Review Article

Using Data Mining Techniques to Predict Chronic Kidney Disease: A Review Study

Abstract

One of the growing global health problems is chronic kidney disease (CKD). Early diagnosis, control, and management of chronic kidney disease are very important. This study considers articles published in English between 2016 and 2021 that use classification methods to predict kidney disease. Data mining models play a vital role in predicting disease. Through our study, data mining techniques of support vector machine, Naive Bayes, and k-nearest neighbor had the highest frequency. After that, random forest, neural network, and decision tree were the most common data mining techniques. Among the risk factors associated with chronic kidney disease, respectively, risk factors of albumin, age, red blood cells, pus cells, and serum creatinine had the highest frequency in these studies. The highest number of best yields was allocated to random forest technique. Reviewing larger databases in the field of kidney disease can help to better analyze the disease and ensure the risk factors extracted.

Keywords: *Classification, data mining, diagnosis, kidney diseases, machine learning*

Mohammad Sattari, Maryam Mohammadi[1]

Health Information Technology Research Center, Isfahan University of Medical Sciences, [1]Department of Management and Health Information Technology, School of Management and Medical Information Sciences, Isfahan University of Medical Sciences, Isfahan, Iran

Introduction

One of the growing global health problems is chronic kidney disease. Chronic kidney disease is associated with an increased mortality and risk of many diseases. More than two million people worldwide undergo dialysis or kidney transplants to survive, but it may represent only 10% of those who need treatment to survive. Every year, more than one million people in 112 low-income countries die from untreated kidney failure. That is why the cost of dialysis or kidney transplant treatment is high.[1,2] In 2017, the number of people with chronic kidney disease worldwide was 69.75 million, causing 1.2 million deaths.[3]

So, early diagnosis, control, and management of chronic kidney disease are very important.[2] One of the best ways to reduce this mortality is early treatment.[4] But in developing countries, patients are treated in critical conditions. It is possible to build an automated system to identify patients with chronic kidney disease before reaching the final stage. To achieve this goal, patients' clinical data such as age, sex, and blood pressure can be used.[5] Much research has been done on the development of artificial intelligence systems that have solutions to the disease. Data mining or machine learning models play a vital role in predicting disease. Data mining models extract patterns by creating some mathematical approaches. Then these patterns are used to diagnose disease.[6-8] Hippisley-Cox J, Coupland presented a case study to predict chronic kidney disease at a local hospital in the UK. The study checks two new data mining algorithms that provide a basis for identifying high-risk patients. It assesses more accurate track closer for reducing the risk of disease.[9] Tangri *et al*. developed prediction models using demographic, clinical, and laboratory data. The model consists of age, sex, glomerular filtration rate, albuminuria, serum calcium, serum phosphate, serum bicarbonate, and serum albumin.[10]

Many studies have applied different data mining techniques to predict chronic kidney disease. Here are some commonly used techniques. Artificial neural network (ANN) is a mathematical model that works like human neurons. Neural networks are non-linear statistical data modeling tools.[11] The perceptron neural network is the simplest neural network and exists as a single-layer and multilayer perceptron (MLP). The single-layer

How to cite this article: Sattari M, Mohammadi M. Using data mining techniques to predict chronic kidney disease: A review study. Int J Prev Med 2023;14:110.

Address for correspondence:
Ms. Maryam Mohammadi, Department of Management and Health Information Technology, School of Management and Medical Information Sciences, Isfahan University of Medical Sciences, Isfahan, Iran.
E-mail: mitmohammadi.89@gmail.com

Access this article online

Website: www.ijpvmjournal.net/www.ijpm.ir

DOI: 10.4103/ijpvm.ijpvm_482_21

Quick Response Code:

