Early Prediction of Chronic Kidney Disease Using Multiple Automated Techniques
Haya Alaskar and Amal Zaki
Pages 26 – 32
DOI: http://dx.doi.org/10.21700/ijcis.2017.xxx

Early Prediction of Chronic Kidney Disease Using Multiple Automated Techniques

Haya Alaskar Amal Zaki
Prince Sattam Bin Abduaziz University, KSA
h.alaskar@psau.edu.sa

Abstract: *Early prediction and protection can keep chronic kidney disease (CKD) from getting inferior. Advancement in Technology offers several automated techniques and algorithms to analyze and classify the test results obtained from the patient medical report.The task of automated technique is to make use of historical data, to discover regular patterns and improve future decisions for early cure of CKD patient. In this work various techniques and algorithms such as machine learning techniques and data mining algorithms have been implemented for classification of patients medical information results for improvement of diagnosis accuracy. Comparison of classification results is carried out to show the most accurate and sensitive technique for early predection of CKD according to its sternness. In this research a 24 attributes (features) have been selected for this experiment. Simulation results are presented for 12 classifiers on the basis of its accuracy, sensitivity, specificity and execution time for CKD prediction. It is shown that the naïve beyes classifier has the highest performance with 100% accuracy. In the meanwhile a data clustering algorithm is implemented to cluster the data sets of patients to five groups which are correlated to the stage of CKD. Prediction of the level or stage of CKD is achieved by utilizing pack-propagation Neural Networks (NN) as a classifier.*

Keywords: *Atomated classification techniques; Chronice kidney disease; Machine learning techniques; Data mining algorithms; data processing and clustering algothims; Pack propagation neural network.*

Received: *July 30, 2016* | **Revised**: *September 10, 2016* | **Accepted**: *January 05, 2017*

1. Introduction

Chronic kidney diseases have become a major public health dilemma. Chronic kidney diseases account for 60% of all deaths worldwide. Eighty percentages of chronic disease deaths worldwide occur in third world countries [1]. Chronic kidney disease (CKD) is known as chronic renal disease. Chronic kidney disease involves situation that damage human kidney and diminish their ability to keep him healthy.

Automated techniques play a vital role in the field of health care applications. These techniques involve data mining techniques and machine learning algorithms.

Data mining techniques has been used as a modern trend for achieving diagnostics results, in medical applications such as cardiology arrhythmia, neurodegenerative diseases as well as kidney disease. Data mining concepts are used to examine a prosperous collection of data from various perspectives medical society and deriving useful information from huge data base [2].

Many researchers have done numerous researches for analysis and classification of CKD by extracting hidden information from large medical test reports from clinic and medical associations. S. Parul et. al [3] have made comparative study of predicting CKD utilizing artificial neural network (ANN) and supervised vector machine (SVM) algorithms.

In [4] the authors have implemented two neural networks techniques as well as WEKA tool to find the best technique among the above three algorithms for Kidney Stone Diagnosis. Their results showed that the back-propagation neural network (NN) significantly improves the conventional classification technique in medical applications.

L.Firmont et al [5] used data mining algorithm to perform exploratory analysis to evaluate factors influencing hemoglobin levels to control CKD.

S. Ramya et al [6] proposed work deals with classification of different stages in CKD using machine learning algorithm. Four classifiers techniques have been considered for predicting chronic kidney disease, Back propagation Neural Network, Radial Basi Function and Random Forest. The models are evaluated with four different measures like Kappa, Accuracy, Sensitivity and Specificity. From their experimental result, the Radial Basis Function is the better accuracy.

A. Vigil et al in [7] design a longitudinal, observational, and prospective cohorts study of a sample of 306 patients derived from primary care with the diagnosis of renal failure based on serum creatinine -estimated glomerular filtration rate eGFR < 90 mL/min/1.73 m2. Rapid kidney function decline is defined as an annual $eGRF_{creat}$ loss > 4 mL/min/1.73 m^2. Patient's data was grouped according to the rate of kidney function decline based on logistic regression model. From this study it is concluded that lower serum albumin, previous cardiovascular disease and higher proteinuria are the main predictors of rapid kidney function decline.

