Data Mining for Chronic Kidney Disease Prediction

Conference Paper · March 2019

CITATIONS
5

READS
1,931

3 authors, including:

Faisal Aqlan
Penn State Behrend
113 PUBLICATIONS **509** CITATIONS
SEE PROFILE

Abdulrahman Shamsan
Binghamton University
9 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Proceedings of the 2017 Industrial and Systems Engineering Conference
K. Coperich, E. Cudney, H. Nembhard, eds.

Data Mining for Chronic Kidney Disease Prediction

Faisal Aqlan and Ryan Markle
Industrial Engineering Department
Penn State Behrend
Erie, PA 16563

Abdulrahman Shamsan
Department of Systems Science and Industrial Engineering
State University of New York at Binghamton
Binghamton, NY 13902

Abstract

Chronic Kidney Disease (CKD) is one of the most widespread illnesses in the United States. Recent statistics show that twenty-six million adults in the United States have CKD and million others are at increased risk. Clinical diagnosis of CKD is based on blood and urine tests as well as removing a sample of kidney tissue for testing. Early diagnosis and detection of kidney disease is important to help stop the progression to kidney failure. Data mining and analytics techniques can be used for predicting CKD by utilizing historical patient's data and diagnosis records. In this research, predictive analytics techniques such as Decision Trees, Logistic Regression, Naive Bayes, and Artificial Neural Networks are used for predicting CKD. Preprocessing of the data is performed to impute any missing data and identify the variables that should be considered in the prediction models. The different predictive analytics models are assessed and compared based on accuracy of prediction. The study provides a decision support tool that can help in the diagnosis of CKD.

Keywords: Chronic kidney disease, diagnostics, data analysis, data mining, predictive analytics

1. Introduction

Chronic Kidney Disease (CKD) is the gradual loss of kidney function over time. CKD, also called chronic kidney failure or chronic or renal disease, can be caused by several factors including high blood pressure, diabetes, and other disorders. According to the National Kidney Foundation (www.kidney.org), there are twenty-six million adults in the United States who have CKD and million others are at increased risk. Kidneys filter the excess fluids and wastes from the blood and as CKD progresses, these wastes and fluids can build in the body and can cause heart and blood vessel disease. Patients who have CKD suffer from symptoms such as lack of energy, fatigue, drowsiness, pain, and pruritus [1]. The factors that increase the risk of Kidney disease include: Diabetes, Hypertension, Smoking, Obesity, Heart Disease, Family History of Kidney Disease, Alcohol Intake, Drug Abuse/Drug Overdose, Age, Race/Ethnicity, and Male Sex [2]. CKD has five different stages of development. Each stage increases in severity as one progresses from Stage 1 to Stage 5. In Stage 1, a person can develop below normal kidney functions and even experience a slight loss in kidney function. During Stage 2, a person can experience slight to moderate loss in kidney function. Stage 3 further intensifies, with a person experiencing moderate to severe loss in kidney function. In Stage 4, a person will experience a severe loss in kidney function. In Stage 5, a person will experience complete kidney failure.

According to [3], there are no predictive instruments that are commonly accepted for CKD. This lack of a common predictive instrument is becoming a bigger issue as the amount of CKD patients continues to grow. It is also stated that the health burden due to CKD is likely to continue to rise with the aging population and world-wide increase in Type 2 Diabetes [4]. CKD is a disease that affects everyone differently and is progressive in some, but not all patients [5]. Even though different approaches for preventing, reducing, halting, and reversing CKD have been described in the medical writings, all related factors have not been identified comprehensively [6]. Data mining techniques are used to investigate renal disease and to analyze the differences among various administrative areas. Data mining methods used in the literature include Adaptive Neuro-Fuzzy Inference, Support Vector Machines,

Artificial Neural Networks, etc. In this study, we will develop predictive analytics models to study and analyze chronic kidney disease. The data set we use has some missing data and we will adopt methods to impute the missing data. According to [7], Imputation is the process used to determine and assign replacement values for missing data items. Imputation is a great data analytics technique to use when a complete data set is needed. There are several different types of data imputation such as Mean Imputation, Multiple Linear Regression, and Hot Deck Imputation. Several data mining techniques will be used to predict CKD based on the input variables. We use IBM SPSS Modeler for implementing this study.