perceptron can only classify discrete linear problems, and for more complex problems, it is necessary to use more layers. A multilayer perceptron consists of components of layers and weights. In general, in multilayer perceptron, there are three types of neural layers which are input layer, hidden layer(s), and output layer.[12] The neurons of the input layer receive the data and transmit it to the neurons of the first hidden layer through weight links. Here, the data are processed and the result is transmitted to the neurons in the next layer. Finally, the neurons of the last layer provide the output of the network.[13] Convolutional neural network (CNN) is a mathematical structure that usually consists of three types of layers (or building blocks): convolution, pooling, and fully connected layers. The first two layers, the convolution and pooling layers, perform feature extraction, while the third layer, a fully connected layer, maps the extracted features to the final output, such as classification.[14] Support vector machine (SVM) technique is an efficient method for classification of linear and non-linear data. This method first uses a non-linear mapping to convert the initial data to high dimensions and then searches for the best separating hyperplane in the new dimension.[15] In fact, the purpose of the support vector machine is to create a decision boundary between two classes that allows the prediction of labels from one or more feature vectors and creates the maximum possible distance for the two classes of support vectors.[16] Random forest (RF) is an ensemble learning method for classification and regression problems that builds several decision trees during training. The random forest algorithm injects a subset of randomly selected data into each of the decision trees. Each of the algorithms performs the learning operation. When a new dataset is given to the algorithm for prediction, each of these trained algorithms predicts a result. Finally, the random forest algorithm can select the category with the most votes by voting and place it as the final category for the classification operation.[17] The decision tree algorithm can predict qualitative variables in addition to quantitative variables. The result of implementing the decision tree algorithm is a set of logical conditions with a tree structure that is used to predict a feature. So that the data placed in the last leaves of this tree are labeled by one of the values of the target attribute. The decision tree algorithm works in such a way that it tries to minimize diversity (in terms of target features) in the nodes. This non-uniformity in the nodes can be measured using impurity measures, the most important and widely used of which is the Gini index. Often, the difference between the types of decision trees is the measurement of impurity, splitting and pruning of tree nodes. Two examples of decision tree algorithms include classification and regression tree (CART) and C4.5.[18,19]

This research reviewed and analyzed studies that have applied data mining techniques for predicting the disease. Finally, it provided a scientific framework for future research in this field.

Table 1: Number of records in each database

Database	The number of records
PubMed	57
Web of Science	137
Science direct	234
Scopus	153
Total	581

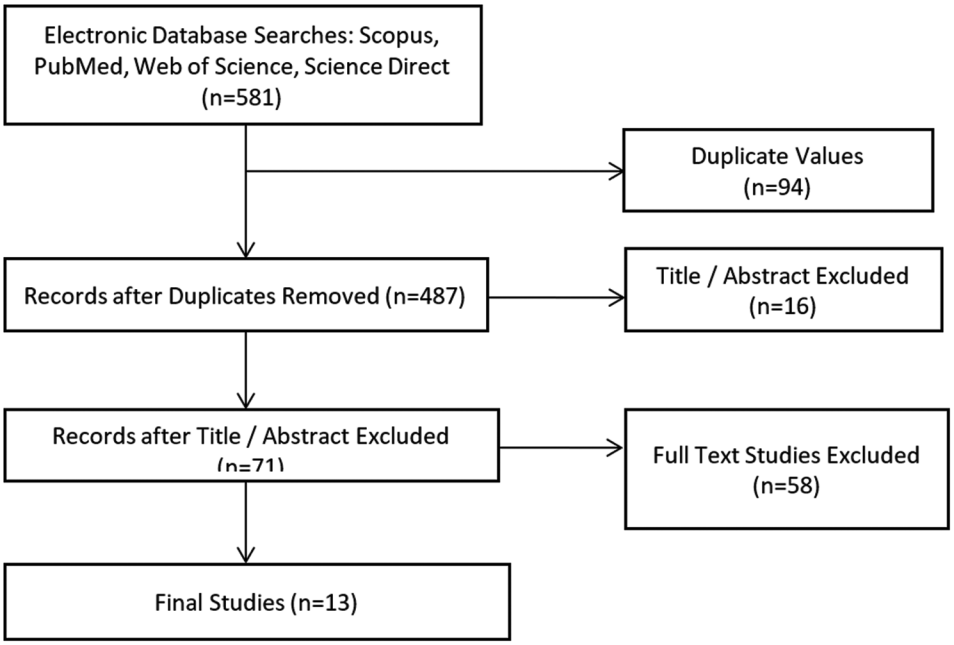

Figure 1: Extraction strategy of studies

Table 2: Characteristics of the reviewed studies including names, techniques, datasets, and criteria