In [8] The medical data sets of CKD collected from Visakhapatnam district with 690 samples and 49 attributes have been analyzed using Weka and Orange CAD tools. Machine learning algorithms such as AD Trees, J48, Random forest, Naïve Bayes, K star are selected for statistical analysis and classification of CKD.

The main objective of this work is to predict kidney function failure as early as possible through the implementation of several automated techniques on the medical test result which are obtained from the medical reports of the patients. The work emphasize on the most efficient technique which leads to reduction of the error of diagnosis and to improve diagnostic accuracy and sensitivity through comparison between several classification techniques. New data sets (medical reports) for 400 samples have been collected from Apollo Hospitals in India with 24 features in each set [9]. Early predication of CKD can be achieved by clustering the data sets of patients to five groups which are correlated to stage of CKD by implementing pack propagation neural network as a classifier.

2. Methodology

The National Kidney foundation determines the different stages of chronic kidney disease based on the presence of kidney damage and glomerular filtration rate (GFR), which is the best overall index to measure the kidney function or determine the CKD stage. If the GFR number is low, this means that the kidneys are not working as well as they should. The earlier kidney disease is detected, the higher of the chance of slowing or stopping its progression. There are five stages of CKD. Table 1 shows the rat of GFR and corresponding stage to it [10].
There are several features that can lead to the determination of level or stage of CKD. These are age, blood pressure, albumin, sugar, red blood cells, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus coronary artery disease, appetite, pedal edema, anemia class count, hypertension, diabetes mellitus, specific gravity.

The methodology for detecting status of the patient is carried out through the following steps:

Step 1:

Dataset: The data sets for diagnosis of chronic kidney disease are obtained from medical reports of the patients collected from Apollo Hospitals, in India it has been provided by Dr P.Soundarapandian M.D and Senior Consultant Nephrologist. There are 400 instances with 24 different attributes related to kidney disease. The data are 400 samples, which divided into 250 samples have chronic kidney disease patients and 150 samples do not have chronic kidney disease persons. However, there are some Missing Attribute Values from medical information in some patients. Therefore, preprocessing methods has been implemented to remove the data set with missing values. Table 2 shows features information required to perform analysis and classification.

Step 2:

Classification: Using pr tools from matlab: PR Tools is a Matlab toolbox for pattern recognition that persistently provides libraries for new and developing statistical techniques. Therefore 12 algorithms from machine learning have been applied some of them are Normal Density Based Classification and other from Neural Networks (NN). These algorithms are implemented as classification task for accurate diagnosis of CKD based on different performance and evaluation measures. The machine learning algorithms utilized in this work are listed below.

Step 3:

Clustering: In this step the data sets of the patients whom are defined from the best classifier (second step) are clustered in to five groups. Each group will include the data sets that have matching features for number of patients. These features are correlated to the GRF factor which determines the stage of CKD. Classification process is performed to the five groups and decision is obtained for prediction of CKD level.

Table 1: Stages of CKD

Stage	CKD Description	Glomerular Filtration Rate(GFR)%
1	*Kidney damage (e.g., protein in the urine) with normal GFR*	90 or above
2	*Kidney damage with mild decrease in GFR*	60 to 89
3	*Moderate decrease in GFR*	30-59
4	*Severe reduction in GFR*	15-29
5	*Kidney failure*	Less than 15

Table 2: The features information (Attributes)

