This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2017.2787480, IEEE Reviews in Biomedical Engineering

Identifying important attributes for early detection of Chronic Kidney Disease

Anandanadarajah Nishanth, and Tharmarajah Thiruvaran, *Member, IEEE*

Abstract Those who are with Chronic Kidney Disease (CKD) are not aware that the medical tests they take for other purposes sometimes contain useful information about CKD disease. This information is sometimes not used effectively to tackle the identification of the disease. Therefore, attributes of different medical tests are investigated to identify what attributes contain useful information about CKD. A database with several attributes of healthy subjects and subjects with CKD are analyzed with different techniques. Common Spatial Pattern (CSP) filter and Linear Discriminant Analysis (LDA) are first used to identify the dominant attributes that could contribute in detecting CKD. Here CSP filter is applied to optimize separation between CKD and non-CKD. Then, classification methods are also used to identify the dominant attributes. These analyses suggest that hemoglobin, albumin, specific gravity, hypertension and diabetes mellitus together with serum creatinine are the most important attributes in the early detection of CKD. Further, it suggests that in the absence of the information of hypertension and diabetes mellitus, the attributes blood glucose random, and blood pressure may be used.

***Index Terms*— Chronic Kidney Disease (CKD), Common Spatial Pattern (CSP) filter, Linear Discriminant Analysis (LDA), Estimated GFR (eGFR) and Serum Creatinine.**

I. Introduction[1]

CHRONIC kidney disease (CKD, also called chronic renal disease) is a condition in which kidneys gradually lose their function. This could cause problem to waste and excess fluids accumulation in the body and affects the functionality of the body, potentially leading to complications. The disease can progress to end-stage renal disease (complete kidney failure). This occurs when kidney function got worsened to a point where dialysis or kidney transplantation is required for survival. People with CKD also have an increased risk of developing Cardiovascular Diseases (CVD) [1]-[2]. Further, the CVD in CKD population is different from those who are not affected by CKD (non-CKD) [1]. That is, CVD is the leading cause of death in individuals who are on dialysis [1] that implies renal disease causes death in a different way.

However, sometimes a person who is affected by early CKD may not feel unwell or notice any prodrome. To identify CKD, specific urine and blood tests should be taken [3]. So identifying CKD at the early stage is not easy without proper tests. On the other hand several attributes of medical tests taken for other purposes contain useful information of CKD. To effectively use these attributes the importance of these attributes of CKD should be studied in detail. There are many studies done to find the risk factors of CKD. In [4] the authors summarized the risk factors as obesity, hypertension, diabetes mellitus, cigarette smoking, established cardiovascular disease, age being greater than 60 years, aboriginal and Torres Strait Islander peoples, Maori and Pacific peoples, family history of stage 5 CKD or hereditary kidney disease in a first or second degree relative, and severe socioeconomic disadvantage.

In order to reduce the chances of CKD leading to dialysis or kidney transplantation**,** early detection of CKD is important. If a person is suspected to have CKD then kidney imaging may be used to confirm the disease. Since the population is high kidney imaging cannot be used by everyone and only those who have high probability to have CKD may be recommended. Thus there should be a way to get a clue of CKD initially. Then that could prompt further tests to confirm CKD. In [5] it is suggested that if CKD is detected earlier then even combined specialized nephrology nurse and primary care clinicians can provide special attention on screening, monitoring and advising on changing the patient's life style to prevent or reduce the development of CKD.

In order to detect CKD, Glomerular Filtration Rate (GFR) was earlier calculated from Cockcroft-Gault formula (developed in 1973) to check the condition of the renal function [6]. This formula especially considers the widely used index, serum creatinine concentration, to measure the renal function [7]. However, serum creatinine concentration does not rely only on GFR but other factors also affect [7]. Then, the equation is modified as Estimated GFR (eGFR) with Medication of Diet in Renal Disease (MDRD). This uses serum creatinine, age, ethnicity, gender, blood urea nitrogen and albumin [7]. Furthermore, Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation

[1] 23-January-2016

A.Nishanth, was with Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Jaffna, Kilinochchi, Sri Lanka. He is now with the Department of Computer Science, University of Louisiana at Lafayette (e-mail: eng_nishanth@eng.jfn.ac.lk)

T. Thiruvaran, was with University of New South Wales, Australia. He is now with the Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Jaffna, Kilinochchi, Sri Lanka. (e-mail: thiruvaran@eng.jfn.ac.lk)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2017.2787480, IEEE Reviews in Biomedical Engineering

was developed in 2009 to estimate GFR from serum creatinine, age, gender, and race then it is updated in 2012, based on cross sectional analysis of 13 studies [8]. The inclusion of age in [8] was motivated by the continuous population studies in [9-14]. The inclusion of gender in [8] was motivated by the meta-analysis in [14-15]. However, the analysis of the importance of age and gender based on eGFR using serum creatinine is, with an implied assumption that serum creatinine is the major factor detecting CKD. KDIGO [16] provides state-of-the-are guidance on evaluation, management and treatment of subjects with CKD. It recommends the following as the markers for kidney damage: Albuminuria, urine sediment abnormalities, electrolyte and other abnormalities due to tubular disorders, abnormalities detected by histology, structural abnormalities detected by imaging and history of kidney transplantation. It further guides the monitoring and the progression of CKD and the complications created by CKD.

Another study suggests neural network technique to identify CKD [17]. The difficulty in this method is, if a certain set of attributes are used to train the neural network then every subject should have all those attributes to test irrespective of their relative importance. The proposed analysis in this paper aims to identify the important attributes that contribute to CKD. So that when people take medical test for some other purposes, they can use those attributes from those tests to get a clue of CKD. Then they may proceed to take proper test to confirm CKD. Further our analysis can provide the information about what attributes should be used for the classification of CKD and non-CKD.

In addition, this study may motivate any improvement of eGFR equation by including other attributes for the detection of CKD. Thus the purpose of this study is to identify the dominant attributes that could be used to improve the specificity and classification accuracy of CKD and non-CKD.