Title	Authors	Year	Dataset	Dataset size	Techniques	Criteria
A Dynamic Pooling-based Convolutional Neural Network Approach to detect Chronic Kidney Disease	Navaneeth, *et al*.[20]	2020	School of Electronics Engineering, VIT University, India	172 instances of CKD patients	SVM - CNN	Accuracy Sensitivity Specificity
Risk Prediction for Early Chronic Kidney Disease: Results from an Adult Health Examination Program of 19,270 Individuals	Shih, *et al*.[3]	2020	Adult Health Examination Dataset	19,270 instances of CKD patients	CART, C4.5, LDA[1], ELM[2],	Accuracy Sensitivity Specificity AUC
Chronic Kidney Disease Prediction Using Machine Learning Methods	Ekanayake, *et al*.[21]	2020	UCI CKD Dataset	400 instances of CKD patients	Decision tree, random forest, XGB[3] extra tree, AdaBoost[4], KNN[5], NN[6], linear SVC[7], LR[8], RBF[9], Gaussian NB[10]	Accuracy Sensitivity
Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease	Alaiad, *et al*.[22]	2020	UCI CKD Dataset	400 instances of CKD patients	NB, SVM, decision tree KNN JRip	Accuracy Sensitivity Specificity
Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms	Aldhyani, *et al*.[23]	2020	Diabetic Disease Dataset Breast Cancer Disease Dataset Kidney Disease Dataset	400 instances of CKD patients	Rough K-means clustering SVM, NB, KNN random forest	Accuracy Sensitivity Specificity
Prediction of Kidney Disease Stages using Data Mining Algorithms	El-Houssainy, *et al*.[2]	2020	UCI CKD Dataset	361 instances of CKD patients	PNN[11] MLP SVM RBF	Accuracy Sensitivity Specificity
Development of an Ensemble Approach to Chronic Kidney Disease Diagnosis	Ayodele, *et al*.[24]	2020	UCI CKD Dataset	400 instances of CKD patients	NB KNN decision tree	Accuracy Sensitivity Specificity
Data Mining to Predict Early-Stage Chronic Kidney Disease	Pintoa, *et al*.[25]	2020	UCI CKD Dataset	400 instances of CKD patients	SVM, KNN, J48	Accuracy Sensitivity Specificity
Optimization of C4.5 Algorithm using Meta Learning in diagnosing of Chronic Kidney Diseases	Nurzahputra, *et al*.[26]	2019	UCI CKD Dataset	400 instances of CKD patients	C4.5, C4.5 and multiboot, C4.5 and Bagging	Accuracy Sensitivity AUC
Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant Miner and J48 Decision Tree	Islam, *et al*.[5]	2019	East West University (Bangladesh)	2800 instances of CKD patients	J48, ant miner, AdaBoost logit boost	Accuracy
Comparative Study of Classifier for Chronic Kidney Disease Prediction using Naive Bayes, KNN, and Random Forest	Devika, *et al*.[27]	2019	UCI CKD Dataset	400 instances of CKD patients	NB, KNN random forest	Accuracy Sensitivity
Early-Stage Chronic Kidney Disease Diagnosis by Applying Data Mining Methods to Urinalysis, Blood Analysis, and Disease History	Akben.[28]	2018	UCI CKD Dataset	400 instances of CKD patients	K-means, KNN, NB, SVM	Accuracy
Diagnosis of Chronic Kidney Disease Using Random Forest Classification	Balakrishna, *et al*.[29]	2017	UCI CKD Dataset	400 instances of CKD patients	Random forest J48 NB tree REP tree	Accuracy

[1]Latent Dirichlet allocation. [2]Extreme learning machine. [3]Extreme gradient boosting. [4]Adaptive boosting. [5]K-nearest neighbors. [6]Neural network. [7]Support vector classifier. [8]Logistic regression. [9]Radial basis functions. [10]Naive Bayes. [11]Probabilistic neural network

Methods

This study considers articles published in English between 2016 and 2021 that use classification methods to predict kidney disease. PubMed, Science direct, Web of Science, and Scopus databases are considered as searchable databases. [Table 1]. Table 1 shows the number of published studies between 2016 and 2021 in the mentioned dataset.