1.Age (numerical)	
2.Blood Pressure (numerical)	bp in mm/Hg
3.Specific Gravity (nominal)	sg - (1.005,1.010,1.015,1.020,1.025)
4.Albumin (nominal)	al - (0,1,2,3,4,5)
5.Sugar (nominal)	su - (0,1,2,3,4,5)
6.Red Blood Cells (nominal)	rbc - (normal,abnormal)
7.Pus Cell (nominal)	pc - (normal,abnormal)
8.Pus Cell clumps(nominal)	pcc - (present,notpresent)
9.Bacteria (nominal)	ba - (present,notpresent)
10.Blood Glucose Random(numerical)	bgr in mgs/dl
11.Blood Urea(numerical)	bu in mgs/dl
12.Serum Creatinine(numerical)	sc in mgs/dl
13.Sodium(numerical)	sod in mEq/L
14.Potassium(numerical)	pot in mEq/L
15.Hemoglobin(numerical)	hemo in gms
16.Packed Cell Volume(numerical)	
17.White Blood Cell Count(numerical)	wc in cells/cumm
18.Red Blood Cell Count(numerical)	rc in millions/cmm
19.Hypertension(nominal)	htn - (yes,no)
20.Diabetes Mellitus(nominal)	dm - (yes,no)
21.Coronary Artery Disease(nominal)	cad - (yes,no)
22.Appetite(nominal)	appet - (good,poor)
23.Pedal Edema(nominal)	pe - (yes,no)
24.Anemia(nominal)	ane - (yes,no)

3. Implementation

There are two experiments to examine the CKD data sets in order to reach specific early diagnosis. The first experiment outcome detects patients and inpatients through classification of the 400 samples data sets from medical test results. While the second experiment represents clustering of the data sets into five groups for only patients date sets that determined from first experiment one. These five groups are related to the stages presented in table 1. A measurement factor GRF value is correlated to each cluster of data sets. Through this experiment a second classifier is implemented to each group for prediction of the patient status and stage of illness. The experiments are explained in the following section.

3.1 Experiment 1

First: There are 400 instances (date sets) with 24 different attributes (features) related to kidney disease, which have been showed in table 2. The data sets are 400 samples, 250 samples have chronic kidney disease patients and 150 samples have not chronic kidney disease inpatients. However, there are some missing attribute values from medical information in some instances. The instances are analyzed and filtered from data sets with missing features information. In the preprocessing the data are filtered using Weka tool []. Therefore samples have been reduced to 153 samples where 43 are chronic kidney disease and 115 are not chronic kidney disease. Each data set has 24 attribute (features).

Second: Data mining algorithms are implemented for classification of data sets to determine the samples that corresponding to patients and samples refer to inpatients. The following algorithms are used to classify the medical test data sets 153 samples that obtained from preprocessing step.
The Linear and Quadratic Classifiers

1. IDC - Normal densities based linear (muli-class) classifier
2. QDC - Normal densities based quadratic (multi-class) classifier
3. KLLDC - Linear classifier based on KL expansion of common cov matrix
4. PCLDC - Linear classifier based on PCA expansion on the joint data
5. PARZENC - Parzen classifier
6. TREEC - Construct binary decision tree classifier
7. DTC - Decision tree classifier, rewritten, also for nominal features
8. NAIVEBE - Naive Bayes classifier
9. BPXNC - Feed forward neural network classifier by back propagation
10. IMNC - Feed forward neural network by Levenberg-Marquardt rule
11. NEURC - Automatic neural network classifier
12. SVC - Support vector classifier

Selection of the results from the best classifier is performed based on the accuracy, sensitivity and time. This output from the classifier will indicate the patient and inpatients samples.

3.2 Experiment 2

This experiment consists of two data processing techniques; these are clustering of data sets of 43 instances of patients and classification of each cluster for prediction of patient status (stage of CKD).
First: All the data sets are clustered in groups. Each group contains specific numbers of instance (data sets). Some features in each data set have correlated

function with GFR. These features are strongly indicating the level of GRF factor. Normal GFR varies according to age, sex, and body size, and declines with age. The Serum creatinine is a dominant feature in the data sets which gives an indication to the level or stage of CKD. This leads to have five groups corresponding to the five stages presented in table 1.

Second: in this process a classification technique is selected to predict the patient's stage of CKD. The pack propagation neural network (NN) technique is utilized to classify the instances in each group and output will be the level of CKD.

Cluster Analysis: is the chore of grouping a set of data in such a way that data in the same group or cluster are more similar in features to each other than to those in other groups (clusters). In data mining clustering is an essential process. The 43 instances refer to the patients are clustered in five group for statistical analysis using K- mean. In this algorithm clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the K cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized [11].

