2016 IEEE International Conference on Healthcare Informatics

Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes

Asif Salekin
Department of Computer Science
University of Virginia
Charlottesville, Virginia
Email: as3df@virginia.edu

John Stankovic
Department of Computer Science
University of Virginia
Charlottesville, Virginia
Email: stankovic@cs.virginia.edu

***Abstract*—Chronic kidney disease (CKD) is a major public health concern with rising prevalence. In this study we consider 24 predictive parameters and create a machine learning classifier to detect CKD. We evaluate our approach on a dataset of 400 individuals, where 250 of them have CKD. Using our approach we achieve a detection accuracy of 0.993 according to the F1-measure with 0.1084 root mean square error. This is a 56% reduction of mean square error compared to the state of the art (i.e., the CKD-EPI equation: a glomerular filtration rate estimator). We also perform feature selection to determine the most relevant attributes for detecting CKD and rank them according to their predictability. We identify new predictive attributes which have not been used by any previous GFR estimator equations. Finally, we perform a cost-accuracy tradeoff analysis to identify a new CKD detection approach with high accuracy and low cost.**

***Index Terms*—Chronic kidney disease, machine learning, feature selection.**

I. Introduction

Chronic kidney disease is a worldwide public health problem with an increasing incidence, prevalence, and high cost. Approximately 2.5-11.2% of the adult population across Europe, Asia, North America, and Australia are reported to have chronic kidney disease [1], where in the USA alone it has affected more than 27 million individuals [2]. According to The National Kidney Foundation about 59% of all American are at risk of developing kidney disease in their lifetime [3].

The increase of CKD is partially explained by the increasing prevalence of diabetes mellitus and hypertension which are the leading risk factors for CKD. CKD promotes hypertension and dyslipidemia, which, in turn, can contribute to the progression of renal failure.

Recent studies suggest that some of these adverse outcomes can be prevented or delayed by early detection and treatment [4]. Awareness of CKD among patients is gradually increasing, but still low. According to the 2003-2004 National Health and Nutrition Examination Survey, less than 5 percent of patients with stage 1 or 2 CKD and less than 10 percent with stage 3 reported having been diagnosed with CKD; only 45 percent of patients with stage 4 were aware of their condition [5].

Since there is a relatively small number of practicing nephrologists, nephrologists cannot exclusively manage all patients with CKD. The burden of CKD management thus falls largely on primary care providers (PCPs). A recent study [6] has shown that awareness of CKD by all types of PCPs is unacceptably low and knowledge of CKD management is particularly poor among family practitioners, especially among those with more than 10 years in clinical practice and who spend more than 50% of their time practicing clinical medicine. Hence an accurate, convenient, and automated CKD detection method is important for clinical practice.

In this paper we develop an automated machine leaning solution to detect CKD and explore 24 parameters related to kidney disease. The dataset used for evaluation consists of 400 individuals and suffers from noisy and missing data. We need a robust classifier that can deal with these issues. Hence, we evaluate solutions with three different classifiers: k-nearest neighbour, random forest and neural nets.

The main contributions of this paper are:

- Our solution, using a random forest classifier and 24 attributes, achieves a detection accuracy of 0.993 according to the F1-measure with a 0.1084 root mean square error. We show that this accuracy is significantly higher than current accepted GFR estimator equations; about 60% and 56% RMSE reduction compared to the Modification of Diet in Renal Disease (MDRD) equation [7] and the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation [8].
- Using a wrapper method from machine learning we identify a set of 12 attributes (down from 24 attributes) which detect CKD with high accuracy. Also, using the LASSO regularization method we rank the attributes according to their predictive capability in detecting CKD and further reduce the predictive attributes set to 10. By adding red blood cells, pus cell clumps, hemoglobin, diabetes mellitus, coronary artery disease, pedal edema, anemia as attributes with currently used serum creatinine and albumin, we achieve a 57% reduction in root mean square error compared to the state of the art solutions.
- We identify a highly accurate and cost effective CKD detection classifier considering only 5 attributes: specific gravity, albumin, diabetes mellitus, hypertension and hemoglobin as features. Using this classifier we have achieved 0.98 F1-measure and 0.11 RMSE with a total of $45.05 cost for patient tests.

978-1-5090-6117-4/16 $31.00 © 2016 IEEE
DOI 10.1109/ICHI.2016.36

IEEE computer society

II. Related Work

Chronic kidney disease (CKD) is defined by the presence of structural or functional abnormalities of the kidney with or without an accompanying reduction in glomerular filtration rate (GFR). Persons with CKD may have one or more of the following: pathologic abnormalities, markers of kidney damage (i.e., imaging abnormalities and abnormalities in serum or urine, including proteinuria and abnormal urinary sediment), or GFR less than 60 mL per minute per $1.73m^2$ for at least three months. Glomerular filtration rate (GFR) is one of the commonly used indexes for early detection of CKD. A five-stage classification system for the disorder has been established by the US National Kidney Foundation's Kidney Disease Outcomes Quality Initiative and adopted internationally by the Kidney Disease: Improving Global Outcomes (KDIGO) initiative to guide identification of cases and facilitate management [9], [10], [11], where glomerular filtration rate (GFR) is the estimator for CKD. Estimation of GFR varies by age, sex, and body size. GFR is approximately 120 to 130 mL per minute per 1.73 m^2 in young adults, and decreases by an average of 1 mL per minute per 1.73 m^2 per year after 30 years of age [12]. A GFR less than 60 mL per minute per 1.73 m^2 represents a loss of at least one-half of normal kidney function; below this level, there is an increased prevalence of CKD complications.