Human body organs are interconnected with each other, so if one organ does not work properly then there will be symptoms due to this improperness. When the kidney is not working properly this would cause some changes in attributes such as serum creatinine, blood pressure, blood sugar and hemoglobin. Therefore this correlation among the attributes can be used to identify CKD. Doctors inherently use these attributes and their inter-relationships from reports such as blood reports and urine reports to identify the diseases.

If the attributes that contribute to CKD are identified clearer then it may be easier to identify CKD. The relationship among the attributes such as serum creatinine, blood pressure, hemoglobin and albumin is different for subjects with CKD and healthy subjects.

Therefore, first the covariance among those attributes for both healthy and CKD subjects has to be estimated. This common relationship is needed to be reduced or eliminated. As an attempt for the above process Common Spatial Patterns (CSP) [18] technique was used in order to maximize the separation between CKD subjects and healthy (non-CKD) subjects. A weight related to CSP filter with Linear Discriminant Analysis (LDA) and subsequent classification of CKD and non-CKD were used to identify the dominant attributes.

The CSP method is generally used in brain computer interface to maximize the difference in the electroencephalograph (EEG) data binary tasks (e.g. left hand and right hand imagination) [19-20]. That is to project in a space so that the separation between the two classes is maximized.

TABLE I
SELECTED DATA TYPE OF ATTRIBUTE

Major attributes	Abbreviation	Data type	Data
Age	age	num	in years
Blood Pressure	bp	num	in mm/Hg
Specific Gravity	sg	nom	(1.005,1.010,1.015,1.020,1.025)
Albumin	al	nom	(0,1,2,3,4,5)
Sugar	su	nom	(0,1,2,3,4,5)
Pus Cell	pc	nom	(normal,abnormal) →(1,0)
Pus Cell clumps	pcc	nom	(present,notpresent) →(1,0)
Bacteria	ba	nom	(present,notpresent) →(1,0)
Blood Glucose Random	bgr	num	in mgs/dl
Blood Urea	bu	num	in mgs/dl
Serum creatinine	sc	num	in mgs/dl
Hemoglobin	hemo	num	in gms
Hypertension	htn	nom	(yes,no) →(1,0)
Diabetes Mellitus	dm	nom	(yes,no) →(1,0)
Coronary Artery Disease	cad	nom	(yes,no) →(1,0)
Appetite	appet	nom	(good,poor) →(1,0)
Pedal Edema	pe	nom	(yes,no) →(1,0)
Anemia	ane	nom	(yes,no) →(1,0)

Here num means numerical, and nom means nominal, nominal attributes are changed to 1 and 0, (for example, data for Anemia is given as yes and no but in analysis it is changed to 1 and 0 accordingly.

II. DATASET SELECTION

The CKD dataset for this study was obtained from [21]. This dataset contains 24 attributes. Among them 13 are nominal and 11 are numerical attributes. There are 400 subjects, among them 250 are with CKD and 150 are healthy (not with CKD) subjects. Unfortunately in the dataset, for most of the subjects not all the attributes are listed. However, for a proper analysis all the subjects are supposed to have all attributes. To impose this uniformity only 12 nominal and 6 numerical attributes (see Table I) were selected for this study. These nominal attributes are represented by either 1 or 0 as mentioned in Table I. Accordingly, only the subjects who have all the 18 attributes were selected for the analysis, thus only 240 subjects were selected from 400 subjects contained in the dataset. That is, these 240 subjects have all the selected attributes. This does not mean that the remaining 6 attributes do not convey any information about CKD. The analysis in this paper tries to find the most important

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2017.2787480, IEEE Reviews in Biomedical Engineering

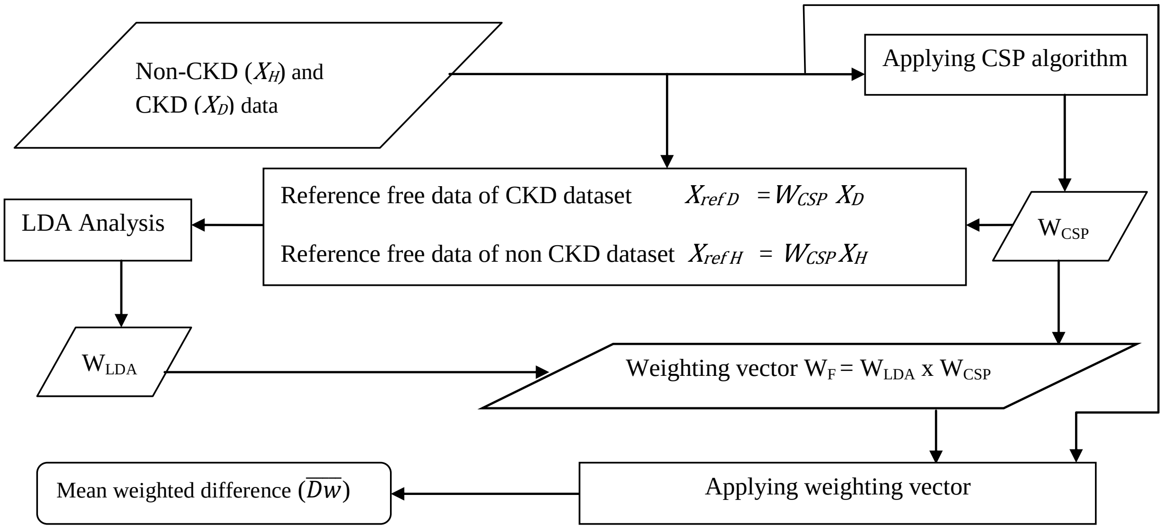
Non-CKD (X_H) and
CKD (X_D) data
Applying CSP algorithm
Reference free data of CKD dataset $X_{refD} = W_{CSP} X_D$
Reference free data of non CKD dataset $X_{refH} = W_{CSP} X_H$
LDA Analysis
W_{CSP}
W_{LDA}
Weighting vector $W_F = W_{LDA}$ x W_{CSP}
Mean weighted difference ($\overline{Dw}$)
Applying weighting vector

Fig. 1. Finding the mean weighted difference between the CKD and the non-CKD attributes

attributes among the selected 18 attributes. This analysis is limited by the dataset used.

