Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods

Hüseyin Polat

J Med Syst (2017) 41:55
DOI 10.1007/s10916-017-0703-x

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods

Huseyin Polat[1] · Homay Danaei Mehr[1] · Aydin Cetin[1]

Received: 28 July 2016 / Accepted: 9 February 2017

Abstract As Chronic Kidney Disease progresses slowly, early detection and effective treatment are the only cure to reduce the mortality rate. Machine learning techniques are gaining significance in medical diagnosis because of their classification ability with high accuracy rates. The accuracy of classification algorithms depend on the use of correct feature selection algorithms to reduce the dimension of datasets. In this study, Support Vector Machine classification algorithm was used to diagnose Chronic Kidney Disease. To diagnose the Chronic Kidney Disease, two essential types of feature selection methods namely, wrapper and filter approaches were chosen to reduce the dimension of Chronic Kidney Disease dataset. In wrapper approach, classifier subset evaluator with greedy stepwise search engine and wrapper subset evaluator with the Best First search engine were used. In filter approach, correlation feature selection subset evaluator with greedy stepwise search engine and filtered subset evaluator with the Best First search engine were used. The results showed that the Support Vector Machine classifier by using filtered subset evaluator with the Best First search engine feature selection method has higher accuracy rate (98.5%) in the diagnosis of Chronic Kidney Disease compared to other selected methods.

Keywords Feature selection · Support vector machine · Chronic kidney disease · Machine learning

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ Huseyin Polat
polath@gazi.edu.tr

[1] Department of Computer Engineering, Faculty of Technology, Gazi University, 06500, Teknikokullar, Ankara, Turkey

Abbreviations

CKD	Chronic Kidney Disease
UCI	University of California Irvine
SVM	Support Vector Machine
GA	Genetic Algorithm
SymmetricUncertAttributesetEval	Symmetrical uncertainty attribute set evaluator
SVEGA	Shapely Value Embedded Genetic Algorithm
KNN	K-nearest Neighbor
GainRatioAttributeEval	Gain ratio attribute evaluator
PrincipalComponentsAttributeEval	Principal components attribute evaluator
SIMCA	Soft Independent Modeling of Class Analogy
AUC	Area Under the roc Curve
TCMSP	Traditional Chinese Medicine Syndrome Prediction method
OSAF	Oscillating Search Algorithm Feature Selection
NotCKD	Without Chronic Kidney Disease
ClassifierSubsetEval	Classifier subset evaluator

Published online: 27 February 2017

WrapperSubsetEval	Wrapper subset evaluator
FilterSubsetEval	Filtered subset evaluator
CfsSubsetEval	Correlation feature selection subset evaluator
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ROC	Receiver Operating Characteristic

Introduction

Chronic Kidney Disease (*CKD*) or chronic renal disease gradually progresses and usually after months or years the kidney loses its functionality. In general it may not be detected before it loses 25% of its functionality. The start of renal failure may not be recognized by the patients since kidney failure may not give any symptoms initially.

Kidney failure treatment targets to control the causes and decelerate the advance of the renal failure. If treatments are not enough, patient will be in the end-stage of renal failure and the last treatment is dialysis or renal transplant. At present, 4 out of every 1000 person in the United Kingdom are suffering from renal failure [1] and more than 300,000 American patients in the end-stage of kidney disease survive with dialysis [2]. Moreover, according to the National Health Service kidney disease is more frequent in South Asia, Africa, than in the other countries. Due to detecting the chronic kidney failure is not feasible until the kidney failure is completely progressed; thus, realizing the kidney failure in the first stage is extremely important. Through early diagnosis, the act of each kidney can be taken under control, which leads to decreasing the risk of irreversible consequences. For this reason, routine check-up and early diagnosis are crucial to the patients, for they can prevent vital risks of renal failure and related diseases [1]. Blood test is one of the steps to detect *CKD*. Therefore, it can be distinguished by measuring factors, and physicians can decide treatment processes, reducing the rate of progression [3].

The purpose of medical diagnosis is to mine useful information from the massive medical datasets which are accumulated frequently [4]. Vast majority of the studies on medical datasets are related to cancer diagnosis. There are only a few studies related to *CKD*. Neves et al., have used Artificial Neural Network to *CKD* diagnose. Results were showed that the sensitivity value of diagnosis was in the range of 93.1%–94.9% and the specificity value of diagnosis were in the range of 91.9%–94.2% [5]. Di Noia et al., have developed a software tool by using artificial neural networks for classification of end-stage kidney disease. Their proposed model has obtained 91.37% accuracy with 70.76% of sensitivity and 70.76% of positive predictive [6]. Chen et al., have used Soft Independent Modeling of Class Analogy (*SIMCA*), Support Vector Machine (*SVM*) and K-nearest Neighbor (*KNN*) on *CKD* dataset for the prediction of CKD risks. They have generated a composite dataset by adding random noise to all samples and then fusing it with original dataset. Consequently, they have created two subsets, training and test sets by adding the different variations, to investigate the ability of noise disturbance tolerance of the used classification algorithms. They have reported that the accuracy rate of *SVM* and *KNN* on original dataset were higher than *SIMCA*. Moreover, *SVM* has achieved higher accuracy rate than the others on two generated composite dataset with random noise and fused datasets of original and composite data [7]. However, authors did not perform feature selection methodologies on CKD dataset.

Classification methods on disease and physicians' diagnosis datasets are useful for experts in medical diagnosis. Classification methods can minimize diagnosis faults that can occur by less experienced physicians and results can be obtained in a short period of time [8]. However, there is not a single or generalized algorithm available in machine learning and decision making models to diagnose all types of diseases. The accuracy rate of classification of high dimensional datasets can be low and features can be at over fitting risks, and computational effort may be high and costly. Generally, low dimensional datasets can lead to higher classification accuracy with lower computational cost and the risk of over fitting can be abated [9].

In this study, we have investigated the accuracy rate of the methods using feature selection by data reduction. Wrapper and filter feature selection methods have been used to reduce the dimension of features for the diagnosis of *CKD.* Then, *SVM* has been used for classification of features. Datasets were obtained from University of California Irvine (*UCI*) machine learning repository.

The article is organized as follows: Section 2 is a review of the used feature selection methods for other diagnoses of disease; Section 3 and 4 present review of feature selection methods and *SVM* algorithm; Section 5 gives a review of used methodology, and Section 6 presents the simulation test results for the used methods, comparing each with one other.

