International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 02 - Issue 03 || March 2017 || PP. 51-61

On the Prediction of Early Stages of Chronic Renal Impairment

Pinar Yildirim[1], Ljiljana Trtica Majnarić[2], Markus Plass[3], Andreas Holzinger[3,4]
[1]*Department of Computer Engineering, Faculty of Engineering & Architecture, Okan University, Istanbul, Turkey*
[2]*Department of Internal Medicine, Family Medicine and History of Medicine, School of Medicine, University J.J. Strossmayer Osijek, Osijek, Croatia*
[3]*Institute for Medical Informatics, Statistics & Documentation Medical University Graz, A-8036 Graz, Austria*
[4]*Institute of Information Systems and Computer Media, Graz University of Technology, A-8010 Graz*

Abstract: It has been recognized that chronic renal impairment can modify the level of the global CV risk. This fact is of the great importance for planning prevention of CV disease, since mild-moderate chronic renal impairment occurs in an adult population with high frequency of 5%-10%. While clinical and biochemical disorders which can contribute to high CV risk of patients with end-stage-renal-disease are well known, these disorders in people with early stages of chronic renal impairment are poorly defined. In order to better clarify this issue, we presented here a predictive approach where data mining algorithms were performed and evaluated on a dataset composed of a large number of clinical parameters,which describe the health status of older subjects, primary healthcare (PHC) attenders, burdened with multiple medical conditions and CV risk factors. Results were compared with evidence on CV risk factors in end-stage-renal-disease.

Keywords: Chronic Renal Impairment, J48, J48-Graft, LMT, RandomTree, REPTree

I. Introduction

The classical CV risk factors have been known for a long time and include: older age, hypertension, cigarette smoking, elevated serum cholesterol, diabetes and obesity [1]. New CV risk factors have emerged over time, including : low-grade inflammation, latent infections caused with Epstein-Barr virus (EBV), Cytomegalovirus (CMV) and Helicobacter Pylori (HP) infections, increased serum homocysteine (a sulfur-containing amino acid) concentrations and complex clinical conditions: the metabolic syndrome (a cluster of CV risk factors typically including dyslipidemia characterized with increased serum triglycerides and decreased HDL-cholesterol), chronic renal impairment and inflammatory rheumatic disease [2,3].

End-stage-renal-disease (ESRD) and haemodialysis (a device dependent treatment for patients with ESRD) are known as conditions with unexpectedly high risk for the development of CV disease [4]. Although classical CV risk factors, hypertension and diabetes, are the leading causes of ESRD, these two conditions, taken alone, are not sufficient to explain a full range of CV risk in patients with ESRD [5]. More over, it has been showed that lowered, rather than elevated values of classical CV risk factors, can provide the strong associations of ESRD with CVD [6]. This paradoxical result is considered to bea consequence of protein-energy malnutrition and chronic low-grade inflammation, two shared and commonly occuring pathophysiologic disorders in patients with ESRD [7]. Decreased clearance of inflammatory and other toxic metabolites, due to decreased renal function, and multiple comorbid conditions, known to accompany ESRD, have been proposed as causes of these pathophysiologic disorders [8]. It has been recently recognized that CV risk progressively increases as the renal function declines and that it is already elevated in the earliest stages of chronic renal impairment [9]. Recommendations have been included, in CV risk estimation systems, that the global CV risk level should be modified for the impact of chronic renal impairment [10,11].This fact could be of a great importance for planning prevention in general population, since mild-moderate chronic renal impairment is a frequent disorder, encompassing 5%-10% of adult population [12]. While ESRD is known for a variety of biochemical, metabolic, endocrine and immunologic disorders, which all may contribute to high CV risk in patients with ESRD,the exact relationships of these disorders with earlier stages of chronic renal impairment and their distribution within the patient populations are still unknown [13].

II. Related Work

Attempts to use data mining methods to comprehensively capture comorbid disorders in patients with chronic renal impairment have not been registered so far. Similar study is that of Shital Shah and others (2003) where they applied data mining on a large set of routinely monitored parameters to predict survival of haemodyalisis patients[14]. Some of the extracted parameters corresponded with the existing knowledge, while

some others were new and can be used to improve future clinical studies` and data collection protocols. Another example is a study performed in Tehran (Iran), where the authors showed, similar as to our study, that simple, easily available clinical parameters can be useful to improve decision making in clinical settings [15]. In this study, data mining method was applied on a large set of health, laboratory and echocardiography parameters to improve selection of patients with coronary artery disease who should undergo angiography.

III. Materials and Methods

1. Data Sources

In order to better cope with the complexity of a chronic disorder such as chronic renal impairment, where many pathogenetic components and their distribution within the patient population are still unknown, we reached out for the innovative multicomponent method based on using data mining algorithms.