Classification: The classifier utilized in the CKD prediction is based on feed forward back-propagation NN [12]. The NN output indicates the stage of CKD of the sample provided in the input which represents the features of the instance belong to specific group.NN classifier is created using newff Matlab function. The traingd is NN training function that updates weight and bias values according to gradient descent, with number of epochs of 100000 used during the training phase. After the training is done the associated weights are calculated. The weights are adjusted to represents the GRF ranges. This leads to define correlation between dominant features and stage of CKD.

4. Simulation Results and Evaluations

4.1 Experiment 1 Results

The data sets are 400 samples, 250 samples have chronic kidney disease patients and 150 samples have not chronic kidney disease inpatients. The data sets are normalized and shown in figure 1. The red color represents the 250 instances of patients and the blue color represents the 150 instances for inpatients. The data set are then filtered from incomplete data sets using weka tool. Figure 2 represents 43 instances for patients (red color) and 115 instances for inpatients (blue color).

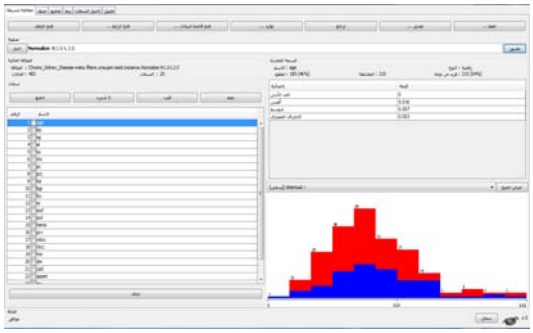

Figure 1. Normalized data sets (Instances)

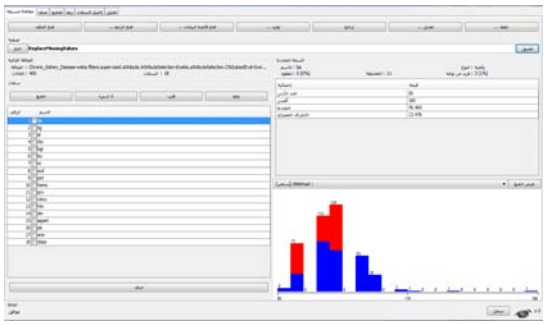

Figure 2. Filtered data sets (Instances)

The 12 algorithms have been implemented using PR Tools from MATLAB tool box. Comparison is done based on each classifier performance such as accuracy, sensitivity, specificity, and total time consumed during process. Figure 1 shows the 12 classifiers accuracy and figure 2 shows the sensitivity of these classifiers. Figure 3 shows the processing time for all classifiers. From the obtained results, it is found that naïvebc contributes the highest performance among all implemented classifiers so it is the best classifier used for detection of CKD. The correctly classified instances (43 patients) obtained (100%) accuracy. The incorrectly classified instances 0 (0%) out of the total instances (43 patients). The correctly classified instances (115 inpatients) obtained (100%) accuracy. The incorrectly classified instances 0 (0%) out of the total instances (115 inpatients). The kappa statistics is 1.The kappa (first syntax) calculates the kappa-statistic measure of interpreter agreement when there are two unique rates and two or more ratings.

The performance evaluation of the 12 classifiers is given in table 3. From table 3, it has been shown the exactness that obtained while running the various algorithms in PR tool box by matlab. The clarification that the best algorithm is naivebc for this Data set where accuracy, specificity, sensitivity are 1 and the ROC values 1 1(very accurate). The execution time is moderate for this algorithm.

Table 3: Performance Evaluation of 12 classifiers

	accuracy	**Errors**	**Specificities**	**Sensitiveities**	**time**
ldc	0.989	0.01087	1	0.958333	0.013485 seconds.
Qdc	0.913043	0.134783	1	0.483333	0.035919 seconds.
Parzenc	0.826087	0.215217	0.935294	0.358333	0.020611 seconds.
dtc	0.987	0.013043	0.997059	0.958333	0.055286 seconds.
pcldc	0.989	0.01087	1	0.958333	0.062611 seconds.
Klldc	0.989	0.01087	1	0.958333	0.005502 seconds.
Treec	0.956522	0.023913	0.994118	0.925	0.005713 seconds.
naivebc	1	0	1	1	0.481730 seconds.
bpxnc	0.989	0.01087	1	0.958333	0.385552 seconds.
lmnc	0.989	0.01087	1	0.958333	1.957721 seconds.
Neurc	0.956522	0.030435	0.985294	0.925	0.058013 seconds.
Svc	0.99	0.004348	1	0.983333	0.020014 seconds.