For experiments, this dataset is divided into training and testing sets. For training 216 subjects and for testing the remaining subjects were used. For all the experiments the training set of 216 subjects were randomly selected and it was repeated 200 times as 200-fold dataset. The average results of these 200 experiments with 200 randomly selected set was used as the final results.

III. Finding the most important attribute for CKD

To find the important attributes a weighted difference method is proposed in this paper. To do that initially a weighting vector, W_F, was obtained that could separate CKD and non-CKD effectively. Fig. 1 illustrates how the weighting vector, W_F, is found.

In Fig. 1 X_D (CKD) and X_H (non-CKD) data were used to calculate CSP projection filter *(W_{CSP})* and then that projection matrix was applied to get the unique reference free components between CKD and healthy (non-CKD) attributes.

Matrix X_D denotes the attribute matrix of CKD with the dimensions of *A* and *S1*, where "*A"* is the number of attributes and "*S1"* is the number of subjects with CKD. Matrix X_H denotes the attribute matrix of healthy subjects (non-CKD) with the dimensions of *A* and *S2* where *"A"* is the number of attributes and *"S2"* is the number of healthy subjects.

Fig. 1 illustrates how the CSP filter (explained briefly in section A), W_{CSP}, was calculated and applied to the CKD and non-CKD data to obtain reference free data.

Then, using this reference free data LDA projection vector *(W_{LDA})* was found by LDA analysis. From W_{CSP} and W_{LDA}, the weighting vector, W_F, was found as shown in Fig. 1. Then weighted difference for each attribute ($D_w(a)$) between the CKD and non-CKD were found as shown in (1), where "*a*" and "*s*" are the indices of attribute and subject respectively.

$$D_w(a) = |W_F(a)\left[\frac{1}{S1}\sum_{s=1}^{S1} X_D(a,s) - \frac{1}{S2}\sum_{s=1}^{S2} X_H(a,s)\right]| \quad (1)$$

Then final mean weighted difference of each attribute averaged over the 200-fold sets was found as shown in (2). That is, (1) and (2) are used to calculate the distance between CKD and non-CKD for different attributes.

$$\overline{D_w^{200}(a)} = \frac{1}{200}\sum_{f=1}^{200} D_W(a,f) \quad (2)$$

Here it was assumed that the attribute corresponding to the highest final mean weighted difference, $\overline{D_w^{200}(a)}$, is the most important attribute in detecting CKD.

A. Common Spatial Pattern Filters

In this analysis CSP projection filter that leads to new components whose variances are best suited for the classification of two populations of attributes related to CKD and healthy (non-CKD) is developed. The CSP spatial filter design is based on simultaneous diagonalization of both classes' covariance matrices [18].

Algorithm details are described below for classifying X_D (CKD) and X_H (non-CKD) subjects. The projection filter, W_{CSP}, was developed in such a way to satisfy the properties shown in (3) and (4).

$$Cov(W_{CSP}X_D)\big/trace(X_D X_D^T) + Cov(W_{CSP}X_H)\big/trace(X_H X_H^T) = I \quad (3)$$

$$Cov(W_{CSP}X_D)\big/trace(X_D X_D^T) = D \quad (4)$$

Here ^{T} denotes the transpose operator, *I* refers the identity matrix, *trace(y)* is the sum of the diagonal elements of *y* and *D* refers diagonal matrix with elements monotonically descending. From (3) and (4) W_{CSP} was found and by filtering by the projection filter, W_{CSP}, the raw attributes were transformed into uncorrelated components. That gave the reference free components between CKD and healthy (non-CKD) attributes using (5) and (6) respectively.

$$X_{refD} = W_{CSP}\, X_D(a, s) \quad (5)$$
$$X_{refH} = W_{CSP}\, X_H(a, s) \quad (6)$$

B. *Finding the Most Important Attributes by Classification*

The idea is to identify the attributes that contribute most in classifying the data into CKD (X_D) and non-CKD (X_H). For classification the CSP filtered data (the reference free data X_{refH} and X_{refD}) were used. Four types of classification based analyses were performed to identify the most important attributes. Omit-one method and four-attribute-combination method together with two classifiers, namely LDA and K-nearest neighbor (KNN) were used to device those four analyses. These four types of analyses were explained below.

1) Omit-one method with LDA classifier

In this method each time one attribute was omitted, and the rest of the attributes were used in the classification. That is, the accuracies were obtained by omitting each attribute at a time and recorded against the omitted attribute. For classification LDA classifier was used.

Here it was assumed that the omitted attribute corresponding to the lowest accuracy contains the most information about CKD.

2) Omit-one method with KNN classifier

This is similar to section III B (1) but using KNN as the classifier instead of LDA. Here KNN classifier checked the 15 nearest neighbors.

3) Four-attribute-combination method with LDA classifier

In this method the attributes were combined into groups of 4 attributes. All possible combinations (C_4^{18}) were formed. For each combination the LDA classifier was used to train and evaluate the dataset and accuracy was obtained.

Here it cannot be simply assumed that all the attributes in the combination with highest accuracy have more information, than all the attributes with lower accuracy. Because there could be cases where only three out of four attributes contribute to the accuracy and the fourth may have not contributed.

Therefore, the important attributes were selected through a different process, where the number of times each attribute comes in the combination with relatively higher accuracy was considered. That is, first the accuracies were ordered in the descending order. Then the top 5% (153 combinations out of C_4^{18} combinations) of combinations starting from the highest accuracy combination were selected. Then the number of times (hits) each attribute was included in different combinations among the top 5% were counted. Here it is assumed that the attributes corresponding to the higher hits contain more information.

However, in addition to the above analysis the attributes corresponding to the best accuracy were also used in the analysis.