Earlier studies focused on plasma creatinine (Pcr) and creatinine clearance as markers of GFR, but Pcr usually does not increase until GFR has decreased by 50% or more, and many patients with normal Pcr levels frequently have lower GFR [13]. Creatinine clearance is also used to estimate the GFR. But, it overestimates true GFR [14] since creatinine is filtered and secreted by the proximal tubules. Generation of creatinine is determined by muscle mass and diet, whereas tubular secretion could be decreased by the use of medications such as trimethoprim and cimetidine (Tagamet). The serum creatinine level is an insensitive marker of GFR early in the course of CKD. A 33% decrease in GFR may raise the creatinine level from 0.8 to only 1.2 $mgperdL(70.72$ to $106.08molperL)$. If the prior creatinine level is not known, this decrease in GFR may go unrecognized. When estimated GFR is suspected to be inaccurate, for example, in patients with severe malnutrition or paraplegia-a 24-hour urine collection should be performed to evaluate creatinine clearance.

Currently, there are three equations commonly used to estimate GFR on the basis of creatinine concentration in serum and demographic features: the Cockcroft-Gault equation [15], the Modification of Diet in Renal Disease (MDRD) equation, [7] and the more accurate Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) [8] formula. Equations 1 and 2 show the MDRD and CKD-EPI equations expressed as single equations where S_{cr} is serum creatinine in mg/dL and α, k are constant values depending on the gender of the patient.

$$\begin{aligned} GFR = 175 \times (S_{cr})^{-1.154} \times (Age)^{-0.203} \\ \times (0.742 if female) \times (1.212 if African American) \end{aligned} \tag{1}$$

Equation	Attributes
Cronic Kidney Disease Epidemiology Collaboration	Age, sex, race, serum creatinine level
Cockcroft-Gault	Age, weight, sex, serum creatinine level
Modification of Diet in Renal Disease	Age, sex, race and serum urea, nitrogen, albumin, creatinine level

TABLE I

EQUATIONS AND ATTRIBUTES FOR GFR ESTIMATION

$$\begin{aligned} GFR = 141 \times min(S_{cr}/k, 1)^{\alpha} \times max(Scr/k, 1)^{-1.209} \\ \times 0.993^{Age} \times 1.018[iffemale] \times 1.159[ifblack] \end{aligned} \tag{2}$$

Table I shows the parameters used by these equations to estimate GFR. The Cockcroft-Gault equation uses age, weight, sex, and serum creatinine level for GFR estimation. MDRD and CKD-EPI equations do not require weight or height variables because the results are reported normalized to accepted average adult surface area.

The Cockcroft-Gault equation systematically overestimates GFR. The MDRD is reasonably accurate in patients with CKD, but it may misidentify persons with normal kidney function as having CKD. The MDRD can also be affected by fluctuations in creatinine production and fluid balance; it gives falsely elevated estimated GFRs in malnourished and overhydrated patients and falsely decreased GFRs due to increased serum creatinine levels in patients taking trimethoprim and cimetidine [16]. Also, its accuracy varies among ethnic groups [17]. Estimation accuracy of GFR using the MDRD equation study has achieved up to a root mean square error of 0.274 [8]. On the other hand, the CKD-EPI formula can estimate GFR with root mean square error of 0.250 [8], hence shows better performance particularly at high rates, and could overcome some of these limitations. Both the MDRD and CKD-EPI equations are based on serum creatinine. Despite modest reduction in bias with the CKD-EPI equation, estimates remain imprecise, with some people showing large differences between the measured and estimated GFR. Like all other creatinine-based estimation equations, they suffer from physiologic limitations of creatinine as a filtration marker [18].

Research presented in [19] has considered 5 attributes: blood pressure, serum creatinine, packed cell volume, hypertension, and anemia to calculate the L-factor and clustered CKD and non-CKD patients based on the L-factor value. According to their evaluation CKD cannot be detected based on their L-factor classifiers. Other works [20], [21] have evaluated machine learning algorithms such as back propagation neural networks, radial basis functions, random forests and SVMs and achieved up to 85.3% accuracy on identifying CKD. Also, [22] performs feature selection techniques such as information gain, gain ratio, or attribute evaluation and fusion based feature selection to identify relevant features, but their evaluation has not presented the relevant selected features. Moreover, this work presented classification accuracy of naive bayes, random forest, J48 classifier and logistic regression classifier without

Attribute	Data type
Age	age in years
Blood Pressure	mm/Hg
Specific Gravity	Nominal
Albumin	nominal(1-5)
Sugar	nominal (1-5)
Red Blood Cells	normal,abnormal
Pus Cell	normal,abnormal
Pus Cell clumps	present,notpresent
Bacteria	present,notpresent
Blood Glucose Random	mgs/dl
Blood Urea	mgs/dl
Serum Creatinine	mgs/dl
Sodium	mEq/L
Potassium	mEq/L
Hemoglobin	gms
Packed Cell Volume	nominal
White Blood Cell Count	cells/cumm
Red Blood Cell Count	millions/cmm
Hypertension	yes, no
Diabetes Mellitus	yes, no
Coronary Artery Disease	yes, no
Appetite	good,poor
Pedal Edema	yes,no
Anemia	yes,no

TABLE II
ATTRIBUTE INFORMATION

mentioning which attributes were used as features for these classifiers.

Hence, the goals of our study are to comprehensively explore parameters which are related to kidney disease and to introduce a cost effective machine learning approach to detect early CKD instead of the GFR estimation equations.

III. DATA SET AND ATTRIBUTES

Our research uses a publicly available dataset [23]: Early stage of Chronic Kidney Disease. This dataset includes 400 patients with 24 attributes collected from each of these patients; 250 of them have CKD. The ages of these patients vary from 2 to 90 with mean of 51.48 and a standard daviation of 17.17. Most of the 24 collected attributes shown in Table II have not been used by previous state of art approaches for CKD detection, e.g., most approaches use only age, serum creatinine, albumin, and urea.

