discriminant analysis classifier. The area under the ROC curve called AUC in the first case was equal to 0.795 and in the second case to 0.944. The larger AUC the better classifier. After many trials it was verified that AUC values and ROC curves are depended on the chosen sample set rows obtained from raw data, for different sets different results and generated classifiers, even very rarely tree classifier was better. But for most cases tree classifier AUC mean was about 0.8 and quadratic discriminant classifier AUC mean – above 0.9 (max. qdc AUC = 1).

SUMMARY

The executed experiments affirm possibilities of creating good classifiers for detecting a proper patient with the help of medical data sets, but only with previous training. Supervised methods, especially clustering, are not even adequate in the case of body water parameters. Maybe for larger sets, with greater number of variables and rows, it would be possible. In near future we are going to measure more variables with the help of more sophisticated medical equipment and evaluate all possible classifiers once again.

In this paper we proved that computer science techniques can be a great help for medical physicians in: improving predictive abilities of different tests, making a better differential diagnosis.

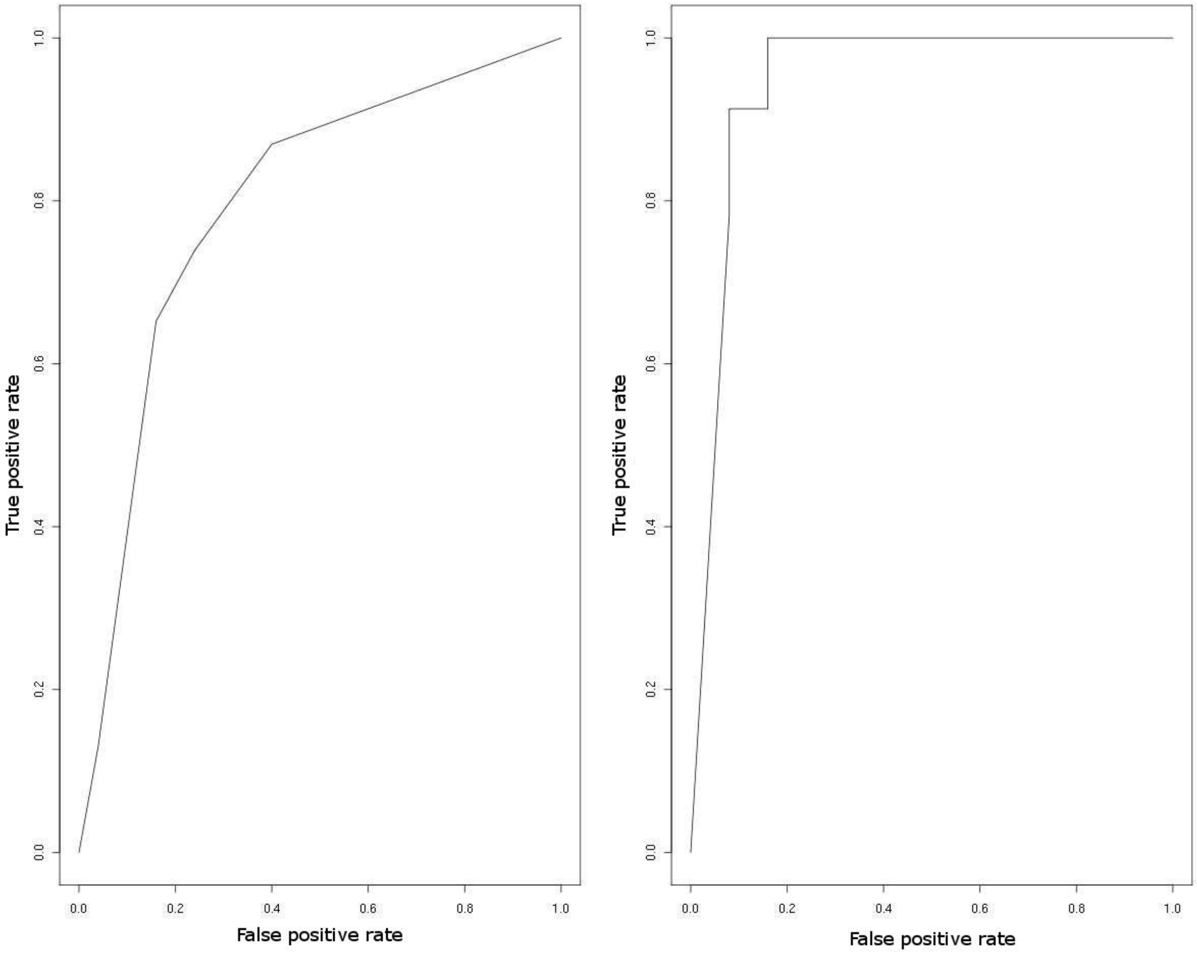

Fig. 6 ROC curves: on the left from decision tree, on the right from the quadratic discriminant analysis classifier.

REFERENCES

[1]. Guida B., De Nicola L., Trio R., Pecaroro P., “Comparison of Vector and Convetional Bioelecrical Impedance Analysis in the Optimal Dry Weight Prescription in Hemodialysis,” American Journal of Nephrology 20, 311-318 (2000).

[2]. Volker Wizemann, Peter Wabel, Paul Chamney, Wojciech Załuska, Ulrich Moissl, Christiane Rode, Teresa Małecka-Massalska, Daniele Marcelli., “The mortality risk of overhydration in haemodialysis patients,” Nephrology Dialysis Transplantation 46, [online] (2009).

[3]. Dumler F., Kilates C., “Use of bioelectrical impedance techniques for monitoring nutritional status in patients on maintenance dialysis,” Journal of Renal Nutrition 3, 116-124 (2003).

[4]. Schwenk A., Beisenherz A., Romer K., “Phase angle from bioelectrical impedance analysis remains an independent predictive marker in HIV-infected patients in the era or highly active antiretroviral treatment,” American Journal of Clinical Nutrition 72, 496-501 (2000).

[5]. Scharfetter H., Wirnsberger G.H., Holzer H., “Influence of ionic shifts during dialysis on volume estimations with multifrequency impedance analysis,” Medical & Biological Engineering & Computing 96-102 (1997).

[6]. Spiegel D.M., Bashir K., Fisch B., “Bioimpedance resistance ratios for the evaluation of dry weight in hemodialysis,” Clinical Nephrology 2, 108-114 (2000).

[7]. Kushner R., “Bioelectrical Impedance Analysis: A Review of Principles and Applications,” Journal of the College of Nutrition 11 (2), (1992).

[8]. Venables, W. N., Ripley, B. D., “Modern Applied Statistics with S”, Fourth edition, Springer, (2002).

[9]. Ripley, B. D., “Pattern Recognition and Neural Networks”, Cambridge University Press, (1996).