Feature selection

Feature selection is the main field in knowledge discovery, pattern recognition and statistical science. The purpose of feature selection is to remove a subset from inputs which are not important. Features do not depend on information about predictive classes. Reducing the dimensions of features and irrelevant features can produce a comprehensive model for classification. The main challenge of feature reduction is recognizing the best subset of features to achieve the best results of classifications [10]. Feature selection can simplify the data realization, decrease overfitting problem and the size of data storage, also it can decrease the cost of train to obtain higher accuracy [11].

Feature selection methods can be categorized into three groups. Filter, wrapper and embedded methods. Fig. 1 shows the schema of filter and wrapper methods for feature selection and *SVM* classification algorithm for classifying the selected subset methods used in this paper.

The filter method does not depend on any learning algorithm. The filter method selects the features whose ranks are the highest among them, and then the selected subset can be prepared for any classification algorithm. After applying feature selection with filter method, various classification algorithms could be evaluated (Fig. 2) [12]. An appropriate feature selection yields to performance improvement of classifier by reducing the computing time and by using optimized data in the dataset [13]. Furthermore, Filter method is a popular method in feature selection due to its fast performance and scalability [14, 15].

Wrapper method calculates scores of feature sets that rely on the estimated power by using a classifier algorithm as a black box [16]. Evaluation of specific subset is achieved by performing test and training on that specific dataset. The wrapped search algorithm around the classifier gains the space of all features of subsets [12]. 2^n (n is the number of features), various assessments are required for a full search in the wrapper method. Although dealing with correlated features and finding the relevant associations are the advantages of this approach, it might leads to the overfitting problems [17] (Fig. 3).

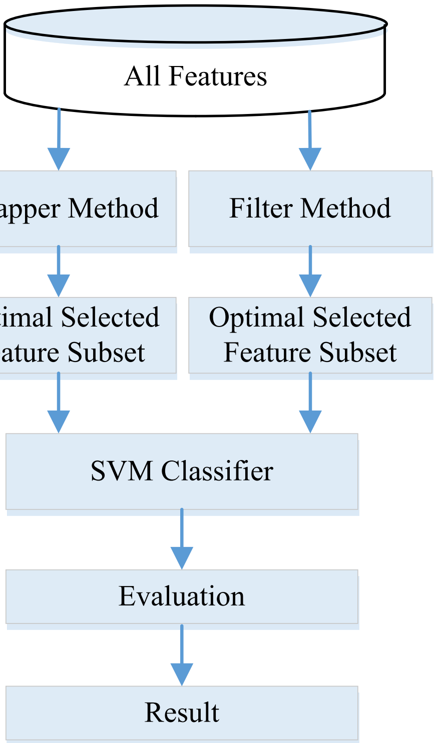

Fig. 1 Schema of SVM classification with filter and wrapper methods

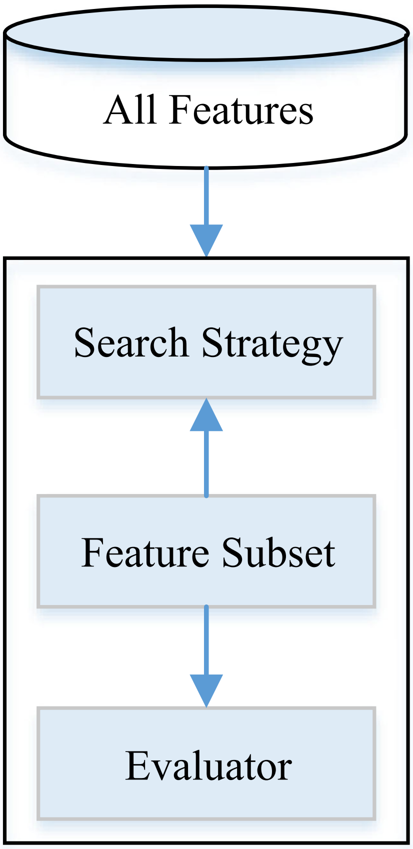

Fig. 2 Schema of the filter feature selection method

In embedded method, feature selection is embedded with structure of classifier (Fig. 4). The advantage of embedded method is that they interact with their classification model, and they do not have complicated computation [18].

Decision support systems use different methods to reduce the features dimension and classification algorithms to diagnose several kinds of diseases.

Lavanya and Rani to analyze the performance of classification before and after using feature selection methods, they have used Decision Tree- CART classifier and different feature selection methods on the same scope of three Breast Cancer datasets. Correlation feature selection subset evaluator *(CfsSubsetEval)*, classifier subset evaluator (*ClassifierSubsetEval*), symmetrical uncertainty attribute set evaluator (*SymmetricUncertAttributesetEval))*, filtered subset evaluator *(FilterSubsetEval)*, gain ratio attribute evaluator *(GainRatioAttributeEval)*, *SVM* attribute evaluator, principal components attribute evaluator *(PrincipalComponentsAttributeEval)* and some other methods were used. In all used evaluators, some searching methods such as Best First, Greedy stepwise, Ranker, Scatter and Genetic were used. Results have shown that in each dataset different evaluators improved