These attributes along with their relation with kidney diseases are described below:

- *Serum creatinine* is a waste product that comes from muscle activity. When kidneys are working well they remove creatinine from the blood. As kidney function slows, blood levels of creatinine rise. According to the studies [7] [8], serum creatinine, *age*, *serum urea*, and *specific gravity* are the most used predictive parameters for CKD detection.
- Studies have shown graded relations between increased *albuminuria* (the presence of albumin in the urine) and kidney outcomes in diverse study populations [24]. Also, data from the general U.S. population indicate that albuminuria is the most typical marker of CKD in young adults [25].
- *High blood pressure* can damage blood vessels in the kidneys, reducing their ability to work properly. When the force of blood flow is high, blood vessels stretch so blood flows more easily. Eventually, this stretching scars and weakens blood vessels throughout the body, including those in the kidneys. If the kidneys' blood vessels are damaged, they may stop removing wastes and extra fluid from the body. Extra fluid in the blood vessels may then raise blood pressure even more, creating a dangerous cycle [26].
- CKD is an independent risk factor for *coronary artery disease (CAD)*. It is the leading cause of morbidity and mortality in patients with CKD [27].
- Study [28] shows that 70% of those with an elevated serum creatinine had *hypertension*. Hence, high blood pressure, CAD and hypertension are good predictive attributes for CKD.
- *Anemia* is a condition in which the body has fewer red blood cells than normal. Red blood cells carry oxygen to tissues and organs throughout the body and enable them to use energy from food. With anemia red blood cells carry less oxygen to tissues and organs, particularly the heart and brain. Anemia commonly occurs in people with CKD having permanent or partial loss of kidney function. Anemia might begin to develop in the early stages of CKD, when someone has 20 to 50 percent of normal kidney function [29]. Anemia is a predictive factor for early renal disease. *Hemoglobin*, *red blood cell count*, *packed cell volume* in the patients blood are used to detect early stage of anemia.
- According to National Kidney Foundation [30] about a third of people with *diabetes* may get CKD. The filtering units of the kidney are filled with tiny blood vessels. If a person has diabetes, high sugar levels in the blood can cause these vessels to become narrow and clogged. Without enough blood, the kidneys become damaged and *albumin* passes through these filters and ends up in the urine where it should not be. Diabetes causes nerve damage which make patient unable to detect if his or her bladder is full. The pressure from a full bladder can cause damage to the kidney. *Blood glucose* is used to screen for diabetes. Hence, diabetes, blood glucose and albumin in urine are good indicators for CKD.
- Though *sodium* and *potassium* are essential for the human body, a person with CKD cannot eliminate excess sodium, potassium and fluid from his body. Eventually sodium, potassium, and fluid buildup in tissues and bloodstream. High sodium increases blood pressure [31]. High potassium in the blood is called hyperkalemia, which may occur in people with advanced CKD. Some of the effects of high potassium are nausea, weakness, numbness and slow pulse. Both sodium and potassium are predictor attributes for CKD.
- *Edema* is the medical term for swelling. Edema results whenever small blood vessels become 'leaky' and release fluid into nearby tissues. The extra fluid accumulates,

causing the tissue to swell. A kidney condition called nephrotic syndrome can result in severe *pedal edema*.

- In the developing world, infectious diseases are also important causes of kidney failure [32] , including infections due to *bacteria* (tuberculosis in India and the Middle East, streptococcal infection in Africa), *viruses* (HIV and hepatitis B and C in Africa), and *parasites* (schistosomiasis in Africa and Latin America, leishmaniasis in Africa and Asia, and malaria in Africa). *Pus cells* in urine indicates infection in the kidney.

IV. Classification Task

The task of classifying data is to decide class membership Y of an unknown data item X based on a data set $D = (x_1, y_1), ...(x_n, y_n)$ of data items x_i with known class memberships y_i. In binary class classification problems the class labels y are either 0 or 1. The x_i are usually m-dimensional vectors, the components of which are called covariates and independent variables or input variables. The relationship between x and y is described by a probability distribution $P(x, y)$; where the data set D contains independent samples from P. From statistical decision theory, it is well known that the optimal class membership decision is to choose the class label y that maximizes the posterior distribution $P(y|x)$. In this research we explore the following 3 different classification algorithms to predict optimal class membership (CKD or not CKD). In this section, as background, we briefly describe and compare these classifiers and discuss their applicability for our dataset.

A. K-Nearest Neighbours

The k-nearest neighbour algorithm [33] uses the data directly for classification without building a model first. As such, no details of model construction need to be considered, and the only adjustable parameter in the model is k, the number of nearest neighbours to include in the estimate of class membership: the value of $P(y|x)$ is calculated simply as the ratio of members of class y among the k-nearest neighbors of x. By varying k, the model can be made more or less flexible. The advantage of the k-nearest neighbours classifier is, it is robust to noisy training data and effective with large training datasets. The major drawback lies in the calculation of the case neighborhood: for this, one needs to define a metric that measures the distance between data items. In most cases it is done by trial and error.

B. Random Forest

The random forest [34] is an ensemble approach that can also be thought of as a form of nearest neighbour predictor. Ensembles [35] are a divide-and-conquer approach used to improve performance. The main principle behind ensemble methods is that a group of ‘weak learners’ can come together to form a ‘strong learner’. The random forest starts with a standard machine learning technique called a ‘decision tree’ which, in ensemble terms, corresponds to our weak learner. The decision tree algorithm repeatedly splits the data set according to a criterion that maximizes the separation of the data, resulting in a tree-like structure. In this algorithm an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. The random forest takes this notion to the next level by combining trees with the notion of an ensemble. Thus, in ensemble terms, the trees are weak learners and the random forest is a strong learner. The advantages of a random forest classifier are that its’ runtimes are quite fast, and that it is able to deal with unbalanced and missing data. Weaknesses of this algorithm are that when used for regression it cannot predict beyond the range in the training data, and it may over-fit data sets that are particularly noisy.