4) Four-attribute-combination method with KNN classifier

This is similar to section III B (3) but using KNN as

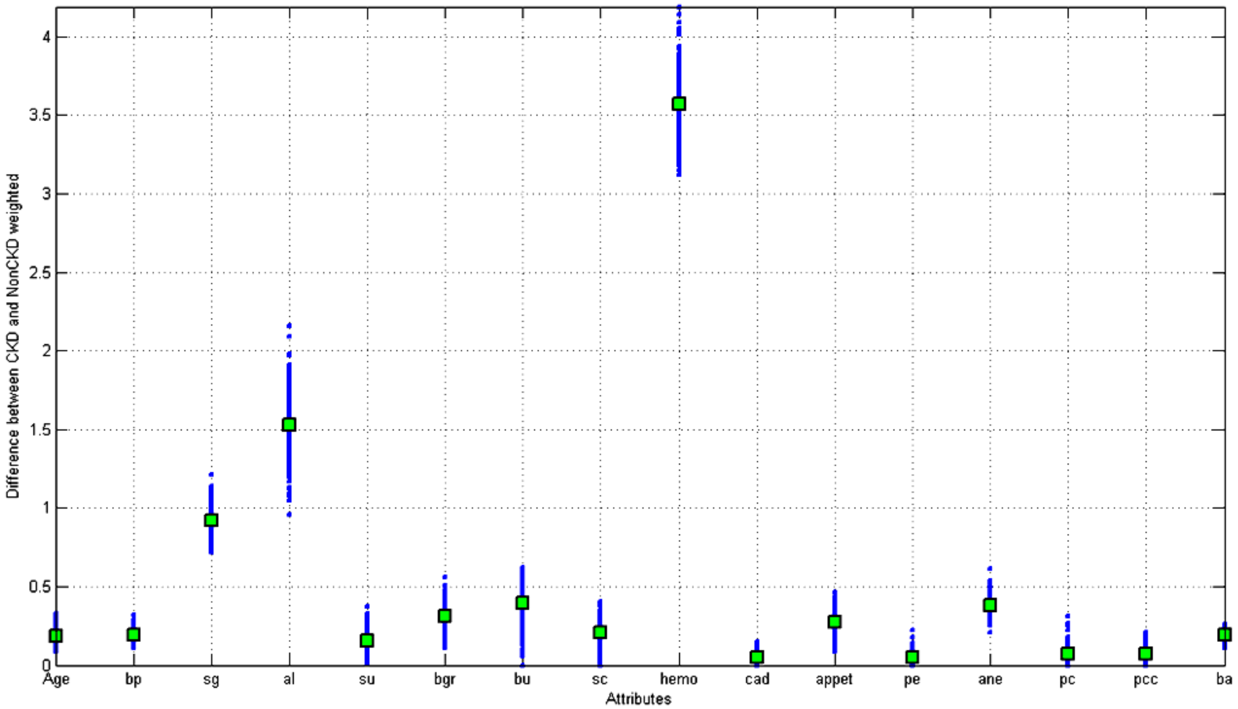

Fig. 2. Weighted difference between CKD and non-CKD of each attribute for 200-fold sets and the mean weighted difference (super imposed)

the classifier instead of LDA. Here KNN classifier checked the 15 nearest neighbors.

IV. Results and Discussion

Fig. 2 shows the weighted difference, $D_w(a)$, and the super imposed mean weighted difference, $\overline{D_w^{200}(a)}$, for each attribute. The attributes with higher, $\overline{D_w^{200}(a)}$ are listed in the descending order in the first column in Table II.

The accuracy of the classification results of omit-one method is shown in Fig. 3 against different omitted attributes for KNN and LDA classifiers. As mentioned earlier, it is assumed that the reduction in performance corresponds to the contribution of the omitted attribute in separating CKD and non-CKD. That is, the lower the performance is the higher the importance of that omitted attribute in detecting CKD. The important attributes found in this analysis (omitted attributes corresponding to lower accuracy) are listed in columns 2 and 3 respectively for LDA and KNN in Table II.

Then the results of four-attribute-combination method are shown in Fig. 4 for LDA and KNN classifiers. The attributes with higher hits are listed in descending order in columns 4 and 5 in Table II for LDA and KNN classifiers respectively.

When observing the order (that corresponds to the importance in detecting CKD) of the attributes in Table II for all the analyses it can be observed that these attributes are not in the same order of importance for different analyses. The order is not the same even for KNN and LDA classifiers. Thus a ranking method was used to finalize the important attributes across different methods.

A. Ranking Method

For each attribute in Table II the rank (the order or the position from the top) is listed in Table III. For example, in Table II hemoglobin is at the 1st position for the weighted difference method and at 11th position for the omit-one method using LDA, thus hemoglobin has the rank of 1 and 11 in the columns 2 and 3 respectively in Table III. In order to find the finalized ranking, the individual rankings across different methods were averaged and the mean ranking of each attribute is shown in the last column in Table III. The attributes are listed in Table III in the ascending order of the mean rank, thus descending order of importance for the detection of CKD. Here it is assumed that the lowest mean rank contains the most information of CKD. Hemoglobin is found to contain the most information of CKD among all 18 attributes.

It is worth noted that the discussion is centered on serum creatinine because generally in clinical practice, estimates of clearance based on the serum level are used to estimate GFR with the implied assumption that serum creatinine is the major attribute in separating CKD with non-CKD.

According to Table III, the risk factor, hypertension, seems more important than serum creatinine. Further,

TABLE II
Important Attributes Found With Different Methods

Weighted difference	Omit-one method C_{17}^{18}		Four-attribute-combination method (C_4^{18})	
	LDA	KNN	LDA	KNN
Hemoglobin	Hypertension	Anemia	Hemoglobin	Age
Albumin	Anemia	Pus Cell	Diabetes Mellitus	Hemoglobin
Hypertension	Coronary Artery Disease	Serum Creatinine	Hypertension	Hypertension
Diabetes Mellitus	Albumin	Blood Glucose Random	Albumin	Diabetes Mellitus
Specific Gravity	Age	Hemoglobin	Appetite	Serum Creatinine
Blood Urea	Pus Cell	Pedal Edema	Sugar	Blood Pressure
Anemia	Blood Glucose Random	Sugar	Pus Cell clumps	Specific Gravity
Serum Creatinine	Sugar	Bacteria	Blood Glucose Random	Blood Urea
Appetite	Pus Cell clumps	Specific Gravity	Pus Cell	Albumin
Pedal Edema	Serum Creatinine	Blood Urea	Pedal Edema	Blood Glucose Random
Bacteria	Hemoglobin	Albumin	Specific Gravity	Pus Cell clumps
Blood Pressure	Pedal Edema	Age	Serum Creatinine	Anemia
Pus Cell clumps	Blood Pressure	Diabetes Mellitus	Blood Pressure	Pedal Edema
Coronary Artery Disease	Specific Gravity	Pus Cell clumps	Bacteria	Coronary Artery Disease
Sugar	Appetite	Blood Pressure	Coronary Artery Disease	Sugar
Pus Cell	Blood Urea	Hypertension	Blood Urea	Appetite
Age	Diabetes Mellitus	Appetite	Anemia	Bacteria
Blood Glucose Random	Bacteria	Coronary Artery Disease	Age	Pus Cell

The attributes are ordered in such a way that the most important attribute in detecting CKD is at the top and gradually reducing the importance towards the bottom of the order.