The rest of this paper is organized as follows: Section 2 reviews the literature related to data mining for kidney disease. In Section 3, we present the proposed data mining framework. In Section 4, implementation of the proposed framework is discussed. In Section 4, we also discuss the imputation of missing data and the predictive analytics models. Finally, Section 5 presents the conclusions and future work.

2. Related Literature

Data mining and analytics are widely used to extract useful information from raw data. Today, data mining has become an important field in healthcare that is used to detect unknown information in healthcare datasets and utilize analytics to predict diseases. For chronic kidney disease (CKD), authors use predictive analytics models to predict the disease based on its causes and develop models for its progression. A survey of the prediction models for CKD was presented in [8]. The authors reviewed thirteen studies that describe 23 predictive models for kidney disease. In [6], the authors discussed a text mining approach for CKD. The proposed approach extracts information from the previously published literature. The use of artificial neural network models to predict kidney stones was presented in [9]. Three different neural network algorithms were used and compared based on their accuracy, time to develop the model, and size of training datasets. The use of data mining techniques for predicting kidney dialysis survivability was discussed in [10]. Three predictive analytics models were used and compared. In [11], data mining classification techniques were used to predict kidney disease. Two classification methods were used, Naïve Bayes (NB) and Support Vector Machines (SVM). The authors found that the performance of SVM is better that NB. There are five stages of CKD and the disease can progress from one stage to another in an irreversible process. Early diagnosis of the disease is very important. To help diagnosing the disease, some authors developed analytical models to predict the different stages of CKD. A metabolomics based predictive analytics model was developed to identify the stages of CKD [12]. A new index was developed to predict the CKD stages with accuracy of 81.3%. In [8], the authors developed a model for predicting the progression of CKD to kidney failure. Regression and discrimination methods were used. A case study for predicting CKD in a local hospital in England was discussed in [4]. The study considered two main outcomes: moderate to severe CKD and end stage failure of the kidney. The predictive models provide a basis for potential identification of high risk patients. In [1], authors studied the relationship between the clusters of symptoms and the patient's quality of life focusing on CKD, stages 2 to 4. In [13], a prediction model using artificial neural networks ensemble was developed to identify the end stage of kidney disease.
In this study, we will develop and compare different predictive analytics models for CKD. Since the collected data includes missing values, an approach for dealing with missing data is proposed first. Then, feature selection is used to identify the significant features that will be included in the predictive models. Six main prediction models are considered and the performance of the models is also compared.

3. Data Mining Framework

The proposed data mining framework for CKD analytics is shown in Figure 1. The raw data is first analyzed and coded and then preprocessed to identify outliers and missing data. To deal with the missing data, we used the following criteria: first, if the variable has greater than 15% missing data we removed it. Second, if the variable has 15% or less missing data we keep it and use missing data imputation methods to impute the missing data. Five missing data imputation techniques are considered: fixed using mean, fixed using mid-range, random uniform, random normal, and Classification and Regression Trees (C&RT) algorithm. The imputation methods are evaluated based on the variability in the imputed data compared to the original data. Feature election is then used to identify the features to be used in the prediction models. The prediction models used are: Neural Networks, Logistic Regression, Bayes Net, Random Trees, Chi-square Automatic Interaction Detector (CHAID), and Support Vector Machines. The models were then evaluated based on accuracy, sensitivity, and specificity. The data for this study was obtained from the University of California – Irvine repository.