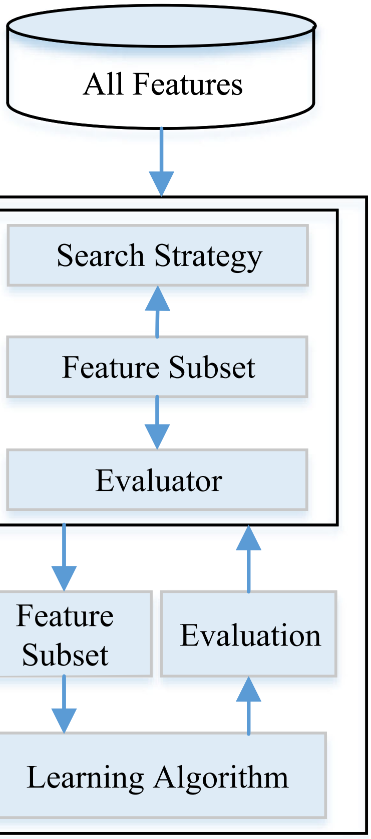

Fig. 3 Schema of the wrapper feature selection method

the classification performance. By comparing, all used evaluators results have demonstrated that accuracy of Decision Tree-CART classifier on Breast Cancer Wisconsin (Original) dataset by using *PrincipalComponentsAttributeEval* by 96.99% was higher than the other two used evaluators on the same dataset. The accuracy of Decision Tree classifier by using *SymmetricUncertAttributesetEval* on Breast Cancer Wisconsin (Diagnostic) dataset by 94.72% was the highest in all used evaluators on mentioned dataset. The highest accuracy rate of classifier on Breast Cancer dataset by using *SVM* attribute evaluator was 73.07% [19]. Jiang et al., have used *CfsSubsetEval* with Best First search as feature selection method and *SVM* classifier to predict the hepatotoxic compounds in traditional Chinese medicine. Results have shown that their using method have achieved 82.41% accuracy rate. Despite six positive compounds were detected false, they claimed that their used method was better than the other previous used methods in detecting the hepatotoxic compounds [20]. S. Sasikala et al., have used Shapely Value Embedded Genetic Algorithm (*SVEGA*) to feature dimension reduction and improve the accuracy rate of the diagnosis of breast cancer. They have reported better accuracy rates than such algorithms as Naive Bayes, *KNN* and *SVM* [21]. Ćosović et al., have used filter and wrapper feature selection method to select the best subset of three datasets of internet attacks i.e., Slammer, Nimda and Code Red I [22–24]. *CfsSubsetEval* and *GainRatioAttributeEval* as filter methods were used. *ClassifierSubsetEval* and wrapper subset evaluator (*WrapperSubsetEval*) as wrapper feature selection methods were used. To discover the internet anomalies they used *SVM* Radial Basis Function, Naïve Bayes and Decision Tree-J48, on selected subsets and compared the performance of the classifiers. Results have shown that Naïve Bayes on reduced Nimda dataset by *ClassifierSubsetEval* has got the highest F-measure and Decision Tree-J48 classifier on reduced Code Red I dataset by *WrapperSubsetEval* has got the highest F-measure. *SVM* classifier on reduced Slammer dataset by *WrapperSubsetEval* achieved the highest all performance measurements. (F-measure, recall rate) [25]. Özçift and Gülten, have proposed a feature selection which is a combination of Genetic Algorithm (*GA*) and Wrapped Bayesian Network and using Best First and Sequential Floating search methods for erythema-squamous diseases diagnosis. It has a higher accuracy rate than the other used methods such as *SVM*, Multi-Layer Perceptron, Simple Logistics and Functional Decision Tree [9]. Akbarisanto et al., have applied *CfsSubsetEval* and *ClassifiersubsetEval* feature selection methods to reduce the features of Bandung dataset which include in people's tweets. To investigate the performance of classification algorithms in mood classification of people's tweets Naïve Bayes and *SVM* classifiers were compared. In this regard after choosing the best subset of dataset by feature selection methods, classifiers were used on reduced dataset. According to their results, Naïve Bayes classifier on reduced dataset by *CfsSubsetEval* evaluator has achieved higher accuracy rate by 89.0411% than *SVM* classifier [26]. Yan et al., have proposed Information Gain and Traditional Chinese Medicine Syndrome Prediction method (*TCMSP*) for feature

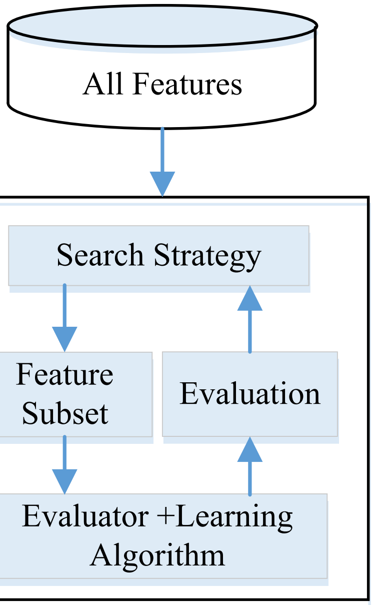

Fig. 4 Schema of the embedded feature selection method

selection and classification of liver cirrhosis dataset. After utilizing feature selection, dataset has been re-trained by different algorithms such as *SVM*, Naive Bayes, etc. It can be interpreted that proposed method improved the accuracy rate of all algorithms for the diagnosis of the diseases. Furthermore, this method is used for heart disease, lung cancer, and iris datasets. Results have demonstrated that the new method was better than other methods on three datasets [27]. R. Chaves et al., have used t-test feature selection method for *SPECT* images of Alzheimer's disease and have got 98.3% accuracy rate for *SVM* classification algorithm to early diagnosis of Alzheimer's disease [28]. Carsten et al., have used Oscillating Search Algorithm Feature Selection (*OSAF*) and *SVM* to reduce features and classification of urinary *RNA* metabolites datasets of breast cancer women. The experimental results have shown sensitivity by 83.5% and specificity by 90.6% and the proposed method has performed well in diagnosis of breast cancer [29]. Peter and Somasundaram have proposed *CFS* with Bayes Theorem as a hybrid feature selection method for classification of heart disease. In this regard, some feature selection evaluators such as *CfsSubsetEval, FilterSubsetEval,* Chi-squared attribute evaluation, Gain ratio attribute evaluation and the hybrid *CFS* with *FilteredSubsetEval* were used. In addition, Naïve Bayes, *SVM*, *KNN* and Multi-Layer Perceptron were used for classification. Results have shown that the proposed feature selection method have increased the accuracy rate of all used classifiers by average accuracy rate of 84.07%. Moreover, the accuracy rate of *KNN* classifier has been increased (from 75.18% to 85.55%) more than the accuracy rate of other used classifiers after using the proposed hybrid *CFS* with Bayes Theorem [30].