In our study, the patient sample consists of 93 subjects, 35 M/58 F, aged 50-89 years (median 69), primary health care (PHC) attenders living in the urban area of the town of Osijek, the north-eastern Croatia, the region with the high prevalence of chronic diseases. Performed dataset contains 61 medical variables, only low-cost, easily available parameters, many of which are routinely collected data used from patients` health records. An idea was to determine the health status of examined patients systematically, by many aspects (Table 2). Hidden clinical contexts and new pathogenetic pathways, otherwise unvisible in clinical studies, are expected to be detected in this way, which, in turn, may facilitate further research.

Table1. Attributes used in the dataset (Num:Numeric, Nom:Nominal) (1-30)

No	Attribute	Description	Type
1	age	age (years)	Num
2	sex	(M=Male, F=Female)	Nom
3	Hyper	Hypertension (yes, no)	Nom
4	DM	Diabetes mellitus (yes IGT=Impaired glucose tolerance No)	Nom
5	Fglu	Fasting blood glucose (mmol/L)	Num
6	HbA1c	Glycosilated Haemoglobin (%) (average blood glucose during last three months)	Num
7	Chol	Total Cholesterol (mmol/L)	Num
8	TG	Triglycerides (mmol/L)	Num
9	HDL	HDL-cholesterol (mmol/L)	Num
10	Statins	Therapy with statins (yes,no)	Nom
11	Anticoag	Therapy with anticoagulant/antiaggregant drugs (yes,no)	Nom
12	CVD	Cardiovascular diseases (yes, no) (myocardial infarction, angina, history of revascularisation, stroke, transient ischaemic cerebral event, peripheral vascular disease)	Nom
13	BMI	Body Mass Index (kg/m2)	Num
14	w/h	Waist/hip ratio	Num
15	Arm cir	Mid arm circumference (mm)	Num
16	skinf	Triceps skinfold thickness (mm)	Num
17	Gastro	Gastroduodenal disorders (yes,no) (gastritis, ulcer)	Nom
18	uro	Chronic urinary tract disorders (yes,no) (recurrent cystitis in women, symptoms of prostatism in men)	Nom
19	COPB	Chronic obstructive pulmonary disease (yes,no)	Nom
20	aller d	Allergy (Rhinitis and/or Asthma) (yes,no)	Nom
21	dr aller	Drugs allergy (yes, no)	Nom
22	analg	Therapy with analgetics/NSAR (yes,no) Patients with osteo-arthritis	Nom
23	derm	Chronic skin disorders (yes,no) (chronic dermatitis, dermatomycosis)	Nom
24	neo	Malignancy, including skin malignancy (yes,no)	Nom
25	OSP	Osteoporosis (yes, no)	Nom
26	Psy	Neuropsychiatric disorders (yes,no) (anxiety/depression, Parkinson`s disease, cognitive impairments)	Nom
27	MMS	Mini Mental Score – test for screening on cognitive dysfunction Max Score =30 Score <24 indicates ognitive impairment	Num
28	CMV	Cytomegalovirus specific IgG antibodies (IU/ml)	Num
29	EBV	Epstein-Barr virus specific IgG (IU/ml)	Num

30	HPG	Helicobacter pylori specific IgG (IU/ml)	Num

\

Table2. Attributes used in the dataset(Num:Numeric, Nom:Nominal) (31-61)

No	Attribute	Description	Type
31	HPA	Helicobacter pylori specific IgA (IU/ml)	Num
32	LE	Leukocytes Number x109/L	Num
33	NEU	Neutrophils % in White Blood Cell differential	Num
34	EO	Eosinophils % in White Blood Cell differential	Num
35	MO	Monocytes % in White Blood Cell differential	Num
36	LY	Lymphocytes % in White Blood Cell differential	Num
37	CRP	C-reactive protein (mg/L)	Num
38	E	Erythrocytes number x1012/L	Num
39	HB	Haemoglobin (g/L)	Num
40	HTC	Haematocrite (erythrocyte volume blood fraction)	Num
41	MCV	Mean cell Volume (fL)	Num
42	FE	Iron (g/L)	Num
43	PROT	Total serum proteins (g/L)	Num
44	ALB	Serum albumin (g/L)	Num
45	HOMCIS	Homocistein (µmol/L)	Num
46	ALFA1	Serum protein electrophoresis (g/L)	Num
47	ALFA2	Serum protein electrophoresis (g/L)	Num
48	BETA	Serum protein electrophoresis (g/L)	Num
49	GAMA	Serum protein electrophoresis (g/L)	Num
50	RF	Rheumatoid Factor level (IU/ml)	Num
51	VITB12	Vitamin B12 (pmol/L)	Num
52	FOLNA	Folic acid (mM/L)	Num
53	INS	Insulin (µIU/L)	Num
54	CORTIS	Cortisol in the morning (nmol/L)	Num
55	PRL	Prolactin in the morning (mIU/L)	Num
56	TSH	Thyroid-stimulating hormone (IU/ml)	Num
57	FT3	Free triiodothyronine (pmol/L)	Num
58	FT4	Free thyroxine (pmol/L)	Num
59	ANA	Antinuclear antibodies (autoantibodies) (µIU/ml)	Num
60	IGE	IgE (kIU/L)	Num
61	Clear	Creatinine clearance (ml/s/1.73m2)	Nom