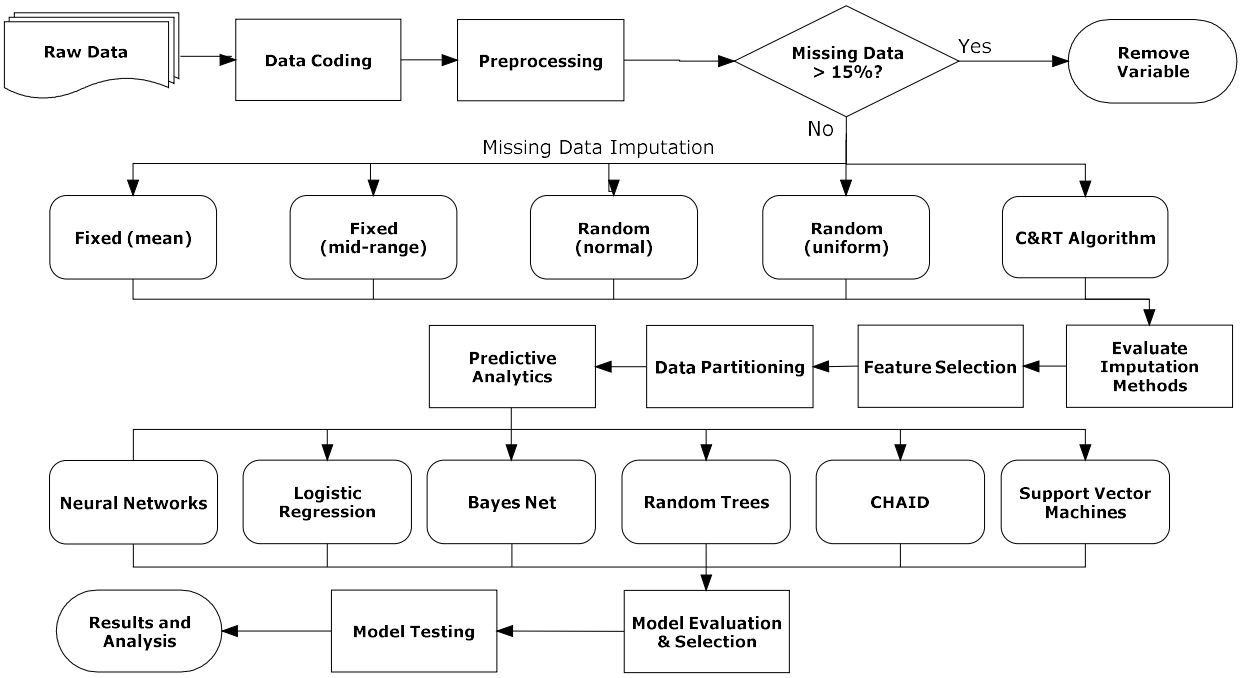


Figure 1: Propose framework for CKD analysis and prediction

4. Framework Implementation

The data set for this research was obtained from UC Irvine data repository. The data set includes 400 data points and 25 attributes. There are 24 input variables and one output variable. Among the input variables, there are 11 continuous or numerical and the rest are nominal. The output variables are a class which has two categories: ckd and notckd. All the input variables have missing data. A summary of the data set is included in Table 1 below.

Table 1: Summary of the CKD data

Attribute	Type	Description, Units, and Values	Min	Max	Mean	Std. Dev	Unique	Valid
Age	Continuous	Age in years	2	90	51.483	17.17	--	391
Blood Pressure	Continuous	Blood pressure in mm/Hg	50	180	76.469	13.684	--	388
Specific Gravity	Nominal	Specific gravity - (1.005,1.010,1.015,1.020,1.025)	--	--	--	--	6	347
Albumin	Nominal	Albumin - (0,1,2,3,4,5)	--	--	--	--	7	348
Sugar	Nominal	Sugar - (0,1,2,3,4,5)	--	--	--	--	7	351
Red Blood Cell	Nominal	Red Blood Cells - (normal, abnormal)	--	--	--	--	3	248
Pus Cell	Nominal	Pus Cells - (normal, abnormal)	--	--	--	--	3	335
Pus Cell Clumps	Nominal	Pus Cell Clumps - (present, notpresent)	--	--	--	--	3	387
Bacteria	Nominal	Bacteria - (present, notpresent)	--	--	--	--	3	396
Blood Glucose Random	Continuous	Blood Glucose Random in mgs/dl	22	490	148.037	79.282	--	346
Blood Urea	Continuous	Blood Urea in mgs/dl	1.5	391	57.426	50.503	--	381
Serum Creatinine	Continuous	Serum Creatinine in mgs/dl	0.4	76	3.072	5.741	--	383
Sodium	Continuous	Sodium in mEq/L	4.5	163	137.529	10.409	--	313
Potassium	Continuous	Potassium in mEq/L	2.5	47	4.627	3.194	--	312
Hemoglobin	Continuous	Hemoglobin in gms	3.1	17.8	12.526	2.913	--	341
Packed Cell Volume	Continuous	Packed Cell Volume	9	54	38.884	8.99	--	329
White Blood Cell Count	Continuous	White Blood Cell Count in cells/cumm	2200	26400	8406.122	2944.474	--	294
Red Blood Cell Count	Continuous	Red Blood Cell Count (millions/cmm)	2.1	8	4.707	1.025	--	269
Hypertension	Nominal	Hypertension - (yes, no)	--	--	--	--	3	398
Diabetes Mellitus	Nominal	Diabetes Mellitus - (yes, no)	--	--	--	--	3	398
Coronary Artery Disease	Nominal	Coronary Artery Disease - (yes, no)	--	--	--	--	3	398
Appetite	Nominal	Appetite - (good, poor)	--	--	--	--	3	399
Pedal Edema	Nominal	Pedal Edema - (yes, no)	--	--	--	--	3	399
Anemia	Nominal	Anemia - (yes,no)	--	--	--	--	3	399
Class	Flag	Class - (ckd, notckd)	--	--	--	--	2	400