Feature subset selection evaluators

Each feature selection method uses feature subset evaluators and searching algorithms to generate the best subset of dataset. In this section, brief descriptions of each evaluators and searching methods that we have used for feature selection are given below:

- *Classifier subset evaluator (ClassifierSubsetEval)* - Evaluates the subsets of features by using a classifier on the training data or on a separate hold out testing subset to find the merit subset of attributes [31].
- *Wrapper subset evaluator (WrapperSubsetEval)* - Uses a classifier to evaluate variable sets and employs cross validation to estimate the accuracy of learning scheme for each set [31].
- *Correlation feature selection subset evaluator (CfsSubsetEval)* – Ranks and selects feature subsets according to both their weak correlation among themselves and the highest correlation with the class to find the correlated features [32].
- *Filtered subset evaluator (FilterSubsetEval)* – Evaluates the values of feature subsets by predicting the redundancy degree among them to select the best suitable subset of features [33].
- *Best First search* - Selects the node that has the best score by using hill climbing augmented with a backtracking facility. It allocates the score for each candidate node by using an evaluation function. There are two lists of nodes. One list includes nodes, which will be explored (open nodes), and another list includes the nodes which were visited (close nodes). The search may start forward search by empty attribute set or start backward search by full attribute set, or start backward and forward search from any node. Open nodes are in priority queue when they are close to the goal. Best First algorithm searches unvisited nodes to choose the best of all nodes instead of using a small subset of neighbor nodes. Searching will not stop when the algorithm achieves a dead-end node. Thus, it is trying to search the best node between the other nodes [34].
- *Greedy stepwise search* - Starting from the empty set, it selects variables by forward selection and eliminates useless variables by backward selection to find the best feature subset. During searching process, new collection of candidate feature subsets was created by adding other features to the available best feature subset. The best feature subset was chosen after evaluating all subsets. The algorithm continues until the new generated collection of subsets does not surpass the best current subset [35, 36].

Advantages and disadvantages of each feature selection methods were shown in Table 1. A brief description of different feature selection evaluators and classifiers on literature was shown in Table 2. In this study, by considering all risks and advantages of using each methods (Table 1) the filter method which could achieve the best subset has been used. Moreover by evaluating different classification algorithms and different feature subset evaluators (Table 2) it could be demonstrated that different methods on different datasets could be used for selecting the best subset. Therefore the accuracy rate of classification algorithms could be increased after using the accurate feature selection method. It is noticeable that the best method for one dataset would not be the best method for other dataset.

Support Vector Machine (*SVM*)

SVM has good capability in classification and prediction problems. It can increase the generalization performance by handling nonlinear classifications with mapping the inputs into

Table 1 Advantages and disadvantages of feature selection methods

Methods	Advantages	Disadvantages
Filter	Fast, Scalable, Independent of the classifier, Better computational complexity	Ignores interaction with classifier
Wrapper	Simple, Interacts with the classifier, Good classification accuracy models feature dependencies	Risk of over fitting, Computationally intensive
Embedded	Interacts with the classifier, Better computational complexity than wrapper methods, Models feature dependencies	Classify with dependent selection

high dimensional areas and solving the of quadratic programming optimization problem. It can discover the optimum disjunctive hyperplane. By considering training samples (Eq. 1), each disjunctive hyperplane must prepare two conditions (Eq. 2) for two classes.

$$\left((xi, yi)|xi{\in}R^{N}, yi{\in}\{-1, 1\}, i = 1, \dots, n\right) \tag{1}$$

$$\begin{gathered}\{\,(W.xi) + b \gg +\, 1{-}\varepsilon_i\ , \text{if } yi = +1 \\ \text{and} \\ (W.\ xi) + b \gg -1 + \varepsilon_i\ , \text{if } yi = -1\ \}\end{gathered} \tag{2}$$

Inequalities of Eq. 2 are similar to Eq. 3. Minimizing Eq. 4 to Eq. 3 can improve the hyperplane separation.

$$yi[(W.xi) + b] \gg +\, 1{-}\varepsilon_i\ , i = 1, \dots, n \tag{3}$$

Table 2 Accuracy rate of different classifiers with feature selection evaluators on different datasets

Dataset	Feature selection evaluators	Classification algorithm	Accuracy (%)	Ref.
Tweeter dataset	*CfsSubsetEval*	Naïve Bayes	89.04	[26]
	ClassifierSubsetEval	SVM	87.98	
Hepatotoxic compounds of traditional Chinese medicine dataset	*CfsSubsetEval*	SVM	82.41	[20]
Heart disease dataset	*CfsSubsetEval*	Naïve Bayes, SVM, KNN, Multi-Layer Perceptron	Average 81.74	[30]
	FilterSubsetEval	Naïve Bayes, SVM, KNN, Multi-Layer Perceptron	Average 80.91	
	CFS + FilteredSubsetEval	Naïve Bayes, SVM, KNN, Multi-Layer Perceptron	Average 83.62	
	CFS + Bayes Theorem	Naïve Bayes, SVM, KNN, Multi-Layer Perceptron	Average 84.07	
	FilterSubsetEval	Naïve Bayes, SVM, KNN, Multi-Layer Perceptron	Average 81.01	
	Chi-squared attribute evaluation	Naïve Bayes, SVM, KNN, Multi-Layer Perceptron	Average 78.97	
	Gain ratio attribute Evaluation	Naïve Bayes, SVM, KNN, Multi-Layer Perceptron	Average 78.60	
	CFS + Bayes Theorem	KNN	85.55	
	Without feature selection	KNN	75.18	
Breast cancer	Without feature selection	Decision tree- CART	69.23	[19]
	CfsSubsetEval		71.32	
	FilterSubsetEval		71.32	
	SVM attribute evaluator		73.07	
Breast cancer Wisconsin (Original)	Without feature selection		94.84	
	CfsSubsetEval		94.84	
	ClassifierSubsetEval		95.13	
	PrincipalComponentsAttributeEval		96.99	
Breast cancer Wisconsin (Diagnostic)	Without feature selection		92.97	
	ClassifierSubsetEval		94.05	
	SymmetricUncertAttributesetEval		94.72	

$$\frac{1}{2}\ \|W\|^2 + C\textstyle\sum_{i=1}^{n}\varepsilon_i \tag{4}$$

In Eq. 4, C parameter can control the balance among complication and accuracy of classifier. In this regard, Lagrange multipliers can find the solution for convex optimization problem and by using suitable replacement it can obtain optimized solution for Eq. 4. Decision function will be provided in Eq. 6 by Lagrange multiplies given in Eq. 5 [37].

$$\text{Maximize}: \textstyle\sum_{i=1}^{n} \propto i - \frac{1}{2}\ \sum_{i.j=1}^{n} \propto i \propto j\ yi\ yj(xi, xj) \tag{5}$$

$$\left\{\mathrm{Subject\ to}: \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} \propto \mathrm{i\ yi} = 0, \propto \mathrm{i} \gg 0, \forall \mathrm{i},\ \mathrm{f(x)} = \mathrm{sgn}\ \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \propto \mathrm{i\ yi(xi, x)} + \mathrm{b}\right)\right\} \tag{6}$$

Methodology

In this study, the data set of *CKD* from *UCI* machine learning repository was used. The dataset includes 24 features excluding the class attribute. A total of 400 instances are included in 250 with *CKD* and 150 instances without *CKD* (*NotCKD*). Attributes of *CKD* dataset were shown in Table 3.