Nominal parameters, used in the dataset, indicate age and sex, diagnoses of the main groups of chronic diseases, information on drugs use and anthropometric measures. A wide set of laboratory tests was also performed (numerical variables), to differentiate between variations in age-related pathophysiologic disorders, including information on: inflammation, the nutritional status, the metabolic status, latent infections, humoral (antibody-mediated) immunity and the neuroendocrine status (Table 1-2). One laboratory parameter, "creatinine clearance", indicating diagnosis of chronic renal impairment, was used as the target variable. The cut-off value of this variable of 1.84, was used to separate patients into two groups: those with good and those with lowered renal function. Namely, "creatinine clearance" is a measure of reduced glomerular filtration rate (GFR). GFR of 65%-60%, corresponding with the cut-off value of serum creatinine clearance of 1.84 ml/s/1.73 m^2, is considered as being still satisfactory for the majority of patients, especially for those of older age (Table 1-2) [16].

2. Classification

Classification is a method to build models from a dataset (numeric and/or categorical variables) with the help of data mining (machine learning). Decision tree (DT) classification is a method to generate a tree (which contains out of nodes and leafs) out of a dataset [17]. The node is a point who evaluates which of the children should be visited next.Other methods for classification are for example rule-based classifiers, neural networks and support vector machines.

The goal of a DT is to predict the leafbased on the other classes; in this case the leaf is "normal renal function" or "renal impairment".DT-generation based on finding for every node the best fitting children. The way how DT was built is nearly the same in all DT-algorithms:
Pseudo-code for building a DT [18]:
Check for base cases
For each attribute *a*
Find the feature that best divides the training data such as information gain from splitting on *a*
Let *a* best be the attribute with the highest normalized information gain
Create *a* decision node node that splits on *a*_best
Recourse on the sub-lists obtained by splitting on *a*_best and add those nodes as children of node

To find the best splitting attribute for example the GINI Index is calculated. Let p(i|t) denote the fraction of records belonging to class i at a given node t and c the number of classes (1).

$$(1) GINI(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

The output of the GINI Index is the impurity a number between 0 and 1. Zero means no impurity, so the class with the smallest GINI Index is the best splitting parameter.
Another way to find the splitting attribute is Entropy (2). This is a degree of disorder in data.

$$(2) Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) * \log_2[p(i|t)]$$

If the Entropy is one, the data is in disorder.
After this the goodness of the splitting class (Information Gain) could be calculated (3).

$$(3) InformationGain = \text{Entropy}(node) - \sum_{j=1}^{k} \frac{N(v_j)}{N} * \text{Entropy(child)}$$

Where $N(v_j)$ is the number of records associated with the child and N the number of records of the node.
The DT can be generated and tested in different way for example "training set","cross-validation" and "Supplied test set".

2. Training set:

With "Training set" the whole dataset D is used to generate the DT. In the evaluation step D is used again to test the tree.

3. Cross-validation:

With "Cross-validation" the dataset D is randomly divided into k (number of folds) subsets D_k. For generating the DT, the algorithm runs k times. In each turn the algorithm is trained on $D\backslash D_k$ and tested on D_k[19]. The accuracy and other parameters were then calculated by the average.

4. Supplied test set:

In this method the algorithm is trained with the dataset D_1 and tested with another dataset D_2. A tool for generating DT is WEKA which is a free Data Mining Software [20] programmed in Java.There are lots of different DT classification algorithms in WEKA. Some of them are J48, J48-Graft, REP Tree, Random Tree and Logistic Model Tree. All of these algorithms can handle nominal and non-nominal attributes.

5. J48 Algorithm

J48 is the Weka implementation of the C4.5 algorithm, based on the ID3 algorithm. The main idea to generate the tree is to use the information entropy as described above.

For each node the most effectively split criteria is calculated and then subsets were generated. To get the split criteria the algorithm looks for the attribute with highest normalized information gain.

The last step is called pruning, in this step the algorithm starts at the bottom of the tree and removes unnecessary nodes, so the height of the tree can be reduced by deleting double information.

6. J48-Graft Algorithm

J48-Graft is an improvement of the J48 algorithm. Grafting tries to add more nodes to the tree. The DT is built like in the J48 algorithm, but at the end there is another process.