C. Neural Network

A neural network [36] is a powerful computational data model that is able to capture and represent complex input/output relationships. The motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform ‘intelligent’ tasks similar to those performed by the human brain. This model differs from the two algorithms above in the sense that it provides a functional form f and parameter vector α to express $P(y|x)$ as $P(y|x) = f(x, \alpha)$. The parameters α are determined based on the data set D, usually by maximum-likelihood estimation. The true power and advantage of neural networks lie in their ability to represent both linear and non-linear relationships and in their ability to learn these relationships directly from the data being modeled. Traditional linear models are simply inadequate when it comes to modeling data that contains non-linear characteristics. The most common neural network model is the multilayer perceptron (MLP). This type of neural network is known as a supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown.

D. Applicability for the Dataset

Our considered dataset [23] with 24 attributes suffers from missing and noisy value. Hence, we need a robust and fast non-linear classifier which can handle both noisy and missing attribute values. All three classifiers in this study create a non-linear decision boundary which is necessary for complex applications like this one. The k-nearest neighbours classifier has the advantage of performing with a small training set and it can adopt new training data at runtime. It has a disadvantage that it is difficult to find an optimal value of k that produces the best performance for a training set with a finite number of training samples. Specially value of k may change with inclusion of new training data. On the other hand, the training of a neural network typically requires a large amount of data in the training set. Both random forest and neural network classifiers have faster speeds of classification compared to k-nearest neighbours. Both k-nearest neighbours and random forest have good adaptivity with missing and noisy data, though they may over-fit particularly on noisy datasets. Feature reduction techniques are used to reduce over-fitting for these classifiers. Although, neural networks are fairly resistant to noise, they are not adaptive to missing data. In these cases, missing values are replaced with pseudo values or instances with missing data are ignored in evaluation.

Since, we want to have a robust and fast classifier which can also handle noisy and missing data, we have evaluated all three of these classifiers on our dataset with both feature reduction (section V-A) and missing value handling approaches to determine the best classification algorithm for this application.

V. Attribute Selection task

Attribute selection is the automatic selection of attributes in data that are most relevant to the predictive modeling problem. It is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method does so by creating new combinations of attributes, where as attribute selection methods include and exclude attributes present in the data without changing them. Attribute selection methods are used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. There are two general classes of attribute selection approaches: the wrapper and embedded approaches. We have evaluated one algorithm from each of these classes on our dataset to determine the most relevant attributes for CKD detection and remove the irrelevant features for cost reduction of the detection approach.

A. Wrapper Approach

This paper performs the wrapper approach to identify the best subset of the 24 attributes, which can be used as features to detect CKD with high accuracy. In the wrapper approach the attribute subset selection is done using the induction algorithm as a black box (i.e. no knowledge of the algorithm is needed, just the interface). The feature subset selection algorithm conducts a search for a good subset using the induction algorithm itself as part of the evaluation function. The accuracy of the induced classifiers is estimated using accuracy estimation techniques. Hence, it is a state space search problem. The wrapper approach conducts a search in the space of possible parameters. In this research we have used 'best first search' as search method due to its robustness. The idea is to select the most promising set we have generated so far that has not already been expanded. Best first search usually terminates upon reaching the goal. Since it is an optimization problem, the search can be stopped at any point and the best solution found so far can be returned. In practice a stale search method is used, where search is terminated if no improved set is found in the last k expansions. An improved node is defined as a node with an accuracy estimation at least ϵ higher than the best one found so far. In the following experiments, k is five and epsilon is 0.1%.

B. Embedded Approach

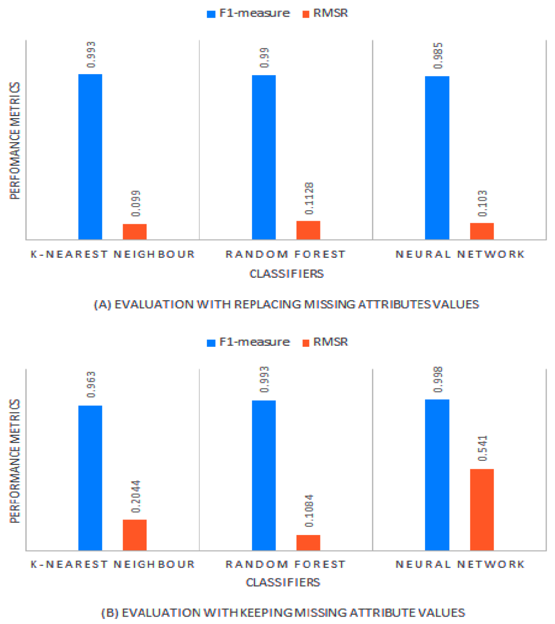
This paper performs the embedded approach to identify and rank the attributes with higher predictability of CKD detection and eliminate unneeded, irrelevant and redundant attributes from consideration. Embedded approaches learn which features best contribute to the accuracy of the detection model during model creation. The most common type of embedded feature selection approaches are regularization methods. Regularization refers to the process of adding additional constraints to a problem that bias the model toward lower complexity. In this research we have used LASSO (Least Absolute Shrinkage and Selection Operator) which is a modified form of least squares regression that penalizes model complexity via a regularization parameter. It achieves better prediction accuracy by shrinkage with ridge regression, but at the same time, it gives a sparse solution, which means that some coefficients are exactly 0. Hence, LASSO is thought to achieve the shrinkage and variable selection simultaneously. LASSO minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. LASSO not only helps to improve the prediction accuracy when dealing with multicolinearity data, but also carries several nice properties such as interpretability and numerical stability. It is a simple non linear dimensionality reduction technique which has efficient solution via coordinate descent with order O(np) where, n is the number of instances in the dataset and p is number of attributes. Also, one of the major advantages of using LASSO for attribute selection is that correlations in predictor attributes are not problematic for LASSO [37].