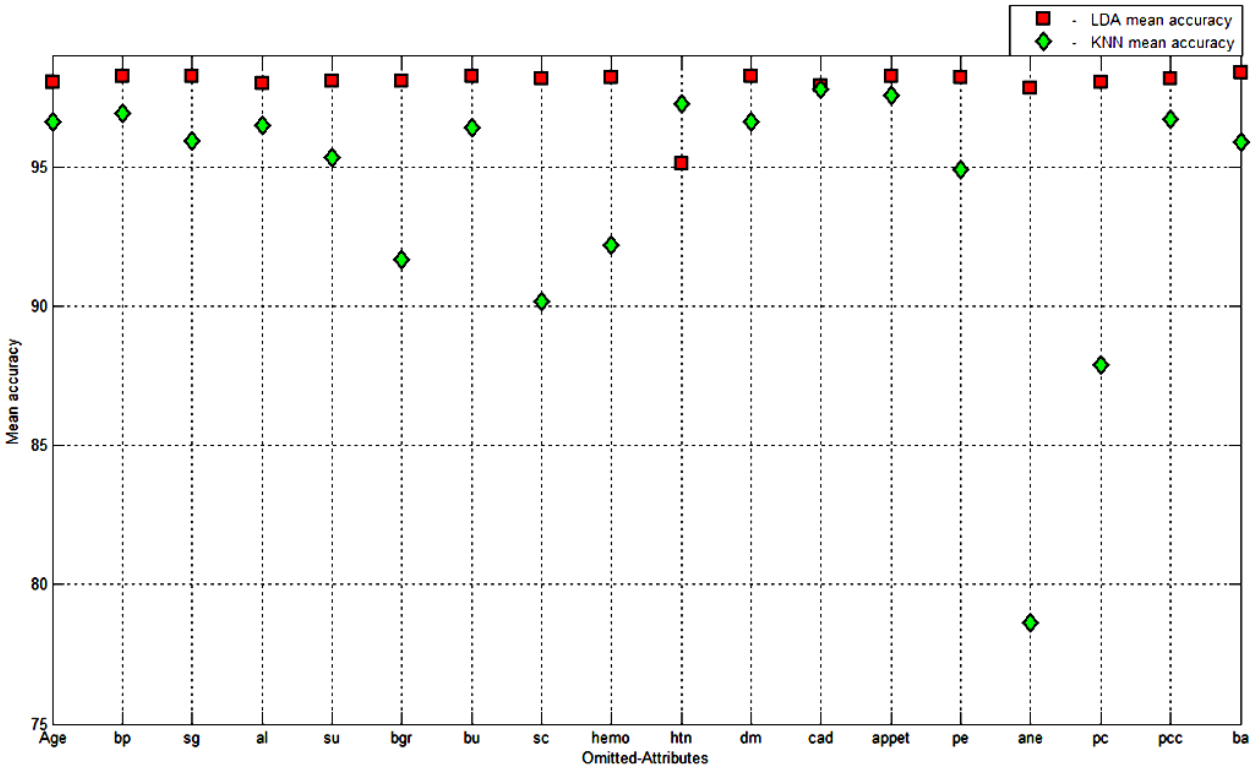

Fig. 3. Omit-one method's mean accuracy for 200-fold sets with omitted attributes for KNN and LDA classifiers

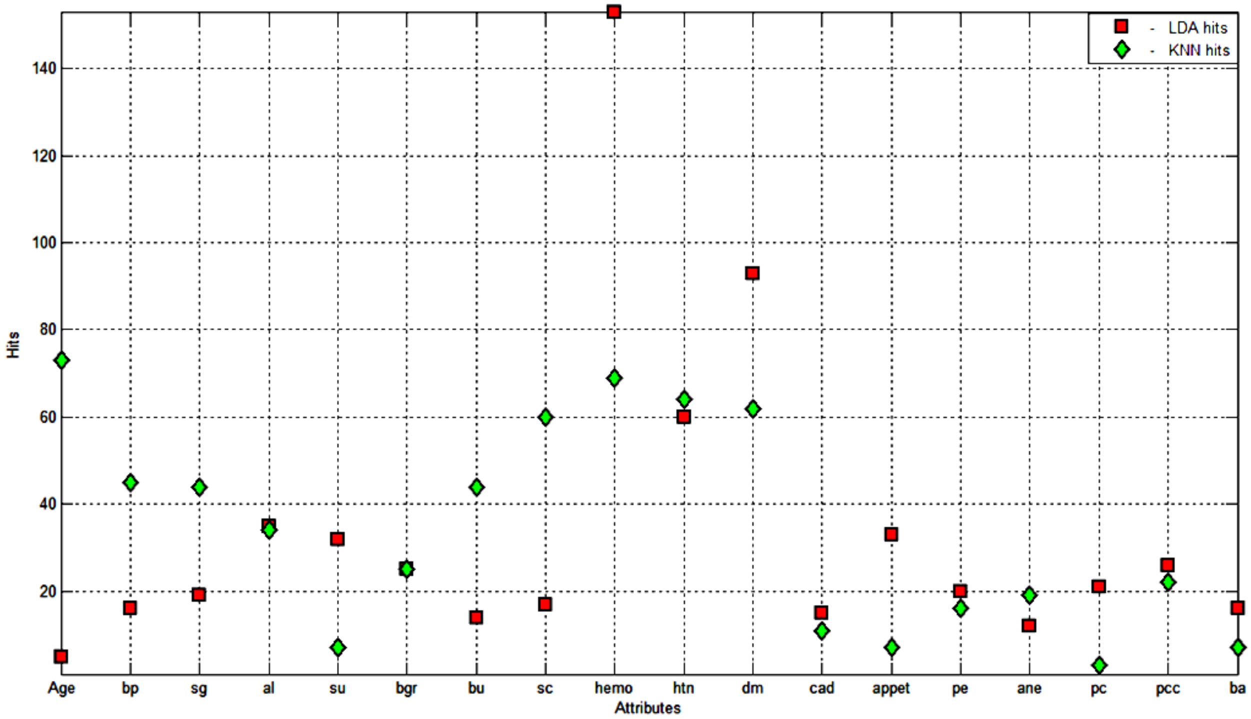

Fig. 4. Hits of attributes for four-attribute-combination method (C_4^{18}) with KNN and LDA classifiers