Wrapper and filter methods based on Best First and Greedy stepwise search were developed to evaluate the feature selection methods and the accuracy of classification algorithms. In this regard, the dataset was classified by *SVM* classification algorithm for the diagnosis of *CKD*; afterward, two methods of wrapper approach and two methods of filter approach were used to feature selection. These methods can reduce dimensional of dataset and they can get higher accuracy of classification in a shorter time. For each subset obtained by feature selection methods, *SVM* algorithm was utilized to diagnose *CKD* and *NotCKD*. Comparison of results will be presented in detail in section 6. In each used evaluator and searching method, the purpose was to find the best methods to obtain highest results for the best diagnosis of *CKD* and *NotCKD*. *ClassifierSubsetEval* with Greedy stepwise search engine and *WrapperSubsetEval* with Best First search engine were used on entire training dataset for the wrapper method. Correlation feature selection subset evaluator *(CfsSubsetEval)* with Greedy stepwise search engine and *FilterSubsetEval* evaluator with Best First search engine were used on full training set for filter method. After performing feature selection methods and reducing dimensional of dataset, *SVM* algorithm was used to classify and diagnose *CKD*. 10 fold cross validation was used in training phase of classification algorithm. In this study, WEKA (version 3.6.13) was used for feature selection and classification.

Table 4 Confusion matrix

Confusion Matrix		Prediction	
		Positive	Negative
Actual	Positive	*TP*	*FN*
	Negative	*FP*	*TN*

Experimental test results

Performance metrics

The confusion matrix is used to describe the performance of classification algorithms by calculating performance metrics

Table 3 Attributes of *CKD* dataset

Attributes of CKD dataset	Explanation	Attributes of CKD dataset	Explanation
age	The age of patient	bacteria	It is a sign of infection in the kidney [41]
red blood cells	How much oxygen tissues could receive [42]	albumin	It is amount of protein in the blood [43]
blood urea	Amount of urea nitrogen in blood [42]	sodium	It indicates a level of sodium in blood [42]
packed cell volume	The volume of the blood cells in a blood circulating [42]	red blood cell count	Indicates amount of red blood cells
coronary artery disease	Diagnose of coronary artery disease which affects the kidneys [42]	pedal edema	Swelling of legs [42]
blood pressure	It has a standard range to find out how blood pressure is high or low [42]	potassium	Level of potassium in urine [42]
pus cell urine	It could indicate infection in the kidney [44]	hypertension	Indicates the high blood pressure [42]
serum creatinine	It indicate the level of creatinine in blood [45]	anemia	Indicates the low level of red blood cells [42]
white blood cell count	Indicates amount of white blood cells	sugar	Level of sugar, which leaks out of blood to urine [42]
appetite	Level of patient's appetite [46]	blood glucose random	Indicates level of glucose in the blood [42]
specific gravity	Concentration of chemical particles in the urine [47]	hemoglobin	Amount of protein in red blood cells [48]
pus cell clumps	It indicates inflammation urine [49]	diabetes mellitus	Indicates the high level of blood sugar [42]

Table 5 Confusion matrix of *SVM* on full dataset and reduced dataset by feature selection methods

Confusion Matrix	Actual	Prediction	
		CKD	NotCKD
SVM without feature selection	CKD	241	9
	NotCKD	0	150
SVM with *ClassifiersubsetEval* with Greedy stepwise	CKD	245	5
	NotCKD	3	147
SVM with *WrapperSubsetEval* with Best first	CKD	246	4
	NotCKD	3	147
SVM with *CfsSubsetEval* with Greedy stepwise	CKD	243	7
	NotCKD	0	150
SVM with *FilterSubsetEval* with Best first	CKD	244	6
	NotCKD	0	150

(Table 4). In this study for measuring the performance of used methods following metrics have been used.

- *True Positive (TP)* - Indicates positive instances correctly classified as positive outputs
- *True Negative (TN)* - Indicates negative instances correctly classified as negative outputs
- *False Positive (FP)* - Indicates negative instances wrongly classified as positive outputs
- *False Negative (FN)* - Indicates positive instances wrongly classified as negative output
- *Classification Accuracy* - Indicates the ability of classifier algorithm to diagnose of classes of dataset (Eq. 7).

$$\text{Accuracy} = \frac{TP + TN}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{7}$$

- *Recall or Sensitivity* - Indicates the accuracy measure of the target class's occurrence (Eq. 8).

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN} \tag{8}$$

True Positive Rate (Eq. 9), False Positive Rate (Eq. 10), Precision (Eq. 11) and F-Measure (Eq.12) are the other metrics calculated by values of confusion matrix.

$$\text{True Positive Rate } (TPR) = \frac{TP}{TP + FN} \tag{9}$$

$$\text{False Positive Rate } (FPR) = \frac{FP}{FP + TN} \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$F\text{−Measure} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{12}$$

- *Receiver Operating Characteristic (ROC) Curve* - The *ROC* curve is a tool for evaluating the test results. It is defined by two dimensional graph (*TPR* as Y-axis and *FPR* as X-axis). The area under the *ROC* curve (*AUC*) presents the accuracy performance of a test to distinguish diagnostic groups or classes. The value of *AUC* ranges from 0 (the classifier diagnosed all classes incorrectly) to 1 (diagnostic performance between classes is perfect) [38, 39].