4.1 Missing Data Imputation

Variables that have more than 15% missing data were removed. Analysis of the missing data n the remaining variables is shown in Figure 2. Five methods were used to impute missing data: fixed (using the mean), fixed (using the range), random (using normal distribution), random (using uniform distribution), and C&RT algorithm. Two examples for the missing data imputation are shown in Tables 2 and 3. For discrete attributes, only one random method is used which is based on empirical distributions for actual data. Moreover, mean is used with continuous variables whereas mode is used with discrete variables. C&RT algorithm was selected to impute the missing data because it produces the less amount of variability in the imputed data when compared to the original data.

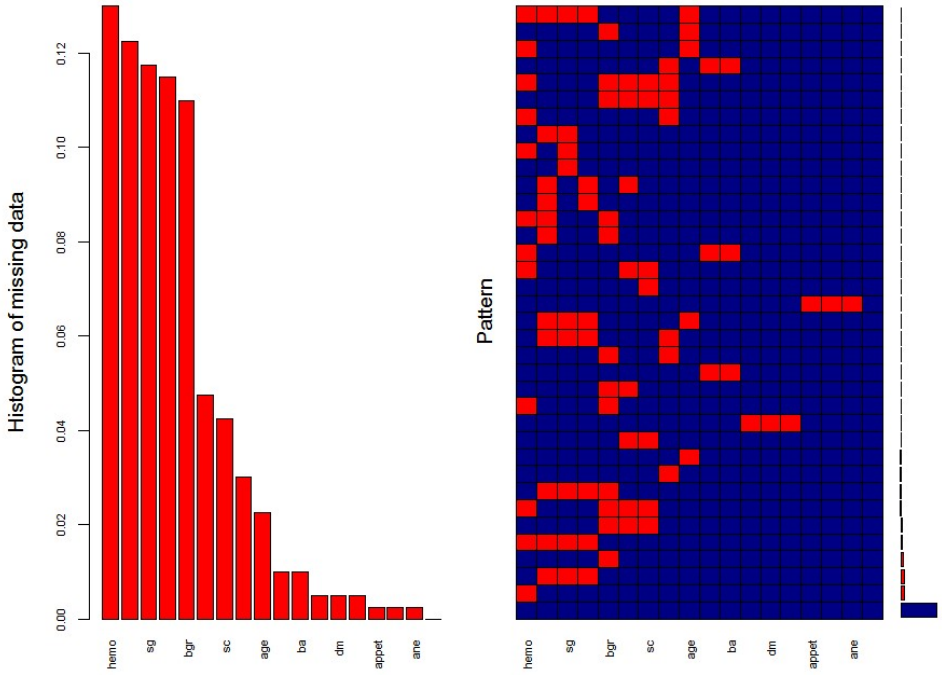
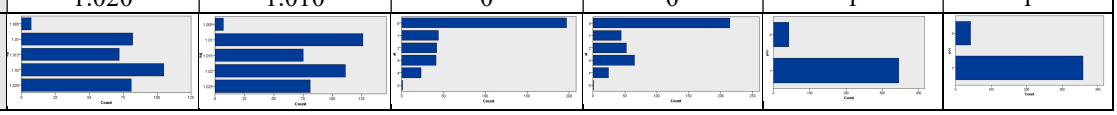


Figure 2: Analysis of missing data

Table 2: Missing data imputation results for continuous variables using Random Uniform method