though the risk factor, diabetes mellitus, is ranked below serum creatinine its individual position in Table II is higher than that of serum creatinine except omit-one method. It is assumed that the information in the omitted attribute is responsible for the reduction in performance. When diabetes mellitus is omitted the information loss due to this may be partly compensated by blood glucose random and that could be the reason why only in omit-one method diabetes mellitus is below serum creatinine in Table II. Thus the hypertension and diabetes mellitus may be considered important than serum creatinine.

TABLE III
THE RANK(POSITION) ACCORDING TO TABLE II AND THE MEAN RANK FOR EACH ATTRIBUTE

Attributes (In descending order of importance according to mean ranking)	Individual ranking					Mean ranking(C_4^{18})
	Weighted difference method	Omit-one method		Four- attribute-combination method		
		LDA	KNN	LDA	KNN	
Hemoglobin	1	11	5	1	2	4
Hypertension	3	1	16	3	3	5.2
Albumin	2	4	11	4	9	6
Serum Creatinine	8	10	3	12	5	7.6
Anemia	7	2	1	17	12	7.8
Diabetes Mellitus	4	17	13	2	4	8
Specific Gravity	5	14	9	11	7	9.2
Blood Glucose Random	18	7	4	8	10	9.4
Pedal Edema	10	12	6	10	13	10.2
Sugar	15	8	7	6	15	10.2
Pus Cell	16	6	2	9	18	10.2
Age	17	5	12	18	1	10.6
Pus Cell clumps	13	9	14	7	11	10.8
Blood Urea	6	16	10	16	8	11.2
Blood Pressure	12	13	15	13	6	11.8
Appetite	9	15	17	5	16	12.4
Coronary Artery Disease	14	3	18	15	14	12.8
Bacteria	11	18	8	14	17	13.6

For each attribute in Table II the rank (position) is listed in this table. For example, in Table II hemoglobin is at 1st position for the weight difference method and at 11th position for the omit-one method using LDA, thus hemoglobin has the rank of 1 and 11 in the columns 2 and 3 respectively in Table III.

However, risk factors, hypertension and diabetes mellitus, are already known as major causes for CKD [4]. In [22] it is mentioned as "Diabetes is the leading cause of end-stage kidney disease worldwide". Thus the study is extended in the next section to find the important attributes except the two risk factors by repeating all the above mentioned experiments without the two risk factors.

B. Analysis without the Risk factors of Hypertension and Diabetes Mellitus

The attributes that contribute most in separating CKD with non-CKD resulted in the repeated experiments without the risk factors are given in Table IV and the corresponding ranking is shown in Table V.

In this analysis the attributes hemoglobin, albumin, specific gravity, blood glucose random, and blood pressure performs better than the serum creatinine.

An interesting observation in Table V is that the blood glucose random and blood pressure are ranked above serum creatinine as opposed to the observation in Table III. This implies that the reason blood glucose random and blood pressure are ranked below serum creatinine is the inclusion of diabetes mellitus and hypertension in the analysis for Table III.

So it may be stated that, when a subject does not have hypertension and diabetes mellitus data, the information of blood pressure and blood glucose random may be used in detecting CKD.

TABLE IV
BEST ATTRIBUTES WITH THE METHODS FOR HYPERTENSION AND DIABETES MELLITUS REMOVED DATA

Omit-one method (C_{15}^{16})	Four-attribute-combination method(C_4^{16})	
LDA	LDA	KNN
Serum Creatinine	Hemoglobin	Hemoglobin
Anemia	Albumin	Age
Pus Cell clumps	Sugar	Specific Gravity
Coronary Artery Disease	Blood Pressure	Blood Urea
Hemoglobin	Blood Glucose Random	Blood Glucose Random
Pus Cell	Specific Gravity	Blood Pressure
Age	Appetite	Albumin
Blood Glucose Random	Pedal Edema	Pus Cell clumps
Appetite	Coronary Artery Disease	Serum Creatinine
Blood Pressure	Pus Cell clumps	Appetite
Pedal Edema	Pus Cell	Pedal Edema
Sugar	Bacteria	Anemia
Blood Urea	Serum Creatinine	Sugar
Specific Gravity	Anemia	Bacteria
Bacteria	Blood Urea	Pus Cell
Albumin	Age	Coronary Artery Disease

The attributes are ordered highest information to lowest order in the table columns.