Search strategy

A combination of keywords and synonyms is searched based on Boolean logic (OR). Search results are combined and searched using Boolean logic (AND).

Inclusion criteria

Studies in English between 2016 and 2021 using classification techniques for predicting kidney disease are included. Keywords searched:

(Kidney disease) AND (Disease OR Diagnosis OR Prediction) AND (Data mining).

Exclusion criteria

Studies which are not in English language are not considered. Also, studies that have used transcript techniques to predict kidney disease are excluded.

Selection of studies

Duplicate records are removed, and then the title and abstract of the remaining studies are considered based on input and output criteria. Unrelated studies were excluded in terms of title and abstract. Then the full text of the articles was considered, and then among the remaining articles, the articles whose full text was not related were deleted. Finally, the remaining articles were considered.

Data extraction and classification

Information on authors' names, year of publication, dataset, risk factors, and techniques used, and evaluation criteria for each technique were extracted from the studies. The factors extracted by the researchers were then analyzed. Five hundred and eighty-one studies were retrieved after the initial search. Finally, 20 cases had the necessary criteria to enter this study. Table 1 shows the number of initial studies retrieved from each database. Ninety-four studies that were duplicates were excluded. The remaining studies (487) were reviewed and evaluated based on the title and abstract, and 416 studies were excluded. After studying the full text of the remaining studies (71 cases), 58 studies were excluded, and finally, 13 studies were selected [Figure 1].

Results

After searching and evaluation, the final analysis was performed on 13 articles. Findings were presented in five sections: risk factors, datasets, data mining technique, the best performance of the techniques in terms of accuracy, sensitivity, specificity, area under the curve, and the criteria used.

In total, after reviewing the articles, 13 articles in the field of data mining on chronic kidney disease were obtained, all of them were in English.

The results of the studies on data mining techniques used in these studies are also presented in Table 2. According to these studies, a total of 24 techniques were used in these studies. Support vector machine (seven studies), Naive Bayes (seven studies), and k-nearest neighbor (seven studies) had the highest frequency. Random forest was used in four studies, and data mining techniques of neural network and decision tree were applied in three studies.

Table 2 shows the criteria used to measure the performance of algorithms in studies. Accuracy had the highest use with 13 studies. Sensitivity with ten studies and specificity with seven studies are the second and third most used measure.

Table 3: Risk factors used in studies and their use

Risk factor	Frequency
Albumin (AL)	12
Age	11
Red blood cells (RBC)	11
Pus cell (PC)	11
Serum creatinine (SC)	11
Blood pressure (BP)	10
Specific gravity (SG)	10
Sugar (SU)	10
Pus cell clumps (PCC)	10
Bacteria (Ba)	10
Blood glucose random (BGR)	10
Blood urea (BU)	10
Hemoglobin (Hgb)	10
Diabetes mellitus (DM)	10
Sodium (Sod)	9
Potassium (Pot)	9
Packed cell volume (PCV)	9
Hypertension (Htn)	9
White blood cell count (WC)	8
Red blood cell count (RC)	8
Appetite (Appet)	7
Coronary artery disease (CAD)	6
Pedal edema (Pe)	6
Anemia	6
Sex/gender	2
High-density lipoprotein cholesterol (HDL-C)	1
(LDL-C) Low-density lipoprotein cholesterol	1
Proteinuria	1
Urine protein and creatinine ratio (UPCR)	1
Total cholesterol (T-CHO)	1
Glomerular filtration rate (GFR)	1
Arterial blood vessel sickness	1
Relative density	1

The area under the curve with two studies had the least application in the studies.

According to Table 2, the UCI dataset (ten studies) had the highest frequency among all datasets. The other datasets were used in only one study.

Also, according to Table 3, a total of 33 risk factors related to chronic kidney disease were considered in the studies. The most frequent measures are albumin (12 studies), age (11 studies), red blood cells (11 studies), pus cells (11 studies), and serum creatinine (11 studies).

The above factors can be classified into three categories. They include demographic characteristics, symptom, examination characteristics, and experimental characteristics.

The results obtained for the number of best performance of data mining techniques in terms of accuracy, sensitivity, specificity, and area under the curve are given in Table 4. Among the studies, the random forest technique (three cases) and the k-nearest neighbor technique (two cases) had the highest number of best performances for the accuracy. The random forest technique (two cases) had the highest number of best practices for sensitivity.