In the grafting process the algorithm looks at the wrong classified parameters (based on the training set) and tries to add a new splitting attribute and adds a complete new branch to reduce the classification error. One of the main problems with grafting is not to over fit the DT.

The combination of pruning and grafting helps to increase the predictive accuracy with the disadvantage of a big DT[21]. This works because pruning only removes local information and grafting uses all attributes to improve the tree.

7. REPTree Algorithm

Reduced Error Pruning Tree (REP Tree) is a bottom-up approach. The algorithm starts with the leaf of the tree and searches the best fitting node, so there is no need for grafting or pruning[22]. This process is repeated till there is one node reaming.

8. RandomTree Algorithm

Random Trees are a combination of single model trees and Random Forest. The main idea is that every node has as subset a linear regression model [23]. So every node has its own linear model. Instead of calculating the best possible split with all attributes for each node, only a random subset is used, this method is increasing the speed of the algorithm.

9. LMT Algorithm

Logistic Model Tree (LMT) combines logistic regression functions with a standard DT[24]. For each numeric node this algorithm generates a logistic regression model. This model can contain more than one attribute.

In the most cases the output is a smaller tree then generated by J48 or similar algorithms.

IV. RESULTS

The number of datasets, the tree is generated of, is 93 and includes 61 attributes. The attributes are shown in Table 1 and Table 2. The outcome variable is "Clear". The numeric value "Clear" is spitted into a nominal class, if "Clear" <1.84 it is classified as renal impairment and if "Clear" ≥ 1.84 it is classified as normal renal function. 35 of the datasets were classified as normal renal function. The mean age of the candidates is 67.65 years (from 47 to 89 years) and 62.37% were female.

To generate the DT, we used for 10 folds cross-validation for different classification algorithms. The advantage of "cross-validation" is that the tested subset D_k is more independent form the output of the algorithm than in "Training set". So the accuracy probably is lower than in "Training set", but it has more reference to clinical practice. In Fig. 1, the DT generated with using J48-Graft Algorithm and 10 folds cross-validation is shown.

1. Evaluation of classification results

Some evaluation indices were used for the evaluation of the classification results, where TP/TN is the number of True Positives/Negatives instances, FP/FN is the number of False Positives/Negatives instances. Figure 2 shows the performance metrics of the algorithms with 10 folds cross-validation.

Sensitivity (also known as 'TP Rate' or 'Recall') is the percentage of positives instances correctly classified (1):

$$(1)\text{Sensitivity} = \frac{TP}{TP + FN}$$

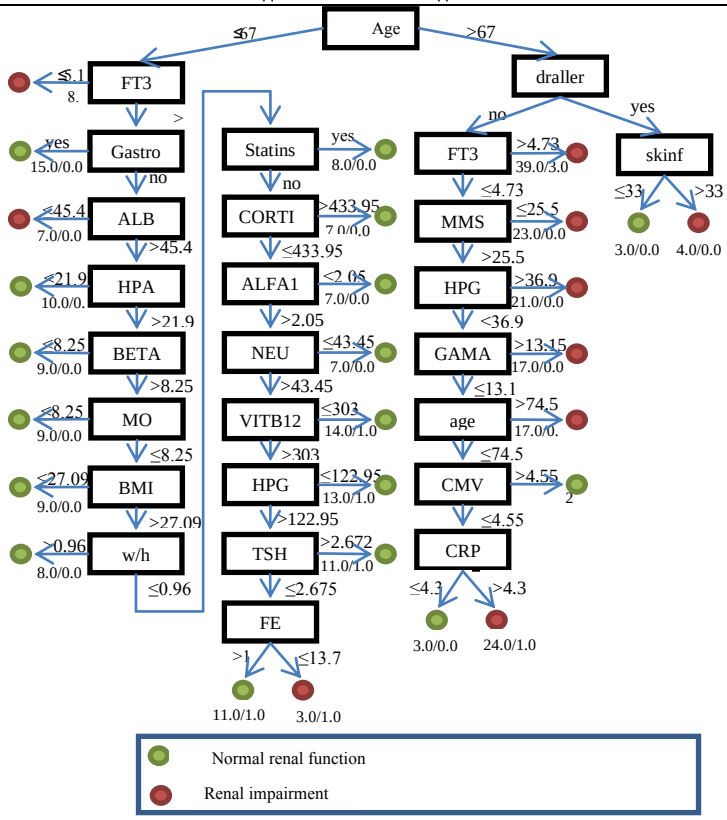

Figure.1. Decision Tree generated by j48-Graft algorithm

Specificity (also known as 'TN Rate') is the percentage of negativesinstances correctly classified (2):

$$(2)\text{Specificity} = \frac{TN}{TN + FP}$$

FP Rate (3) is the percentage of instances wrong classified (renal impairmentclassified as normal renal function).