TABLE V
RANK (POSITION) ACCORDING TO TABLE IV AND THE MEAN RANK FOR EACH ATTRIBUTE FOR HYPERTENSION AND DIABETES MELLITUS REMOVED DATA

Attributes (In descending order of importance according to mean ranking)	Individual ranking				Mean ranking
	Weighted difference method	Omit-one method	Four- attribute combination method (C_4^{16})		
		LDA	LDA	KNN	
Hemoglobin	1	5	1	1	2
Blood Glucose Random	6	8	5	5	6
Specific Gravity	3	14	6	3	6.5
Albumin	2	16	2	7	6.75
Blood Pressure	10	10	4	6	7.5
Serum Creatinine	8	1	13	9	7.75
Appetite	7	9	7	10	8.25
Anemia	5	2	14	12	8.25
Pus Cell clumps	14	3	10	8	8.75
Age	11	7	16	2	9
Blood Urea	4	13	15	4	9
Sugar	12	12	3	13	10
Coronary Artery Disease	15	4	9	16	11
Pus Cell	13	6	11	15	11.25
Pedal Edema	16	11	8	11	11.5
Bacteria	9	15	12	14	12.5

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2017.2787480, IEEE Reviews in Biomedical Engineering

TABLE VI
HIGHEST ACCURACY COMBINATION OF FOUR ATTRIBUTE COMBINATION METHOD

Combination with classifier		Best attribute combination			
With Hypertension and Diabetes Mellitus data	LDA	specific gravity	albumin	diabetes mellitus	hemoglobin
	KNN	specific gravity	sugar	serum creatinine	hemoglobin
Hypertension and Diabetes Mellitus removed data	LDA	bacteria	albumin	appetite	hemoglobin
		specific gravity	albumin	blood glucose random	hemoglobin
	KNN	specific gravity	sugar	serum creatinine	hemoglobin

C. Importance of Age

In the development of the calculation of eGFR in [9-14] age is considered as an important factor; however by looking at Table II the age is always lie at lower importance level for all the methods except four-attribute-combination method using KNN as classifier and omit-one method using LDA classifier. It should be noted that our analyses do not rely on eGFR calculation and it is an independent study without using eGFR (serum creatinine based estimation) based analysis.

D. Important Attributes

From all these analyses and results (particularly from Table III and V) it can be concluded that the most important attributes in detecting CKD are hemoglobin, albumin, specific gravity, serum creatinine, hypertension and diabetes mellitus. Further, when hypertension and diabetes mellitus are not available, blood glucose random and blood pressure fills that gap.

E. Highest Accuracy Combination of Four Attribute Combination Method

The combination of attributes that gave the highest accuracy in the four-attribute combination method is shown in Table VI. It can be observed that most of the attributes identified in section IV (D) came in that combination listed in Table VI.

F. Additional Experiments with Selected Attributes.

The best attributes found as important in Section IV (D) are grouped in selected combinations as listed under Table VII and classification experiments are performed with LDA and KNN classifier. The corresponding mean accuracy of the 200-fold dataset are tabulated in Table VII. The highest accuracy is obtained when hemoglobin, specific gravity, albumin, hypertension, and diabetes mellitus are grouped together. Thus it can be concluded that the combination of the above attributes contain significant information about CKD. It can be noted that the above combination does not have serum creatinine.

TABLE VII
MEAN ACCURACY OF SELECTED ATTRIBUTES COMBINATIONS

	With serum creatinine			without serum creatinine		
	Group1	Group2	Group5	Group3	Group4	Group6
LDA	95.79	98.09	97.21	96.25	98.81	98.03
KNN	47.33	71.65	54.94	73.25	50.13	58.19

Set 1: hemoglobin, specific gravity, albumin
Set 2: hypertension, diabetes mellitus
Set 3: blood glucose random, blood pressure

Attribute combinations details:
Group1: Set 1 and serum creatinine
Group2: Set 1, Set 2 and, serum creatinine
Group3: Set 1
Group4: Set 1 and Set 2
Group5: Set 1, Set 3 and, serum creatinine
Group6: Set 1 and Set 3

TABLE VIII
PERFORMANCE FACTORS OF LDA CLASSIFIER WITH 200-FOLD DATASET

	With serum creatinine			without serum creatinine		
	Group1 Set 1	Group2 Set 1 & Set 2	Group 5 Set 1& Set 3	Group3 Set 1	Group4 Set 1 & Set 2	Group 6 Set 1& Set 3
TypeI error rate	0.05	0.04	0.04	0.04	0.02	0.03
TypeII error rate	0.03	0.00	0.01	0.03	0.00	0.01
sensitivity	0.95	0.96	0.96	0.96	0.98	0.97
specificity	0.96	1.00	0.99	0.96	1.00	0.99
accuracy	0.96	0.98	0.97	0.96	0.99	0.98
F_score	0.66	0.67	0.66	0.66	0.67	0.67
kappa	0.91	0.96	0.94	0.92	0.98	0.96

Attribute combinations details are same as shown in Table VII.

In addition, the performance factors, Type I error rate, Type II error rate, sensitivity, specificity, accuracy, F-score and Kappa, as mentioned in [17] is also analysed for the experiment with LDA model (since LDA classifier performs better than KNN classifier, only LDA was selected for this analysis). The corresponding results are tabulated in Table VIII. From that the Kappa value becomes nearest to one when hemoglobin, albumin, specific gravity, and risk factors hypertension and diabetes mellitus attributes are grouped together, thus this combination is again proved to be the most important attributes in detecting CKD. Again serum creatinine is not in the above group.

Thus serum creatinine may not the only best attribute for eGFR calculation which contributes CKD classification. Because serum creatinine is not only relying on kidney function but due to secretion as well, so it may cause false detection of CKD [23]. That is, when GFR decreases, the percentage of excretion due to secretion increases, thus substances that block distal tubule secretion of creatinine may cause the serum level to increase abruptly, while GFR remain the same as explained in [23]. In this scenario considering the serum creatinine to calculate eGFR may not give the correct measurement of the kidney function.

G. *Classifier to classify CKD and non-CKD subjects*

From Table VII it can be seen that the LDA classifier performed better than KNN classifier and the accuracy in three of the cases are more than 98%. Among the three cases the classifier details for the case of group 2 that includes serum creatinine (as in Table VII) is given below. This classifier can be considered as a suboptimal classifier obtained as a byproduct of this study. (It is suboptimal because only two classifiers are compared and to find an optimal classifier it requires an extensive research). The reason to select this case among the three cases with greater than 98% is because this case includes the widely used attribute serum creatinine. The classifier weights and LDA classifier parameters for this case are given below. Attributes values (*A_V*) should be in the given order as shown in (7) and the corresponding units are as shown in Table I.