Attribute → / Statistic ↓	Age		Blood Glucose Random		Hemoglobin	
	Original data	Imputed data	Original data	Imputed data	Original data	Imputed data
Count	391	400	346	400	341	400
Mean	51.483	51.495	148.043	161.427	12.576	12.233
Minimum	2	2	22	22	3.100	3.100
Maximum	90	90	490	490	17.80	12.800
Std. deviation	17.170	17.239	79.780	95.734	2.882	3.286
Median	55	55	121	124	12.7	12.500
Mode	60	60	99	99	15.000	15.000

Table 3: Missing data imputation results for discrete variables using CR&T algorithm

Attribute → / Statistic ↓	Specific Gravity		Albumin		Pus Cell Clumps	
	Original data	Imputed data	Original data	Imputed data	Original data	Imputed data
Count	347	400	348	400	387	400
Minimum	1.005	1.005	0	0	0	0
Maximum	1.025	1.025	5	5	1	1
Unique	5	5	6	6	2	2
Mode	1.020	1.010	0	0	1	1
Graph						

4.2 Predictive Analytics Models for CKD

In order to predict CKD, six different analytics methods were used: Neural Networks (NN), Logistic Regression (LR), Bayes Net (BN), Random Trees (RT), Discriminant Analysis (DA), and Support Vector Machines (SVM). Three performance metrics are used to evaluate the analytics models: accuracy, sensitivity, and specificity. Definitions of the three metrics along with their descriptions are shown in Table 3. The confusion matrix is shown below. Positive classification is when the person has CKD and negative classification of when the person does not have CKD.

Table 4: Performance metrics for evaluating the analytics models

Metric	Description	Equation
Accuracy	Measures the ability of the model to correctly predict the class label of new or unseen data.	$\frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	Measures the proportion of positives (or Yes's) that are correctly identified as such.	$\frac{TP}{TP + FN}$
Specificity	Measures the proportion of negatives (or No's) that are correctly identified as such.	$\frac{TN}{TN + FP}$

Abbreviation	Name	Description
TP	True Positives	Number of correct classifications predicted as positive (or Yes)
TN	True Negatives	Number of correct classifications predicted as negative (or No)
FP	False Positive	Number of examples that are incorrectly predicted as positive when it is actually negative
FN	False Negative	Number of examples that are incorrectly predicted as negative when it is actually positive

Table 5: Confusion matrix

Outcome of the Diagnostic Test		Predicted	
		Positive (1)	Negative (0)
Observed	Positive (1)	TP	FP
	Negative (0)	FN	TN

The data was partitioned into 80% for training and 20% for testing. Training data set includes 320 data points (201 with CKD and 119 without CKD). Testing data set includes 80 data points (49 with CKD and 31 without CKD). The results are shown in Table 5. For both training and testing data, the accuracy of the prediction models is very high and the model that has the highest accuracy for both data sets is Random Trees. This is also true for the Sensitivity and Specificity measures. Illustrations of Bayesian Network and K-means clustering are shown in Figure 3.

Table 6: Performance measures for the training data

Method	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Neural Network	200	118	1	1	99.38	99.50	99.16
Logistic Regression	201	119	0	0	100.00	100.00	100.00
Bayes Net	199	118	3	1	98.75	99.50	97.52
Random Trees	**201**	**119**	**0**	**0**	**100.00**	**100.00**	**100.00**
CHAID	195	117	6	2	97.50	98.99	95.12
Support Vector Machines	194	119	7	0	97.80	100.00	94.44

Table 7: Performance measures for the test data

Method	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Neural Network	48	31	1	0	98.75	100.00	96.88
Logistic Regression	49	28	0	3	96.25	94.23	100.00
Bayes Net	47	31	2	0	97.50	100.00	93.94
Random Trees	**49**	**31**	**0**	**0**	**100.00**	**100.00**	**100.00**
CHAID	46	31	3	0	96.25	100.00	91.18
Support Vector Machines	47	31	2	0	97.50	100.00	93.94