LASSO regularization cannot handle categorical attributes, which are present in our dataset. Hence, we have converted each categorical attribute to $k-1$ dummy attributes, where k is the number of categories present in that variable. We have performed group LASSO regularization [38], where all the dummy attributes created from one categorical attribute are grouped for attribute reduction.

VI. Evaluation

The evaluation is divided into four sets. In subsection VI-A we present the results of the evaluation on the dataset [23] for detecting CKD using all 24 features with three different classifiers: k-nearest neighbours, random forest, and neural networks. Since, previous state of the art works [8] have used the root mean square error (RMSE) to estimate detection accuracy, this paper considers accuracy, F1-measure, and the root mean square error (RMSE) as performance metrics to compare the classifiers. All the evaluations were done using 10-fold cross validation with 20% of the data as test data.

Since it is not always practical to use 24 features, using the wrapper approach we find the best subset of these 24 attributes that provide excellent accuracy and report these results in subsection VI-B. In section VI-C we rank and identify the attributes with predictability of CKD using LASSO regularization. Finally, in section VI-D we perform a cost analysis to identify a attribute set to detect CKD with high accuracy at a low cost.



(A) EVALUATION WITH REPLACING MISSING ATTRIBUTES VALUES

(B) EVALUATION WITH KEEPING MISSING ATTRIBUTE VALUES

Fig. 1. Detection performance of classifiers with and without replacing missing values.

A. *Predicting CKD with 24 Attributes*

To run the classifiers on the dataset we must first address missing values. We have evaluated two approaches to handle these missing values: replacing all missing values for nominal and numeric attributes in our dataset with the modes and means or medians from the training data. For the k-nearest neighbours approach, we have used the IBk algorithm considering 2 nearest neighbours. $IBk's$ distance computation method assigns maximum distance when there is a missing value encountered in one of the instances. In our random forest algorithm we have used C4.5 [39] trees. In C4.5 the missing values are not replaced in the dataset. Instead, an impurity function computed takes into account the missing values by penalizing the impurity score with the ratio of missing values. On test set the evaluation in a node which has a test with missing values, the prediction is built for each child node and aggregated later (by weighting). The neural network algorithm ignores the missing values during classifier training. Figure 1 (A), (B) shows detection performance of classifiers with and without replacing missing values. In this evaluation we have considered all 24 features as input for the classifier.

Figure 1 shows that, detection accuracy for the k-nearest neighbours approach (IBk algorithm) decreases significantly during training the classifier with missing attribute values. In the training phase, IBk penalizes instances with missing attribute values, which biases the classifier. Both the neural network and random forest algorithm perform better when trained with missing attribute values. The random forest algorithms' (C4.5 tree) impurity function computation adopts the missing values better compared to neural networks ignoring the missing value strategy. Hence, we achieved a highest detection accuracy of 0.993 according to the F1-measure with a 0.1084 root mean square error (RMSE) using the random forest classifier with 100 $C4.5$ trees trained with missing attribute values. This is a 56% RMSE reduction compared to the state of the art solution (the CKD-EPI formula).

Attributes of the best predictive subset
Specific Gravity
Albumin
Red Blood Cells
Pus Cell clumps
Serum Creatinine
Sodium
Hemoglobin
Diabetes Mellitus
Coronary Artery Disease
Appetite
Pedal Edema
Anemia

TABLE III
ATTRIBUTES OF THE BEST PREDICTIVE SUBSET USING WRAPPER APPROACH

B. *Best Subset of Attributes*

We use the wrapper approach to identify the best subset of the 24 attributes, that can detect CKD with high accuracy. In this approach we have used random forest as induction algorithm and the 'best first search' as the search method and 'stale search' after 5 node expansions. Table III shows the best predictive subset of the 24 attributes selected by the wrapper approach.

Using a random forest classifier with these 12 predictive attributes as input we achieve a .99 F1-measure, 99% precision and 0.107 root mean square error (RMSE). Using this approach we have achieved 57% and 61% RMSE reduction compared to the CKD-EPI and MDRD formulas for GFR estimation, respectively. Given the high accuracy achieved, these results imply that only these 12 features are necessary.

Compared to previous approaches shown in Table I, instead of sex, age and weight, we see that specific gravity works as a good predictive attribute. All of the previous equations have considered 'serum creatinine' in their equations, additionally MDRD has considered urea, nitrogen and albumin. Our analysis has also identified serum creatinine and albumin as predictive attributes for CKD. Additionally red blood cells, hemoglobin, diabetes, coronary artery diseases, sodium, pus cell clumps and pedal edema are identified as good predictive attributes for CKD which were not considered in any of the previous approaches.

C. *Assessing Impact of Each Feature*

LASSO penalizes regression models with L1 norms that have sparse solutions: many of their estimated coefficients are zero. Higher coefficients values indicate higher predictive capability for a feature and if the value is zero, we eliminate that attribute. Figure 3 shows the importance of attributes. Using the random forest classier with these 10 predictive attributes as input we achieve a 0.99 F1-measure and a 0.111

Attribute	Pearson's correlation
Hemoglobin	0.729
Packed Cell Volume	0.69
Red Blood Cell Count	0.591
Hypertension	0.5904
Diabetes Mellitus	0.5591
Albumin	0.477
Blood Glucose Random	0.4014
Appetite	0.3933
Pus Cell	0.3752
Pedal Edema	0.372
Specific Gravity	0.372
Blood Urea	0.35
Sodium	0.343
Anemia	0.3254
Sugar	0.3
Serum Creatinine	0.2941
Blood Pressure	0.2906
Red Blood Cells	0.2826
Pus Cell clumps	0.2653
Coronary Artery Disease	0.2361
Age	0.2254
White Blood Cell Count	0.2053
Bacteria	0.1869
Potassium	0.0769

TABLE IV
PEARSON'S CORRELATION BETWEEN ATTRIBUTES AND THE CLASS

root mean square error (RMSE) which is 56% and 60% RMSE reduction compared to the CKD-EPI and MDRD formulas for GFR estimation, respectively.