A_V= [specific gravity, albumin, serum creatinine,
hemoglobin, hypertension, diabetes mellitus] (7)

$$W = [\,-92.2828 \quad 0.4110 \quad 0.0204 \quad -0.5943 \quad 1.3764 \quad 1.7119\,] \tag{8}$$

$$class = \underset{c\in[CKD,Non\_CKD]}{\arg\min} \sum_{i=1}^{6} \frac{\left(A\_V(i).W(i) - M_{class}(c)\right)^2}{V_{class}(c)} \tag{9}$$

Here,

$$[M_{class}(CKD), M_{class}(Non\_CKD)] = [-97.35,\ -103.57] \tag{10}$$

$$[V_{class}(CKD), V_{class}(Non\_CKD)] = [5.37, 0.9] \tag{11}$$

As shown in (9) the values of the two classes are calculated and the class with the minimum value is considered as the class. This model can be used if the subject has all these 6 attributes.

V. Conclusion

If the important attributes that could help to detect CKD is known then even people who are not diagnosed CKD also may get a clue of the condition of their kidney from the medical test that they took for some other purposes. Then they may proceed to properly check for CKD. A weighting vector based on CSP filter and LDA analysis and then classification analysis using LDA and KNN classifiers were used to identify the dominant attributes. It is found that hemoglobin, albumin, specific gravity, serum creatinine, hypertension and diabetes mellitus are the most important attributes in detecting CKD. Further, these analyses suggest that when hypertension and diabetes mellitus are not available, blood glucose random and blood pressure can be used.

References

[1] S. Ardhanari, M. A. Alpert, and K. Aggarwal, "Cardiovascular disease in chronic kidney disease: risk factors, pathogenesis, and prevention," *Adv Perit Dial*, vol. 30, pp. 40–53, 2014.

[2] M. J. Sarnak, A. S. Levey, A. C. Schoolwerth, *et al.*, "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," *Hypertension*, vol. 42, no. 5, pp. 1050–1065, Nov. 2003.

[3] "National Chronic Kidney Disease Fact Sheet", June. 6, 2016. [Online]. Available: Centers for Disease Control and Prevention [June. 10,2016].

[4] D. Johnson. (2012, July). "Risk factors for early chronic kidney disease." [Online]. Available: http://www.cari.org.au/CKD/CKD%20early/Risk_Factors_Early_CKD.pdf [June 10, 2016].

[5] R. Walker, M. Marshall, and N. Polaschek, "Improving self-management in chronic kidney disease: a pilot study," *Ren. Soc. Australas. J.*, vol. 9, no. 3, pp. 116–125, 2013.

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

[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 : a new prediction equation," *Ann. Intern. Med.*, vol. 130, no. 6, pp. 461–470, 1999.

[8] L. A. Inker, C. H. Schmid, H. Tighiouart, *et al.*, "Estimating Glomerular Filtration Rate from Serum and Cystatin C," *N. Engl. J. Med.*, vol. 367, no. 1, pp. 20–29, Jul. 2012.

[9] J. Coresh, B. C. Astor, T. Greene, G. Eknoyan, and A. S. Levey, "Prevalence of chronic kidney disease and decreased kidney function in the adult US population: Third National Health and Nutrition Examination Survey.," *Am. J. Kidney Dis. Off. J. Natl. Kidney Found.*, vol. 41, no. 1, pp. 1–12, Jan. 2003.

[10] G. Manjunath, H. Tighiouart, J. Coresh *et al.*, "Level of kidney function as a risk factor for cardiovascular outcomes in the elderly," *Kidney Int.*, vol. 63, no. 3, pp. 1121–1129, 2003.

[11] M. G. Shlipak, P. A. Heidenreich, H. Noguchi, G. M. Chertow, W. S. Browner, and M. B. McClellan, "Association of renal insufficiency with treatment and outcomes after myocardial infarction in elderly patients.," *Ann. Intern. Med.*, vol. 137, no. 7, pp. 555–562, Oct. 2002..

[12]N. T. Raymond, D. Zehnder, S. C. H. Smith, J. A. Stinson, H. Lehnert, and R. M. Higgins, "Elevated relative mortality risk with mild-to-moderate chronic kidney disease decreases with age," *Nephrol. Dial. Transplant.*, vol. 22, no. 11, pp. 3214–3220, Jun. 2007.

[13]A. M. O'Hare, D. Bertenthal, K. E. Covinsky *et al.*, "Mortality risk stratification in chronic kidney disease: one size for all ages?", *J. Am. Soc. Nephrol. JASN*, vol. 17, no. 3, pp. 846–853, Mar. 2006.

[14]D. Johnson, "Diagnosis, classification and staging of chronic kidney disease," *Kidney Health Aust.*, pp. 5–7, 2012.

[15]T. H. Jafar, C. H. Schmid, P. C. Stark *et al.*, "The rate of progression of renal disease may not be slower in women compared with men: a patient-level meta-analysis," *Nephrol. Dial. Transplant.*, vol. 18, no. 10, pp. 2047–2053, Oct. 2003.

[16] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney inter., Suppl. 2013; 3: 1–150.

[17]L. J. Rubini and P. Eswaran, "Generating comparative analysis of early stage prediction of Chronic Kidney Disease." *International Journal Of Modern Engineering Research*, vol. 50, pp. 49–55, Jul. 2015.

[18]H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement*," IEEE Transactions On Rehabilitation Engineering*, Vol. 8, NO. 4, pp. 441–446, 2000.

[19]M. Akcakaya, B. Peters, M. Moghadamfalahi *et al.*, "Noninvasive Brain-Computer Interfaces for Augmentative and Alternative Communication," *IEEE Reviews In Biomedical Engineering*, VOL. 7, pp. 31–49, 2014.

[20]B. Reuderink and M. Poel, "Robustness of the common spatial patterns algorithm in the BCI-pipeline," 2008, Univ. of Twente.

[21] L. Jerlin Rubini. (2015). UCI Chronic Kidney Disease [Online]. Available:https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. Irvine, CA: University of California, School of Information and Computer Science.

[22] L. Zhang, J. Long, W. Jiang, *et al.*, "Trends in chronic kidney disease in China," *N. Engl. J. Med.*, vol. 375, no. 9, pp. 905–906, 2016.

[23]"Diagnosis and Management of Chronic Kidney Disease", Nov, 2008. [Online]. Available:http://www.mayomedicallaboratories.com/articles/communique/2008/11.html[Oct. 18, 2016].