Test results

Feature selection methods which include filter and wrapper methods were used to reduce the dimensional of dataset and improve the accuracy of *CKD* diagnosis. All the used methods can produce a new dataset by lower dimensional than the original dataset. The used *CKD* dataset had 25 attributes. The first method of wrapper approach *ClassifierSubsetEval*, evaluator with Greedy stepwise search engine, reduced dataset dimension to 7 attributes. The second used method of wrapper approach, *WrapperSubsetEval* with Best First search engine reduced dataset dimension to 11 attributes. The first used method of filter approach, *CfsSubsetEval* with Greedy stepwise search engine reduced the dataset dimension to 16 attributes. The second method of filter approach FilterSubsetEval with Best First search engine, reduced the dataset dimension to 13 attributes. The confusion matrix of *SVM* classifier on original dataset of *CKD* without any feature selection and with all used feature selection methods are shown in Table 5. Table 5 indicates that without any feature selection, *SVM* algorithm could classify all *FP* (instances without *CKD*) correctly. However, it has a high number of *FN* (diagnosis of *CKD* instances are incorrectly 9).

SVM classifier on reduced dataset by *ClassifierSubsetEval* with Greedy stepwise feature selection has got 3 for *FP* and 5 for *FN*, and *SVM* classifier on reduced dataset by *WrapperSubsetEval* with Best First search engine feature selection, has got 3 for *FP* and 4 for *FN*. Therefore, *CKD*

Table 6 Accuracy of *SVM* classification without feature selection methods

SVM classification without feature selection	TP Rate	FP Rate	Precision	Recall	ROC Area	Class
	0.964	0	1	0.964	0.982	CKD
	1	0.036	0.943	1	0.982	NotCKD
Weighted Avg.	0.978	0.013	0.979	0.978	0.982	

Table 7 Accuracy of *SVM* classification with feature selection methods

SVM classification by using feature selection methods	TP Rate	FP Rate	Precision	Recall	ROC Area	Class
ClassifierSubsetEval with Greedy stepwise search engine	0.98	0.02	0.988	0.98	0.98	CKD
	0.98	0.02	0.967	0.98	0.98	NotCKD
Weighted Avg.	0.98	0.02	0.98	0.98	0.98	
WrapperSubsetEval with Best first search engine	0.984	0.02	0.988	0.984	0.982	CKD
	0.98	0.016	0.974	0.98	0.982	NotCKD
Weighted Avg.	0.983	0.019	0.983	0.983	0.982	
CfsSubsetEval with Greedy stepwise search engine	0.972	0	1	0.972	0.986	CKD
	1	0.028	0.955	1	0.986	NotCKD
Weighted Avg.	0.983	0.011	0.983	0.983	0.986	
FilterSubsetEval with Best first search engine	0.976	0	1	0.976	0.988	CKD
	1	0.024	0.962	1	0.988	NotCKD
Weighted Avg.	0.985	0.009	0.986	0.985	0.988	

diagnoses by wrapper feature selection approach has got high *FP*. Despite reducing feature dimensions by wrapper method, *SVM* classifier has fumbled, in diagnosis of all instances without *CKD* (3 for *FP*). Moreover *SVM* classifier on reduced dataset by *CfsSubsetEval* with Greedy stepwise search feature selection has got 7 for *FN* (*CKD* instances have diagnosed incorrectly)*; FilterSubsetEval* with Best First search engine has got 6 for *FN*, and both of them could diagnose all *CKD* correctly and *FP* is 0. By comparison of *FN* values of all used methods, the value of *FN* in *SVM* classifier by using *WrapperSubsetEval* with Best First search engine feature selection method is less than the other used methods. However, it could not be the best method, because of its high *FP* result.

The accuracy of *SVM* classifier without and with the used feature selection methods are shown in Table 6 and Table 7. For each *CKD* and *NotCKD* classes recall, precision, *FP* rate, *TP* rate and the value in the area under the *ROC* curve were calculated. For each of the previously noticed measures weighted average claculated by Eq. 13 (Tables 6 and 7). In Eq. 13, n is the number of classes, A_i is the value of each calculated parameters for i^{th} class, C_i is the instances count for i^{th} class and N is the total count of instances in the dataset [40].

$$\text{Weighted Avg} = \frac{1}{N}\sum_{i=1}^{n} A_i \times C_i \tag{13}$$

It is evident from Table 6 and Table 7 that the accuracy value of *CKD* diagnosis of *SVM* algorithm on reduced dataset by *FilterSubsetEval* with Best First search engine feature selection are most acceptable. It has the highest weighted average value of *TP* rate and the lowest weighted average value of *FP* rate. In addition it has the highest value of precision, recall and the highest value in the area under the *ROC* curve. It is noticeable that the lowest value of *FP* rate (0.009) and the highest value of *TP* rate (0.985) cause to increase the Precision and recall values in Table 6 and Table 7.

In *ROC* area if *AUC* is equal to one or near to one it means the classifier could diagnose the classes of dataset perfectly. Therefore, *SVM* classifier on reduced dataset by *FilterSubsetEval* with Best First search engine has higher value in area under the *ROC* curve (0.988) in both *CKD* and *NotCKD* classes than the other used methods (Tables 6 and 7).

The summary of *SVM* classifier results value with and without using feature selection methods are shown in Table 8. *SVM* classifier, without using feature selection, has the least accuracy rate (97.75%) in diagnosis of *CKD*. From Table 8, it can be demonstrated that the accuracy rate of SVM classifier by using *WrapperSubsetEval* with Best First and *CfsSubsetEval* with Greedy stepwise feature selection methods were the same by 98.25%. Despite the lowest reduced dimension by *ClassifierSubsetEval* with Greedy stepwise search engine (7 attributes selected from 25 attributes),

Table 8 Summary of *SVM* classifier results value with and without using feature selection methods

Method	Incorrectly Classified Instances (%)	Number of Features	Accuracy Rate (%)
SVM without feature selection	2.25	25	97.75
SVM with *ClassifierSubsetEval* with Greedy stepwise	2	7	98
SVM with *WrapperSubsetEval* with Best First	1.75	11	98.25
SVM with *CfsSubsetEval* with Greedy stepwise	1.75	16	98.25
SVM with *FilterSubsetEval* with Best First	1.5	13	98.5

SVM classifier has an accuracy rate of 98%. It is higher than the accuracy rate of *SVM* for 25 dimensions of dataset. However, it is not the best feature selection method because other used methods have higher accuracy rates. Finally, *SVM* classifier on dataset whose dimension has been reduced by using *FilterSubsetEval* with Best First search method (13 selected attributes), has the highest accuracy rate (98.5%). In addition, it has got the lowest incorrectly classified instances (1.5%).