One limitation of L1-based sparse models is that faced with a group of very correlated features they do not select all of those features; which limits their ability to achieve optimal accuracy. Table IV shows the Pearson's correlation between the attributes and CKD, where higher value means high correlation. LASSO regularization selects most of the highly correlated attributes except blood glucose random, pus cell, blood urea and pedal Edema. Blood glucose random and diabetes mellitus have Pearson's correlation of 0.526; pus cell has 0.542 and 0.548 Pearson's correlation with hemoglobin and red blood cell count respectively; blood urea has 0.62 and 0.58 Pearson's correlation with hemoglobin and red blood cell count respectively; pedal edema has 0.455, 0.454, 0.43 and 0.42 Pearson's correlation with red blood cell count, packed cell volume, albumin and hemoglobin. Since, blood glucose random, pus cell, blood urea and pedal Edema have high correlation with attributes which have higher correlation with CKD compare to them, LASSO regularization has not included these attributes to predict CKD. Compared to previous approaches shown in Table I, diabetes mellitus, hypertension, hemoglobin, red blood cell count are good predictive attributes for CKD which were not considered in any of the previous approaches.

D. Cost analysis

This section presents a cost-accuracy trade-off analysis considering the 24 attributes used detect CKD. Table V shows the test names and approximate lowest test costs [40]–[49] for the 24 attributes. Figure 2 list these attributes on the x-axis in order of predictive power. The y-axis displays the RMSE as we incrementally add attributes. The curve is labeled with the cumulative cost in dollars of using these attributes. For example, according to figure 2 using all 24 attributes to detect CKD will cost $451.36 and the accuracy is 0.107 RMSE; using the top 20 predictive attributes (i.e., all the tests up to and including blood urea) has a cost of $294.72. Importantly, considering only the top 5 predictive attributes: specific gravity, albumin, diabetes mellitus, hypertension and hemoglobin as features for our classifier, we achieve .98 F1-measure and 0.11 RMSE (essentially the same accuracy with all the attributes), but with *only* a $45.05 cost. This is a very important result because patients need only be subjected to a few tests at very low cost.

Attribute	Name of the test	Lowest Cost (USD)
Blood Pressure	Blood Pressure Test	Free
Specific Gravity		Free
Albumin	Serum Albumin Test	25
Sugar	Fasting Blood Sugar Test	20
Red Blood Cells	RBC Count, CBC Test	39
Pus Cell	Urinalysis	30
Pus Cell clumps	Urinalysis	30
Bacteria	Blood Culture	50
Blood Glucose Random	Random Blood Glucose Test	20
Blood Urea	Blood Urea Nitrogen Test	11.85
Serum Creatinine	Serum Creatinine Test	14
Sodium	Serum Sodium Test or Sodium Urine Test	3.2
Potassium	Potassium lab test	49
Hemoglobin	HGB1	1.65
Packed Cell Volume	Hematocrit Test	1.62
White Blood Cell Count	Complete Blood Count Test	30
Red Blood Cell Count	Complete Blood Count Test	30
Hypertension		Free
Diabetes Mellitus	Diabetes Assessment	18.4
Coronary Artery Disease	Electrocardiogram	50
Appetite		Free
Pedal Edema		Free
Anemia	Anemia Assessment	27.64

TABLE V
TESTS AND TEST COSTS OF ALL ATTRIBUTES

VII. CONCLUSION

We have introduced a novel approach to detect CKD using machine learning techniques. We have performed an evaluation on a dataset of 400 patients, 250 among them have early stage of CKD. This dataset contains some noisy and missing values. Hence, we need a classification algorithm with the capability of handling missing and noisy values. We evaluated three classifiers: k-nearest neighbours, random forest, and neural networks to find a good solution for this application. To reduce over-fitting as well as to identify the most important predictive attributes for CKD, we have performed feature reduction using

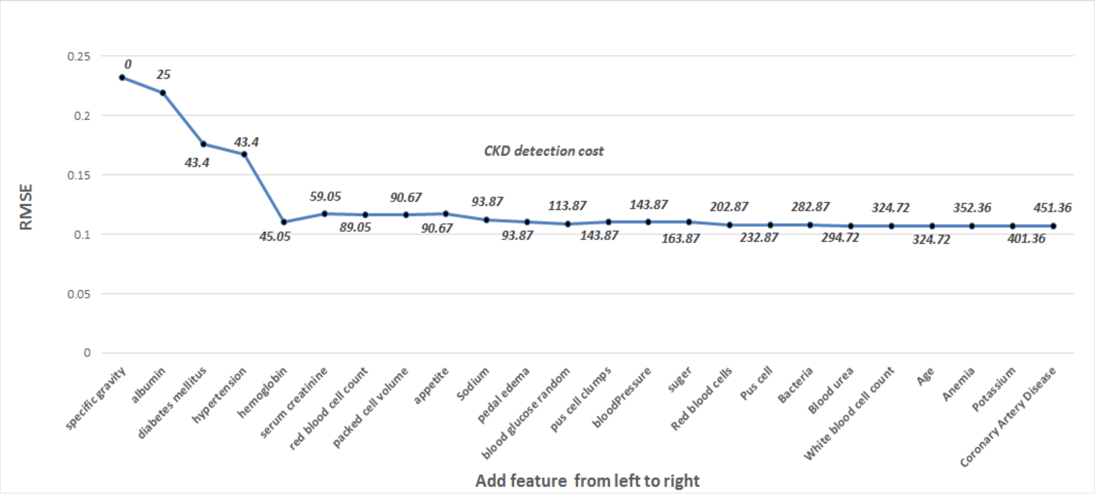


Fig. 2. Change of RMSE and cost of CKD detection with increase of predictive attributes used for classifier