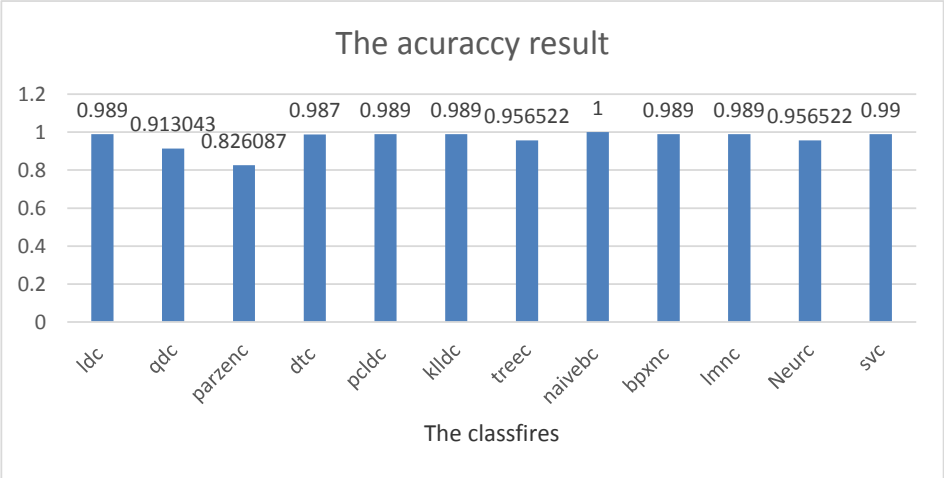

Figure 3. The classifiers accuracy performance.

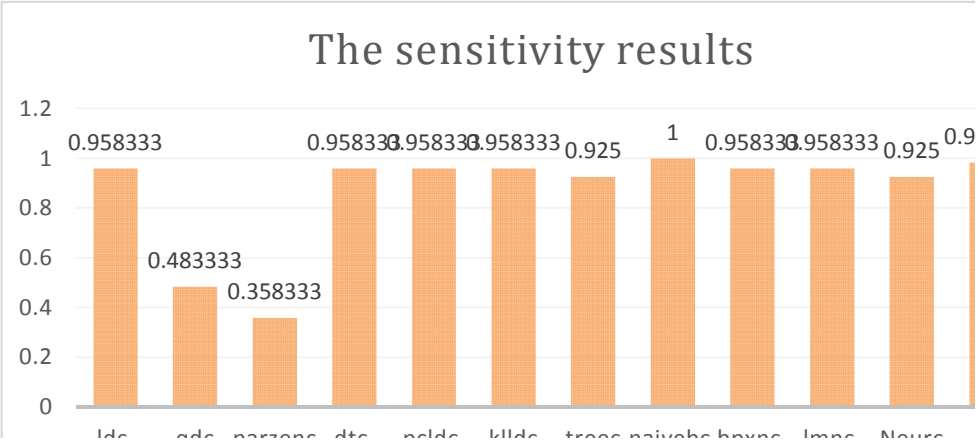

Figure 4. The classifiers sensitivity performance.

Early Prediction of Chronic Kidney Disease Using Multiple Automated Techniques
Haya Alaskar and Amal Zaki
Pages 26 – 32
DOI: http://dx.doi.org/10.21700/ijcis.2017.xxx

4.2 Experiment 2 Results

Clustering Results: The results of data sets clustering using K-mean algorithm are shown in figures 5, 6 and 7. The attribute standard of most dominants features of the data sets of the 43 patients are presented in figure 5. Figure 6 shows the clustering results of the data sets in five groups.