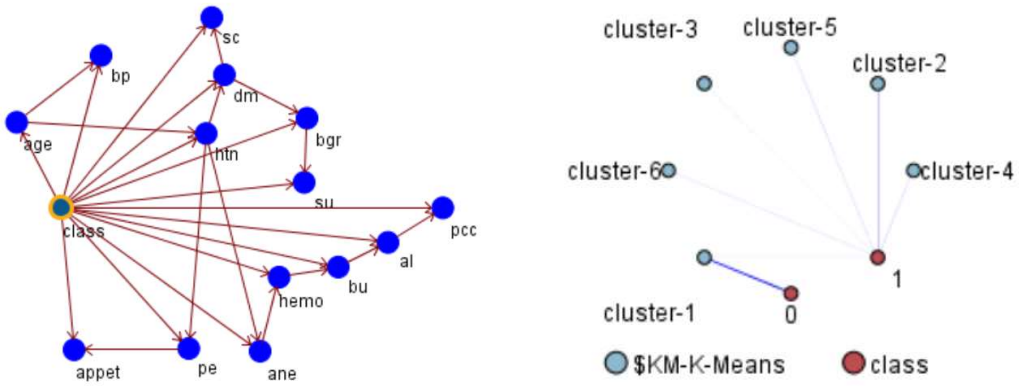


Figure 3: Illustration of the Bayesian Network (left) and K-means clustering (right)

5. Conclusions

In this paper, we discussed the application of data mining and analytics techniques to predict CKD. Since the data includes missing values, several missing data imputation methods were utilized. The method, i.e. C&RT algorithm, that minimizes the variability in the imputed data was selected. Six predictive analytics methods were utilized to predict CKD and it was found that Random Trees results in the best method. The data used in this study includes only two classes for the output variable, ckd and notckd. However, the ckd class can be further classified into five classes as indicated by National Kidney Foundation. Clustering of the data using K-means (see Figure 3) into six clusters shows that all Class 0 (notckd) cases were cluster into one cluster (cluster-1). Class 1 (ckd) cases were distributed into 5 clusters with different percentages. Future work will focus on using regression and clustering methods to study and predict the five stages of CKD.

References

1. Lee, S.J., and Jeon, J.H., 2015, "Relationship between Symptom Clusters and Quality of Life in Patients at Stages 2 to 4 Chronic Kidney Disease in Korea," Applied Nursing Research, 28(4), 13-19.
2. Bala, S., and Kumar, K., 2014, "A Literature Review on Kidney Disease Prediction Using Data Mining Classification Technique," International Journal of Computer Science & Mobile Computing, 3(7), 960-967.
3. Tangari, N., Kitsios, G.D., et al., 2013, "Risk Prediction Models for Patients with Chronic Kidney Disease" Annals of Internal Medicine, 158(8), 596-603.
4. Hippisley-Cox, J., and Coupland, C., 2010, "Predicting the Risk of Chronic Kidney Disease in Men and Women in England and Wales: Prospective Derivation and External Validation of the QKidney® Scores," Hippisley-Cox and Coupland BMC Family Practice, 11-49.
5. Ziyad, A., 2013, "Prediction of Renal End Points in Chronic Kidney Disease," Kidney International, 83(2), 189-191.
6. Kostoff, R.N., and Patel, U., 2015, "Literature-Related Discovery and Innovation: Chronic Kidney Disease," Technological Forecasting and Social Change, 91, 341-351.
7. Hernandez-Pereira, E., Alvarez-Estevez, D., and Moret-Bonillo, V., 2015, "Automatic Classification of Respiratory Patterns Involving Missing Data Imputation Techniques," Biosystems Engineering, 138, 65-76.
8. Tangri, N., Stevens, L., et al., 2011, "A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure," Journal of American Medical Association,305 (15), 1553-1559.
9. Kumar, K., and Abhishek, 2012, "Artificial Neural Networks for Diagnosis of Kidney Stones Disease", International Journal of Information Technology and Computer Science, 7, 20-25.
10. Lakshmi, K.R., Nagesh, Y., and VeeraKrishna, M., 2014, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability," International Journal of Advances in Engineering and Technology, 7(1), 242-254.
11. Vijayarani, S., and Dhayanand, S., 2015, "Data Mining Classification Algorithms for Kidney Disease Prediction," International Journal on Cybernetics and Information, 4(4), 13-25.
12. Kobayashi, T., Yoshida, T., et al., 2014, "A Metabolomics-Based Approach for Predicting Stages of Chronic Kidney Disease," Biochemical and Biophysical Research Communications, 445, 412–416.
13. Noia, T.D., Ostuni, V.C., et al., 2013, "An End Stage Kidney Disease Predictor Based on Artificial Neural Networks Ensemble," Expert Systems with Applications, 40, 4438-4445.