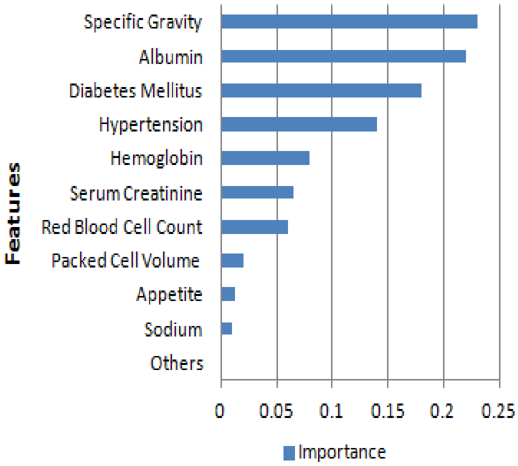


Fig. 3. Importance of attributes using lasso.

two methods: the wrapper method and LASSO regularization. Through our evaluation we find that, the random forest algorithm with a reduced attribute set of 12 members can detect CKD with highest accuracy of .998 using the F1-measure and with a 0.107 root mean square error, which is a 57% RMSE reduction compared to the state of the art solutions. Through our evaluation we find hemoglobin which is an indicator of anemia, diabetes mellitus, specific gravity, hypertension etc. along with previously explored serum creatinine, and albumin are highly predictive attributes for CKD. Also, through cost analysis considering all 24 attributes we identify a cost effective highly accurate detection classifier using only 5 attributes: specific gravity, albumin, diabetes mellitus, hypertension and hemoglobin. Importantly, results of this study introduce new factors to be used by classifiers for more accurately detecting CKD than the state of art using formulas.

VIII. Acknowledgement

This work was supported, in part, by DGIST via funding from the Ministry of Education, Science and Technology of the Rerepublic of Korea.

References

[1] Q.-L. Zhang and D. Rothenbacher, “Prevalence of chronic kidney disease in population-based studies: systematic review,” *BMC public health*, vol. 8, no. 1, p. 117, 2008.

[2] M. Baumgarten and T. Gehr, “Chronic kidney disease: detection and evaluation,” *American family physician*, vol. 84, no. 10, p. 1138, 2011.

[3] V. A. Moyer, “Screening for chronic kidney disease: Us preventive services task force recommendation statement,” *Annals of internal medicine*, vol. 157, no. 8, pp. 567–570, 2012.

[4] G. Remuzzi, P. Ruggenenti, and N. Perico, “Chronic renal diseases: renoprotective benefits of renin–angiotensin system inhibition,” *Annals of internal medicine*, vol. 136, no. 8, pp. 604–615, 2002.

[5] L. C. Plantinga, L. E. Boulware, J. Coresh, L. A. Stevens, E. R. Miller, R. Saran, K. L. Messer, A. S. Levey, and N. R. Powe, “Patient awareness of chronic kidney disease: trends and predictors,” *Archives of internal medicine*, vol. 168, no. 20, pp. 2268–2275, 2008.

[6] L. C. Plantinga, D. S. Tuot, and N. R. Powe, “Awareness of chronic kidney disease among patients and providers,” *Advances in chronic kidney disease*, vol. 17, no. 3, pp. 225–236, 2010.

[7] A. S. Levey, J. P. Bosch, J. B. Lewis, T. Greene, N. Rogers, and D. Roth, “A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation,” *Annals of internal medicine*, vol. 130, no. 6, pp. 461–470, 1999.

[8] A. S. Levey, L. A. Stevens, C. H. Schmid, Y. L. Zhang, A. F. Castro, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene *et al.*, “A new equation to estimate glomerular filtration rate,” *Annals of internal medicine*, vol. 150, no. 9, pp. 604–612, 2009.

[9] A. S. Levey, J. Coresh, E. Balk, A. T. Kausz, A. Levin, M. W. Steffes, R. J. Hogg, R. D. Perrone, J. Lau, and G. Eknoyan, “National kidney foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification,” *Annals of internal medicine*, vol. 139, no. 2, pp. 137–147, 2003.

[10] S. Moe, T. Drueke, J. Cunningham, W. Goodman, K. Martin, K. Olgaard, S. Ott, S. Sprague, N. Lameire, and G. Eknoyan, “Kidney disease: Improving global outcomes (kdigo). definition and classification of chronic kidney disease: a position statement from kidney disease: Improving global outcomes (kdigo),” *Kidney Int*, vol. 67, pp. 2089–2100, 2005.

[11] A. Levin, B. Hemmelgarn, B. Culleton, S. Tobe, P. McFarlane, M. Ruzicka, K. Burns, B. Manns, C. White, F. Madore *et al.*, "Guidelines for the management of chronic kidney disease," *Canadian Medical Association Journal*, vol. 179, no. 11, pp. 1154–1162, 2008.

[12] J. D. Kopple, "National kidney foundation k/doqi clinical practice guidelines for nutrition in chronic renal failure," *American journal of kidney diseases*, vol. 37, no. 1, pp. S66–S70, 2001.

[13] R. D. Perrone, N. E. Madias, and A. S. Levey, "Serum creatinine as an index of renal function: new insights into old concepts." *Clinical chemistry*, vol. 38, no. 10, pp. 1933–1953, 1992.

[14] S. Giovannetti and G. Barsotti, "In defense of creatinine clearance," *Nephron*, vol. 59, no. 1, pp. 11–14, 1991.