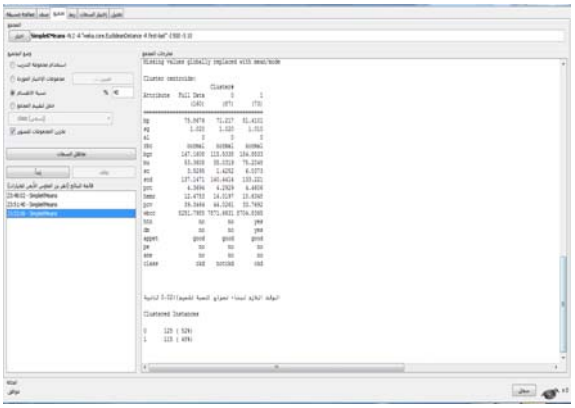

Figure 5. k-means in (40% of total samples)

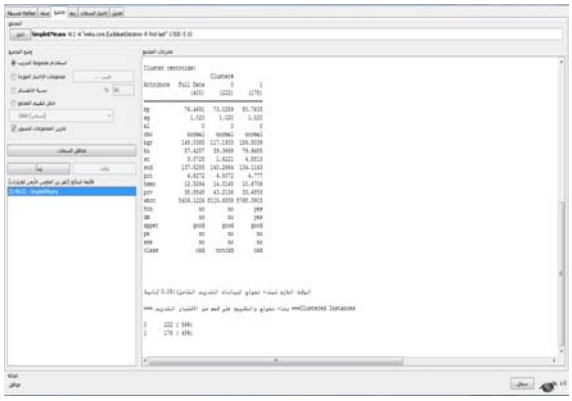

Figure 6. k-means in (100% of total samples)

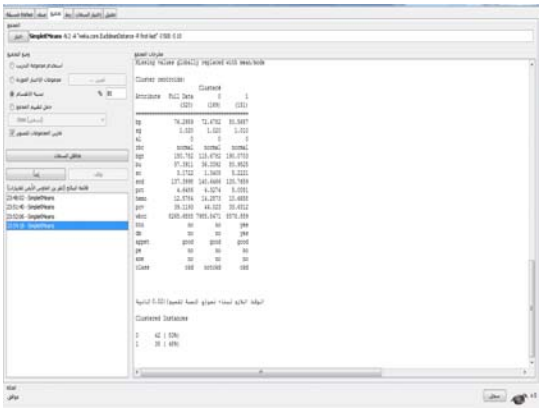

Figure 7. k-means in (80% of total samples)

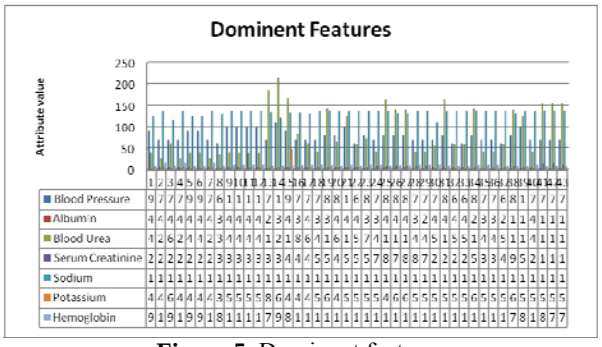

Figure 5. Dominent features

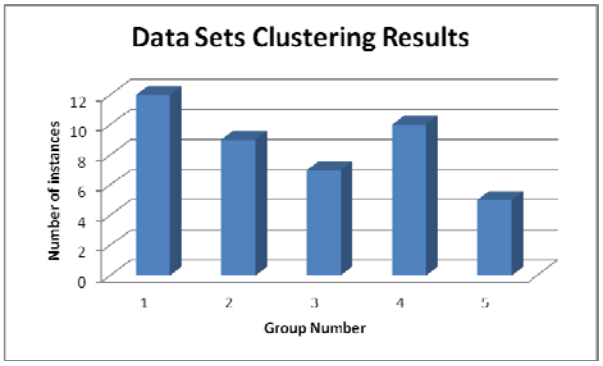

Figure 6. Five clustering results of 43 instances

Classifier Results: Matlab simulation for the pack propagation neural network is done by running the Matlab Simulink NN model and read the output data from Matlab workspace.