Conclusions

In this study, wrapper and filter methods have been utilized on data set of *CKD*. Two different evaluators have been used for each method. For filter approach, *CfsSubsetEval* with Greedy stepwise search engine and *FilterSubsetEval* with Best First search engine have been used. In addition to wrapper approach, *ClassifierSubsetEval* with Greedy stepwise search engine and *WrapperSubsetEval* with Best First search engine have been used. The accuracy rate of *SVM* classifier on full training set has been compared with its accuracy rate on 4 reduced datasets which have been gained by feature selection methods. The results show that after reducing dimension of *CKD* dataset, in all 4 methods accuracy rate of diagnosis have been improved. The accuracy rate of *SVM* classification on reduced dataset by *FilterSubsetEval* with Best First search engine (98.5%) is more than the other used methods. It has obtained the highest values of other comparable results such as *TP* rate, correctly classified instances. Moreover, it has lowest number of important values such as incorrectly classified instances and *FP* rate. It is noticeable that, preparing the dataset with the lowest dimension by feature selection methods could not lead to the highest accuracy rate of classification in perpetuity. For instance, the accuracy rate of *SVM* on the lowest dimension of *CKD* dataset by 7 attributes by *ClassifierSubsetEval* with Greedy stepwise search engine is not the highest accuracy rate (98%), however the accuracy rate of *SVM* classifier on 13 attributes of *CKD* dataset, by using *FilterSubsetEval* with Best First feature selection method, has got the most accuracy rate (98.5%) in *CKD* diagnosis. Furthermore, with different methods of feature selection and classification algorithms, on distinct datasets of disease, classification results can be different in accuracy rates.

References

1. Nordqvist, C., *Chronic kidney disease: causes, symptoms and treatments*. IOP Publishing medicalnewstoday, 2016 http://www.medicalnewstoday.com/articles/172179.php. Accessed 14 Jan 2016.
2. Go, A.S., Chertow, G.M., Fan, D., McCulloch, C.E., and Hsu, C.-y., Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *N. Engl. J. Med.*, 2004. doi:10.1056/NEJMoa041031.
3. Kathuria, P., and Wedro, B., *Chronic kidney disease quick overview*. IOP Publishing emedicinehealth, 2016 http://www.emedicinehealth.com/chronic_kidney_disease/page2_em.htm#chronic_kidney_disease_quick_overview. Accessed 23 Feb 2016.
4. Huang, M.-J., Chen, M.-Y., and Lee, S.-C., Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Syst. Appl.* 32:856–867, 2007. doi:10.1016/j.eswa.2006.01.038.
5. José, N., Rosário Martins, M., Vilhena, J., Neves, J., Gomes, S., Abelha, A., Machado, J., and Vicente, H., A soft computing approach to kidney diseases evaluation. *J. Med. Syst.* 39:131, 2015. doi:10.1007/s10916-015-0313-4.
6. Di Noia, T., Claudio, V., Ostuni, F.P., Binetti, G., Naso, D., Schena, F.P., and Di Sciascio, E., An end stage kidney disease predictor based on an artificial neural networks ensemble. *Expert Syst. Appl.* 40:4438–4445, 2013. doi:10.1016/j.eswa.2013.01.046.
7. Chen, Z., Zhang, X., and Zhang, Z., Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. *Int. Urol. Nephrol.* 48:2069–2075, 2016. doi:10.1007/s11255-016-1346-4.
8. Akay, M.F., Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst. Appl.* 36:3240–3247, 2009. doi:10.1016/j.eswa.2008.01.009.
9. Özçift, A., and Gülten, A., Genetic algorithm wrapped Bayesian network feature selection applied to differential diagnosis of erythemato-squamous diseases. *Digital Signal Processing.* 23: 230–237, 2013. doi:10.1016/j.dsp.2012.07.008.
10. Singh, R.K., and Sivabalakrishnan, M., Feature selection of gene expression data for cancer classification: a review. *Procedia Computer Science.* 50:52–57, 2015. doi:10.1016/j.procs.2015.04.060.
11. Chao-Ton, S., and Yang, C.-H., Feature selection for the SVM: an application to hypertension diagnosis. *Expert Syst. Appl.* 34:754–763, 2008. doi:10.1016/j.eswa.2006.10.010.
12. Kumari, B., and Swarnkar, T., Filter versus wrapper feature subset selection in large dimensionality micro array: a review. *International Journal of Computer Science and Information Technologies.* 2(3):1048–1053, 2011.
13. Villacampa, O., *Feature selection and classification methods for decision making: a comparative analysis*. CEC Theses and Dissertations. College of Engineering and Computing. Nova Southeastern University, Florida, USA, 2015.
14. Karegowda, A.G., Jayaram, M.A., and Manjunath, A.S., Feature subset selection problem using wrapper approach in supervised learning. *Int. J. Comput. Appl.* 1(7):13–17, 2010. doi:10.5120/169-295.
15. Cho, B.H., Yu, H., Kim, K.-W., Kim, T.H., Kim, I.Y., and Kim, S.I., Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. *Artif. Intell. Med.* 42:37–53, 2008. doi:10.1016/j.artmed.2007.09.005.
16. Ladha, L., and Deepa, T., Feature selection methods and algorithms. *Int. J. Comput. Sci. Eng.* 3(5):1787–1797, 2011.
17. Mousin, L., Jourdan, L., Marmion, M.-E., and Dhaenens, C., Feature selection using tabu search with learning memory: learning Tabu Search. 10th International Conference. LION 10. Ischia, Italy, 2016. doi:10.1007/978-3-319-50349-3_10.
18. Ma, S., and Huang, J., Penalized feature selection and classification in bioinformatics. *Brief. Bioinform.* 9:392–403, 2009. doi:10.1093/bib/bbn027.

19. Lavanya, D., and Usha Rani, K., Analysis of feature selection with Classfication: breast cancer datasets. *Indian Journal of Computer Science and Engineering (IJCSE).* 2(5):756–763, 2011.

20. Jiang, L., He, Y., and Zhang, Y., Prediction of hepatotoxicity of traditional Chinese medicine compounds by support vector machine approach. The 8th International Conference on Systems Biology (ISB). Qingdao, China, 2014. doi:10.1109/ISB.2014.6990426.