[15] D. W. Cockcroft and M. H. Gault, "Prediction of creatinine clearance from serum creatinine," *Nephron*, vol. 16, no. 1, pp. 31–41, 1976.

[16] L. Stevens and A. Levey, "National kidney foundation. frequently asked questions about gfr estimates," 2008.

[17] A. Levey, J. Bosch, J. B. Lewis, T. Greene, N. Rogers, D. Roth *et al.*, "Modification of diet in renal disease study group: A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation," *Ann Intern Med*, vol. 130, no. 6, pp. 461–470, 1999.

[18] O. Shemesh, H. Golbetz, J. P. KRIss, and B. D. Myers, "Limitations of creatinine as a filtration marker in glomerulopathic patients," *Kidney Int*, vol. 28, no. 5, pp. 830–838, 1985.

[19] A. Dubey, "A classification of ckd cases using multivariate k-means clustering." *International Journal of Scientific and Research Publications (IJSRP)*, vol. 5, August 2015.

[20] P. Sinha and P. Sinha, "Comparative study of chronic kidney disease prediction using knn and svm," *International Journal of Engineering Research and Technology*, vol. 4, no. 12, 2015.

[21] S. Ramya and N. Radha, "Diagnosis of chronic kidney disease using machine learning algorithms," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 1, 2016.

[22] B. Mohammed Siyad and M. Manoj, "Fused features classification for the effective prediction of chronic kidney disease," *International Journal for Innovative Research in Science and Technology*, vol. 2, no. 10, 2016.

[23] L. Rubini, "Early stage of chronic kidney disease UCI machine learning repository," 2015. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

[24] C. K. D. P. Consortium *et al.*, "Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis," *The Lancet*, vol. 375, no. 9731, pp. 2073–2081, 2010.

[25] R. T. Gansevoort, K. Matsushita, M. van der Velde, B. C. Astor, M. Woodward, A. S. Levey, P. E. de Jong, J. Coresh, M. El-Nahas, K.-U. Eckardt *et al.*, "Lower estimated gfr and higher albuminuria are associated with adverse kidney outcomes. a collaborative meta-analysis of general and high-risk population cohorts," *Kidney international*, vol. 80, no. 1, pp. 93–104, 2011.

[26] "National institution of diabetes and digestive an kidney disease," http://1.usa.gov/1Y5WeWt, accessed: 2015-12-11.

[27] Q. Cai, V. K. Mukku, and M. Ahmad, "Coronary artery disease in patients with chronic kidney disease: a clinical update," *Current cardiology reviews*, vol. 9, no. 4, p. 331, 2013.

[28] J. Coresh, G. L. Wei, G. McQuillan, F. L. Brancati, A. S. Levey, C. Jones, and M. J. Klag, "Prevalence of high blood pressure and elevated serum creatinine level in the united states: findings from the third national health and nutrition examination survey (1988-1994)," *Archives of internal medicine*, vol. 161, no. 9, pp. 1207–1216, 2001.

[29] Y. Yang, B. Yu, and Y. Chen, "Blood disorders typically associated with renal transplantation," *Frontiers in cell and developmental biology*, vol. 3, 2015.

[30] "Diabetes - a major risk factor for kidney disease," https://www.kidney.org/atoz/content/diabetes, accessed: 2015-12-11.

[31] "Sodium and chronic kidney disease," goo.gl/SPmYkG, accessed: 2015-12-11.

[32] T. Jafar, M. Islam, N. Poulter *et al.*, "Chronic kidney disease in the developing world," 2006.

[33] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[34] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[35] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple classifier systems*. Springer, 2000, pp. 1–15.

[36] M. T. Hagan, H. B. Demuth, M. H. Beale *et al.*, *Neural network design*. Pws Pub. Boston, 1996.

[37] M. Hebiri and J. Lederer, "How correlations influence lasso prediction," *Information Theory, IEEE Transactions on*, vol. 59, no. 3, pp. 1846–1854, 2013.

[38] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[39] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[40] "Ekgs and exercise stress tests," http://www.choosingwisely.org/patient-resources/ekgs-and-exercise-stress-tests/, accessed: 2016-3-11.

[41] "Walgreens introduces daily testing for cholesterol, blood glucose and a1c at more than 1,400 stores in 33 states and washington, d.c." http://news.walgreens.com/press-releases/community-news/walgreens-introduces-daily-testing-for-cholesterol-blood-glucose-and-a1c-at-more-than-1400-stores-in-33-states-and-washington-dc.htm, accessed: 2016-3-11.

[42] "Microalbumin/creatinine ratio random urine test," http://www.walkinlab.com/diabetes-tests/microalbumincreatinineratiorandomurinetest.html, accessed: 2016-3-11.

[43] "Walk-in-lab," http://www.walkinlab.com/kidney-tests/albumin.html, accessed: 2016-3-11.

[44] "Health testing centers: Anemia," http://goo.gl/Xxj43u, accessed: 2016-3-11.

[45] "Healthone: Urinalysis, complete," http://goo.gl/kBXvaR, accessed: 2016-3-11.

[46] B. Perl, N. P. Gottehrer, D. Raveh, Y. Schlesinger, B. Rudensky, and A. M. Yinnon, "Cost-effectiveness of blood cultures for adult patients with cellulitis," *Clinical infectious diseases*, vol. 29, no. 6, pp. 1483–1488, 1999.

[47] "What does a bun test cost?" https://goo.gl/GL6ZbL, accessed: 2016-3-11.

[48] "Theranos," https://www.theranos.com/test-menu, accessed: 2016-3-11.

[49] "The new york times health guide for swelling," http://www.nytimes.com/health/guides/symptoms/swelling/overview.html, accessed: 2016-3-11.