The NN has been trained on 24 instances (data sets) of patients and the validation sets used are 16 instances (data sets) of patients. These 19 instances are selected randomly from the five clusters separately. The NN classifier is formed using newff Matlab function to create a feed-forward back propagation network The traingd is the matlab function used to train NN. This training function updates weight and bias values according to gradient descent, with number of epochs of 100000 used during the training phase.

The simulation for the NN classifier is done for instance number 40 as a validation sample. It is used as input instance to the NN and the output is refers to the level or stage of CKD. Matlab simulation result for the neural network output for instance No 40 indicates that the input features to the neural network corresponding to a patient within stage 5 of CKD. Table 5 shows the Confusion matrix of the NN performance. NN classifier performance achieves total accuracy of 89.2 %.

Table 5: Confusion matrix

Number of Instances	*Stage #*	*Correct*	*Incorrect*	*Accuracy (%)*
6 validation sets	1	5	1	80%
4 validation sets	2	3	1	66%
3 validation sets	3	3	0	100%
4 validation sets	4	4	0	100%
2validation sets	5	2	0	100%
Total				89.2%

5. Conclusion

In this work; several automated techniques and algorithms are presented to analyze and classify the test results obtained from the CKD patients medical reports. Many classifiers have been implemented to classify the patients and inpatients from the data sets of the medical reports. New 400 data sets (samples) have been collected with 24 attributes (features) to examine the CKD for early diagnosis. The clarification that the best algorithm is naivebc for this Data set where accuracy, specificity, sensitivity are 1 and the ROC values 1 1(very accurate). For early prediction of the stage of CKD the data sets have been clustered in 5 groups according to the GRF stage. Another classifier has been implemented using pack-propagation NN technique. It has been concluded that early prediction in serious and chronicle diseases is better than detection of the survival of such disease without complete knowledge of stage and situation of infirmity of patients. More over it improve future decisions for early cure of patient.

References

[1] https://www.kidney.org.

[2] S. Vijayarani and S. Dhayanand, Data Mining Classification Algorithms for Kidney Disease Prediction, in *International Journal on Cybernetics & Informatics (IJCI)* Vol. 4, No. 4, August 2015.

[3] S. Parul, and S. Poonam, Comparative Study in Chronic Kidney Disease Prediction using KNN and SVM, *in International Journal of Engineering Research & Technology* Volume. 4 - Issue. 12, December - 2015.

[4] A. Abhishek, G. Sundar M. Thakur, and D. Gupta, Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis, *in International Journal of Computer Science and Information Technologiey,* Vol. 3 (3) , pp 3900-3904, 2012.

[5] L.Frimat, D. Pau, G. Raymond, and G Shoukroun, Data Mining Based On Real World Data In Chronic Kidney Disease Patients Not On Dialysis: The Key Role Of Early Hemoglobin Levels Control, *in Journal of International* Pharmacoeconomics and outcomes Research Volume 18, Issue 7, Page A508, Nov 2015.

[6] 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, Issue 1, January 2016.

[7] A. Vigil, et al, Predictors of a Rapid Decline of Renal Function in Patients with Chronic Kidney Disease Referred to a Nephrology Outpatient Clinic: A Longitudinal Study, *in Journal of Advances in Nephrology*, Volume 2015 (2015), Article ID 657624, 8 pages, Nov, 2015.

[8] P. Swath, and T. Panduranga, Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms, *International Journal of Engineering Research & Technology (IJERT),* Vol. 4 Issue 07, July 2014.

[9] http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

[10] https://www.kidney.org/atoz/content/gfr.

[11] K Tapas, et al., An efficient k-means clustering algorithm: Analysis and implementation, *Pattern Analysis and Machine Intelligence, IEEE Transactions,* PP 881- 892, July 2015.

[12] P. Korpipää, et al., Managing context information in mobile devices, *IEEE Pervasive Computing, Mobile and Ubiquitous Systems*, vol. 2, no. 3, pp. 42-51, July-September 2003.

[13] K. Kumar and A. Abhishek, "Artificial Neural Networks for Diagnosis of Kidney Stones Disease*", International Journal Information Technology and Computer Science,* Vol. 7, No. 3, pp. 20-25, 2012.