$$(3)\text{FP Rate} = 1 - \text{Specificity} = 1 - \frac{TN}{TN + FP}$$

Precision (4) is the Number of instances that are truly of a class (normal renal function) divided by the total instances classified as that class.

$$(4)\text{Precision} = \frac{TP}{TP + FP}$$

F-Measure (5) this parameter is the harmonic mean of Precision and Sensitivity.

$$(5)\text{F} - \text{Measure} = 2 * \frac{\text{Precision} * \text{Sensitivity}}{(\text{Precision} + \text{Sensitivity})}$$

Receiver operating characteristics(ROC) Area is the area under the ROC-Curve (the comparison between True Positive Rate and False Positive Rate). This area is a measurement for the accuracy of the prediction. The value is between 0.5 for a worthless accuracy and 1.0 for a perfect accuracy[25].

According to Table 3, generallyJ48-Graft has the best prediction output with high performance metrics. The ROC area in LMT algorithm is greater than in the other algorithms, but it has its weakness in the other measures.

In Table 3 the output of the error measures are shown. Weka uses all instances and classes for calculating the error measures.

The mean absolute error (6) is calculated by the following formula [26]:

$$(6)\text{Mean absolute error} = \frac{\sum|X_t - F_t|}{m}$$

Where $X_t$is the is the real value, F_t is the forecasted value an m is the number of instances. With the root mean squared error (7) it is possible to give more weight to bigger errors (this is done by the squaring function):

$$(7)\text{Root mean absolute error} = \sqrt{\frac{\sum(X_t - F_t)^2}{m}}$$

As seen in the performance metrics of the error measures, J48-Graft algorithm has the lowest value of mean absolute error with 0.3326. Considering rooting mean squared error measure, LMT algorithm generated the best results with 0.4758.

Table 3. Performance metrics of algorithms

	J48	**J48-Graft**	**REPTree**	**RandomTree**	**LMT**
Sensitivity	0.663	0.674	0.652	0.63	0.63
Specificity	0.584	0.59	0.522	0.53	0.575
FP Rate	0.416	0.41	0.478	0.47	0.425
Precision	0.653	0.664	0.636	0.614	0.627
F-Measure	0.654	0.664	0.622	0.615	0.628
ROC Area	0.625	0.631	0.607	0.58	0.704

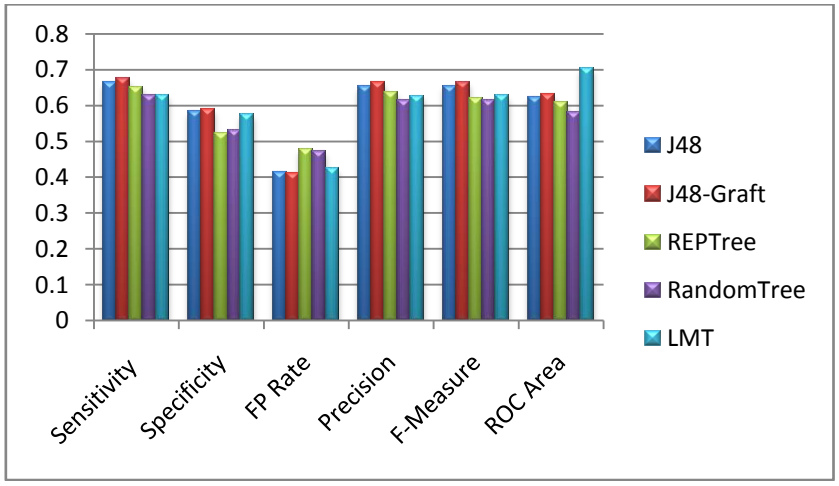

Figure. 2. Performance metrics of algorithms

Table 4. Error measures of algorithms

	J48	**J48-Graft**	**REPTree**	**RandomTree**	**LMT**
Mean absolute error	0.3399	0.3326	0.417	0.3587	0.3794
Root mean squared error	0.5675	0.5632	0.5027	0.5989	0.4758