21. Sasikala, S., Appavu alias Balamurugan, S., and Geetha, S., A novel feature selection technique for improved survivability diagnosis of breast cancer. *Procedia Computer Science.* 50:16–23, 2015. doi:10.1016/j.procs.2015.04.005.

22. Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S., and Weaver, N., Center for applied internet data analysis. IEEE Security and Privacy article, 2003. http://www.caida.org/publications/papers/2003/sapphire/. Accessed 2 Feb 2017.

23. Poore, K., Nimda worm–why is it different?. SANS Institute, 2001. http://www.sans.org/reading-room/whitepapers/malicious/nimda-worm-different-98. Accessed 2 Feb 2017.

24. Center for Applied Internet Data Analysis., UCSD network telescope – code-red worms dataset. Center for Applied Internet Data Analysis, 2016. http://www.caida.org/data/passive/codered_worms_dataset.xml. Accessed 2 Feb 2017.

25. Ćosović, M., Obradović, S., and Trajković, L., Performance evaluation of BGP anomaly classifiers. *IEEE.*, 2015. doi:10.1109/DINWC.2015.7054228.

26. Akbarisanto, R., Akbarisanto, R., and Purwarianti, A., Analyzing bandung public mood using twitter data. Fourth International Conference on Information and Communication Technologies (ICoICT). Bandung, Indonesia, 2016. doi:10.1109/ICoICT.2016.7571910.

27. Wang, Y., Maa, L., and Liu, P., Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine. *Comput. Methods Prog. Biomed.* 95:249–257, 2009. doi:10.1016/j.cmpb.2009.03.004.

28. Chaves, R., Ramírez, J., Górriz, J.M., López, M., Salas-Gonzalez, D., Álvarez, I., and Segovia, F., SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting. *Neurosci. Lett.* 461:293–297, 2009. doi:10.1016/j.neulet.2009.06.052.

29. Henneges, C., Bullinger, D., Fux, R., Friese, N., Seeger, H., Neubauer, H., Laufer, S., Gleiter, C.H., Schwab, M., Zell, A., and Kammerer, B., Prediction of breast cancer by profiling of urinary RNA metabolites using support vector machine-based feature selection. *BMC Cancer.* 9:104, 2009. doi:10.1186/1471-2407-9-104.

30. John Peter, T., and Somasundaram, K., Study and development of novel feature selection framework for heart disease prediction. *Int. J. Sci. Res. Publ.* 2(10):577–583, 2012.

31. Randa Oqab Mujalli, de Juan Oña (2011) A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. J. Saf. Res. 42: 317–326. doi:10.1016/j.jsr.2011.06.010

32. Onik, A.R., Haq, N.F., Alam, L., and Mamun, T.I., An analytical comparison on filter feature extraction method in data mining using J48 classifier. *Int. J. Comput. Appl.* 124(13):1–8, 2015.

33. Yeom, J.S., *Textile fingerprinting for dismount analysis in the visible, near, and shortwave infrared domain*. Thesis. Department of The Air Force. Air Force Institute of Technology. Wright-Patterson Air Force Base, Ohio, USA, 2014.

34. Dechter, R., and Pearl, J., Generalized best-first search strategies and the optimality of a*. *J. Assoc. Comput. Mach.* 32(3):505–536, 1985.

35. Sadeghi, R., Zarkami, R., Sabetraftar, K., and Van Damme, P., Application of genetic algorithm and greedy stepwise to select input variables in classification tree models for the prediction of habitat requirements of *Azolla filiculoides* (lam.) in Anzali wetland, Iran. *Ecol. Model.* 251:44–53, 2013. doi:10.1016/j.ecolmodel.2012.12.010.

36. Wald, R., Khoshgoftaar, T.M., and Napolitano, A., *Optimizing wrapper-based feature selection for use on bioinformatics data*. In Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, Florida, USA, 2014.

37. Xie, J., and Wang, C., Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Syst. Appl.* 38:5809–5815, 2011. doi:10.1016/j.eswa.2010.10.050.

38. Fawcett, T., An introduction to ROC analysis. *Pattern Recogn. Lett.* 27:861–874, 2006. doi:10.1016/j.patrec.2005.10.010.

39. Hajian-Tilaki, K., Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med.* 4(2):627–635, 2013.

40. V. Mohan Patro, Manas Ranjan Patra (2014) Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy. Ransactions on Machine Learning and Artificial Intelligence. 2(4): 77–91. doi:10.14738/tmlai.24.328

41. MAYO CLINIC., Kidney infection. MAYO CLINIC, 2016. http://www.mayoclinic.org/diseases-conditions/kidney-infection/basics/definition/con-20032448. Accessed 2 Feb 2017.

42. Healthline., Red Blood Cell Count (RBC). Healthline. http://www.healthline.com/health/rbc-count#Overview1, 2016. Accessed 2 Feb 2017.

43. DPC Education Center., Albumin and Chronic Kidney Disease. DPC Education Center, 2016. http://www.dpcedcenter.org/albumin-and-chronic-kidney-disease. Accessed 2 Feb 2017.

44. NLDA., Pus cells in urine: causes, symptoms, treatment and best home remedies. NLDA, 2016. https://www.nlda.org/pus-cells-in-urine-causes-symptoms-treatment-and-best-home-remedies/. Accessed 2 Feb 2017.

45. Charles Patrick Davis., Creatinine blood test. MedicineNet.com, 2016. http://www.medicinenet.com/creatinine_blood_test/page2.htm. Accessed 2 Feb 2017.

46. DAVITA., Stage 4 of chronic kidney disease (CKD). DAVITA, 2016. https://www.davita.com/kidney-disease/kidney-disease/symptoms-and-diagnosis/stage-4-of-chronic-kidney-disease-(ckd)/e/686. Accessed 2 Feb 2017.

47. Medline plus., Urine specific gravity test. Medline plus, 2015. https://medlineplus.gov/ency/article/003587.htm. Accessed 2 Feb 2017.

48. DPC Education Center., What you need to know about anemia and kidney disease. DPC Education Center, 2016. http://www.dpcedcenter.org/what-you-need-know-about-anemia-and-kidney-disease. Accessed 2 Feb 2017.

49. Medical-base.com., Pus cell in urine–causes, symptoms & treatment of pus cells. Medical-base.com, 2016. http://medical-base.com/pus-cell-in-urine-causes-symptoms-treatment-of-pus-cells. Accessed 2 Feb 2017.