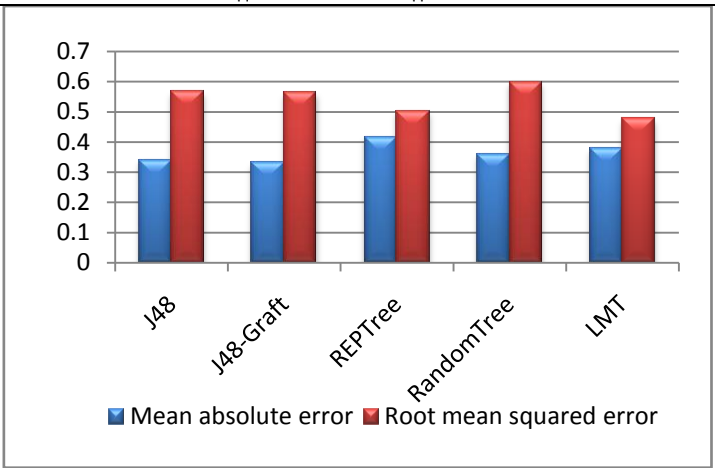

Figure.3. Error measures of algorithms

V. Discussion

The DT diagram (Fig. 1) shows the most effective clinical features which can be used to select subjects with chronic renal impairment, in the population of older PHC patients burdened with chronic diseases and CV risk factors. An important observation,made upon these results, is the cut-off point of age of 67 years, when chronic renal impairment is likely to emerge. This result is not surprising, because evidence indicates the increasing frequency of this disorder with age. Although, an information on the exact cut-off point is new and may be due to the risk factors accumulation, notably diabetes and hypertension, in this age [12]. Alternatively, this cut-off point of age might be reflective of the impact of age *per se* on the development of renal dysfunction. In this regard, evidence shows that prevalence hypertension in general population steadily increases with age, in the age group 60-74 years reaching the value of 50% (the median age of patientsin our sample is 69 years). Related to diabetes, statistical facts also indicate a substantial increase in prevalence in older age, with over 25% of population aged ≥ 65 years having diabetes and 50% pre-diabetes (impaired glucose metabolism) [27]. That there might be an impact of advanced age on the development of renal dysfunction, which occurs independently of these main CV risk factors, it is supported by the evidence indicating that there is a stochastic process of decline of renal GFR [28]. This conclusion, on the dominant impact of age on the decline of renal function, is likely to be supported by our results, as illustrated with the result showing that the majority of nominal variables from the input, including diagnoses of diabetes and hypertension, did not pass the classification process, while the variable "age" showed the strong distinctive power, dividing the DT diagram into the two main streams according to the cut-off point of age of 67 years (Fig. 1).

When the left-side stream of the DT diagram, assigned with the cut-off point of age of ≤ 67 years, is considered, it can be realized that there is no rationalefor screening people of this age on chronic renal impairment, because there are only three outcomes in this part of the diagram, allowing selection of only a small proportion of subjects (Fig. 1).

When the right-side DT diagram is considerated, several implications of practical importance can be drown out. As the first, many of the selected parameters, including: "draller", "HPG", "GAMA", "CMV" and "CRP", are likely to indicate chronic inflammation and the immune system dysfunction (Fig. 1). On the contrary to what is expected according to the evidence, parameters indicating metabolic disorders, known to have a major role in the development of CV disease in patients with ESRD, are missing here, which means that there might be differences between ESRD and earlier stages of chronic renal impairment in clinical expression of CV risk factors.It is possible, as based on our results, that the non-conventional CV risk factors, including inflammation and the immune system dysfunction, are the prevailing ones in early stages of chronic renal impairment, while metabolic disorders might give greater contribution in ESRD. Alternatively, there may be two separate groups of patients from the earliest stages of chronic renal impairment: one group with increased level of inflammation and decreased level of metabolic CV risk factors and another one characterized with increased level of metabolic risk factors. The rational for these explanationscan be found inthe recently publishedpapersindicatingthat low, rather than increased values of classical CV risk factors, are likely to provide the strongest associations of ESRD with CV disease[6].

Another interesting result is related to the branching of the right-side DT diagram into the two main streams, the short-armed and the long-armed (Fig. 1). The critical check points, in the short-armed branch, include variables indicating allergy to drugs (the variable "draller") and malnutrition (the variable "skinf", an anthropometric measure of muscle mass loss). The meaning of this result might be in the possible causal

relationships between these two disorders, in patients with chronic renal impairment,which is a new and intriguing idea, but not without sense, when taking into account the known fact on the associations between genetic variations inthe activity level of the drug-metabolizing enzymes andmedical conditions characterized with immunodeficiency and muscle wasting [29].

There could be one more interesting observation, arising from this part of the DT diagram. This observation isrelated to the results showing that the two variables, "skinf", indicating protein malnutrition, and "CRP", indicating inflammation, both elements of the unique "inflammation-malnutrition syndrome", found in patients with ESRD, are separately selected here, that means, within the two independent branches (Fig. 1). This observation is likely to indicate a substantial pathogenetic distance between inflammation and malnutrition, specifically in patients with early stages of chronic renal impairment, which is opposite to what was found in patients with ESRD. As based on our results, chronic inflammation is likely to be associated with chronic antigenic stimulation and the immune system dysfunction, probably due to the persistence of the chronic latent infections (as indicated by the position of the variable "CRP" within the same branch where there are variables indicating these infections) (Fig. 1). On the contrary, there are no associations, either of chronic inflammation (the parameter "CRP"), or muscle wasting (the parameter "skinf"), with metabolic disorders (as parameters indicating metabolic disorders were not selected) (Fig.1). These results may be due to the relative domination of particular disorders, associated with inflammation, as well as the relative contribution of inflammation and malnutrition in the "inflammation-malnutrition syndrome", in different patient samples and according to stages of chronic renal impairment [30].

As based on the longer arm of the right-side DT diagram, several parameters can be used as critical clinical features, to support decision making for selection of patients with early stages of chronic renal impairment (Fig. 1). These results might be of a great practical importance, especially in PHC setting, where a huge amount of data is collected and managed and where some of parameters, selected here, are frequently performed and may be used as surrogate markers for screening patients on chronic renal impairment. Their availability might be especially important for the reason that testing on renal function is not always easy to perform, as associated with 24^h urine specimen collection. These parameters of the proposed practical utility include: "FT3", indicating decreased thyroid gland function, "MMS", indicating cognitive impairment, "HPG", indicating chronic gastritis caused with Helicobacter pylori infection, "age over 74.5 years" and "CRP>4.3", indicating low-grade inflammation (Fig. 1). Some of these associations, suggested with our results, including association of the thyroid gland hormones disturbation and cognitive impairment with chronic renal impairment, have already been reported in previous studies [31,32]. However, the influence of chronic renal impairment on maintaining Helicobacter pylori infection is a new information which deserves further elaboration.

VI. Conclusion

Presented approach, where data mining algorithms were performed on the large dataset composed of clinical parameters which describe the health status of older subjects by many aspects, enabled many interesting insights into clinical characteristics of patients with mild-moderate stages of chronic renal impairment. The results can help physicians in decision making when screening general population on patients with earlier stages of chronic renal impairment and in understanding the relationships between multiple comorbid disorders, associated with renal function decline. In more general terms, this approach can be used as the first step approach, to identify targets for future research.

References

[1]. D.J. Lerner and W.B. Kannel, Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population, *American Heart Journal, 111(2),* 1986, 383-390.

[2]. I.J. Kullo, G.T. Gau and A.J. Tajik, Novel risk factors for atherosclerosis, *Mayo Clinic Proceedings, 75(4),* 2000, 369-380.

[3]. Emerging risk factors collaboration, Major lipids, apolipoproteins, and risk of vascular disease, *Jama, 302(18),* 2009, 1993-2000.

[4]. M.J. Sarnak, A.S. Levey, A.C. Schoolwerth, J. Coresh, B. Culleton, L.L. Hamm, P.A. McCullough, B.L. Kasiske, E. Kelepouris and M.J. Klag, Kidney disease as a risk factor for development of cardiovascular disease, A statement from the American Heart Association Councils on kidney in cardiovascular disease, high blood pressure research, clinical cardiology, and epidemiology and prevention, *Circulation, 108(17),* 2000, 2154-2169.

[5]. Q. Yao, R. Pecoits-Filho, B. Lindholm and P. Stenvinkel,Traditional and non-traditional risk factors as contributors to atherosclerotic cardiovascular disease in end-stage renal disease. *Scandinavian Journal of Urology and Nephrology, 38(5),* 2004, 405-416.

[6]. J. Park, S.F. Ahmadi, E. Streja, M.Z. Molnar, K.M. Flegal, D. Gillen, C.P. Kovesdy and K. Kalantar-Zadeh, Obesity paradox in end-stage kidney disease patients, Progress in Cardiovascular Diseases, 56(4), 2014, 415-420.

[7]. M. Suliman, P. Stenvinkel, A.R. Qureshi, K. Kalantar-Zadeh, P. Bárány, O. Heimbürger, E.F. Vonesh, B. Lindholm, The reverse epidemiology of plasma total homocysteine as a mortality risk factor is related to the impact of wasting and inflammation, *Nephrology Dialysis Transplantation, 22(1),* 2007, 209-217.

[8]. K. Kalantar-Zadeh, Recent advances in understanding the malnutrition-inflammation-cachexia syndrome in chronic kidney disease patients: what is next? *Seminars in Dialysis, 18(5),* 2005, 365-369.

[9]. G. Leoncini, F. Viazzi, D. Parodi, E. Ratto, S. Vettoretti, V. Vaccaro, M. Ravera, G. Deferrari and R. Pontremoli, Mild renal dysfunction and cardiovascular risk in hypertensive patients, *Journal of the American Society of Nephrology, 15(suppl 1),* 2004, S88-S90.

[10]. Ž. Reiner, A.L. Catapano, G. De Backer, I. Graham, M.-R.Taskinen, O. Wiklund, S. Agewall, E. Alegria, M.J. Chapman and P. Durrington, ESC/EAS Guidelines for the management of dyslipidaemias, The Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and the European Atherosclerosis Society (EAS), *European Heart Journal, 32(14),* 2011, 1769-1818.

[11]. G. Leoncini, E. Ratto, F. Viazzi, N. Conti, V. Falqui, A. Parodi, C. Tomolillo, G. Deferrari and R. Pontremoli, Global risk stratification in primary hypertension: the role of the kidney, *Journal of Hypertension, 26(3),*2008,427-430.

[12]. R.C. Atkins, The epidemiology of chronic kidney disease, *Kidney International, 67(S94),* 2005, S14-S18.

[13]. A. Luger, I. Lang, J. Kovarik, H.K. Stummvoll and H. Templ, Abnormalities in the hypothalamic-pituitary-adrenocortical axis in patients with chronic renal failure, *American Journal of Kidney Diseases, 9,*1987, 51-54.

[14]. A. Kusiak, B. Dixon and S. Shah, Predicting survival time for kidney dialysis patients: a data mining approach, *Computers in Biology and Medicine, 35(4),*2005, 3113-27.

[15]. R. Alizadehsani, J. Habibi, Z. Alizadeh-Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh and F.Alizadeh-Sani, Diagnosing Coronary Artery Disease via Data Mining Algorithms by Considering Laboratory and Echocardiography Features,*Research in Cardiovascular Medicine, 2(3),* 2013, 133-139.

[16]. L.A. Stevens, J. Coresh, T. Greene and A.S. Levey, Assessing kidney function - measured and estimated glomerular filtration rate, *New England Journal of Medicine, 354,*2006, 2473-2483.

[17]. V.P. Bresfelean, Analysis and predictions on students' behavior using decision trees in Weka environment, *Proc.29th IEEE conf. on Information Technology Interfaces*, Cavtat, Croatia, 2007, 51-56.

[18]. S.B. Kotsiantis, I. Zaharakis and P. Pintelas, Supervised machine learning: a review of classification techniques, *Informatica, 31,* 2007, 249-268.

[19]. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proc.14th Int. Joint Conf. on Arteficial intelligence*, Montreal, Quebec, Canada, 1995, 1137-1143.

[20]. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, The WEKA data mining software: an update, *ACM SIGKDD explorations newsletter, 11(1),* 2009, 10-18.

[21]. G.I. Webb, Decision tree grafting, *Proc.15th Int. Joint Conf. on Artificial Intelligence*, Nagoya, Japan, 1997, 846-851.

[22]. T. Elomaa and M. Kooriainen, An analysis of reduced error pruning, *Journal of Artificial Intelligence Research, 15,* 2001, 163-187.

[23]. B. Pfahringer, Semi-random model tree ensembles: an effective and scalable regression method, in D. Wang (Ed.), *Proc. Australasian Joint Conf. on Artificial Intelligence,* Advances in Artificial Intelligence, Lecture Notes in Computer Science 7106 (Springer,2011) 231-240.

[24]. N. Landwehr,M. Hall and E. Frank, Logistic model trees, *Machine Learning, 59(1),* 2005,161-205.

[25]. H. Jiawei, P. Jian and K. Micheline, *Data Mining, Southeast Asia Edition: Concepts and Techniques*, 2 (Berlington, Massachusetts, USA: The Morgan Kaufmann; 2006).

[26]. S.G. Makridakis and H. Michele, *Evaluating accuracy (or error) measures*, Working papers INSEAD, 18 (Windsor, Ontario: Fontaineblau Public Library, 1995).

[27]. Centers for Disease Control and Prevention, *High Blood Pressure Fact Sheet, 24(7),* 2016.

[28]. C.I. Johnston, J. Risvanis, M. Naitoh and I. Tikkanen, Mechanism of progression of renal disease: current hemodynamic concepts. *Journal of Hypertension, 16(S4),* 1998, S3-S7.

[29]. Nebert, D.W., MCkinnon, R.A., Puga, A.,(1996). Human drug-metabolizing enzyme polymorphisms: effects on risk of toxicity and cancer. DNA and cell biology, 15,273.

[30]. Lj. Trtica Majnarić, P. Yildirim, A. Holzinger, Discovery of characteristics of patients with increased level of inflammation,*Medicinal Chemistry, 5(12)*, 2015, 512-520.

[31]. J.M. Tibaldi and M.I. Surks, Effects of nonthyroidal illness on thyroid function, *Medical Clinics of North America, 69(5)*, 1985, 899-911.

[32]. J.L. Holley, The hypothalamic-pituitary axis in men and women with chronic kidney disease, *Advances in Chronic Kidney Disease, 11(4)*,2004, 337-341.