Discussion

Albumin factor has been used the most in data mining studies, and this factor has been used in 92% of studies. Albumin determines the amount of protein excreted in the urine of a person that if the excretion amount of this substance is not normal, it will increase the risk of kidney disease. Age along with red blood cells, pus cells, and serum creatinine was used in 84% of the studies. Studies show that people with kidney failure have shorter lifespan of red blood cells than others. In fact, it can be said that the study of these blood cells can be an effective factor in identifying this disease.[30] Studies have also shown that older people are more prone to kidney disease.[31] Other factors, such as salt intake, have been used in approximately 70% of studies. The kidney is the part of the body that suffers the most from salt intake. In fact, too much salt can cause the kidneys to excrete waste products and lead to kidney failure.

In this study, UCI dataset is the most common datasets and was used in more than 50% of the articles. The distinguishing feature of this collection from other collections is the easy access and freeness. That is why many researchers are willing to use this collection.

Table 4: Data mining techniques and number of best performances

Technique	The number of the best performance in terms of accuracy	The number of the best performance in terms of sensitivity	The number of the best performance in terms of specificity	The number of the best performance in terms of AUC
SVM with RKM	1	1	1	1
SVM-CNN	1	1	1	1
NB	-	-	-	-
KNN	2	1	1	-
Random subspace ensemble technique and KNN	1	1	1	-
Random forest	3	2	-	-
PNN	1	1	1	-
NN	-	-	-	-
Decision tree	-	-	-	-
C4.5	1	1	1	-
j48	1	1	1	-
RBF	-	-	-	-
K-Means	-	-	-	-
LR	-	-	-	-
CART	-	-	-	-
MLP	-	-	-	-
LDA	-	-	-	-
ELM	-	-	-	-
XGB extra trees	1	1	-	-
AdaBoost	-	-	-	-
JRip	-	-	-	-
C4.5 and multiboot	-	-	-	-
C4.5 and bagging	1	1	1	-
Ant miner	1	1	-	-
AdaBoost	-	-	-	-
Logit boost	1	1	-	-
REP tree	-	-	-	-

The point, of course, is that facilitating data access can be an important step in using data mining techniques. Twenty-seven data mining techniques have been applied in 15 studies. Of course, techniques such as support vector machine, Naive Bayes, and k-nearest neighbor have been used in more than 50% of studies. The results show that besides seeking to extract different patterns in renal patients, the studies also sought to improve the performance of data mining techniques in terms of accuracy. So, it can be said that the techniques finally had the best performance in one or two studies, and it is not possible to suggest a specific technique for all datasets.

Conclusions

Techniques such as random forest and k-nearest neighbor have the highest number of best performances than other techniques for predicting kidney disease, although they are not very far from other techniques. The use of these techniques along with techniques such as support vector machines that have performed well in various fields of medicine can provide a good basis for clinicians in the field of kidney disease to study and evaluate risk factors.

In fact, in the future it is possible to design decision support systems that take risk factors such as age, salt intake, albumin, red blood cells, pus cells, and serum creatinine from the input. The system then predicts based on these factors how much a person will be at risk for kidney disease in the future. Also, reviewing larger databases in the field of kidney disease can help to better analyze the disease and ensure the risk factors extracted.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

Received: 16 Nov 21 **Accepted:** 17 May 23
Published: 28 Aug 23

References

1. Couser WG, Remuzzi G, Mendis S, Tonelli M. The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. Kidney Int 2011;80:1258-70.
2. Rady E-HA, Anwar AS. Prediction of kidney disease stages using data mining algorithms. Informatics Med Unlocked 2019;15:100178.
3. Shih CC, Lu CJ, Chen GD, Chang CC. Risk prediction for early chronic kidney disease: Results from an adult health examination program of 19,270 individuals. Int J Environ Res Public Health 2020;17:1-11.
4. Ho CY, Pai TW, Peng YC, Lee CH, Chen YC, Chen YT, *et al*. Ultrasonography image analysis for detection and classification of chronic kidney disease. In2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems 2012 p. 624-9. IEEE.
5. Arif-Ul-Islam, Ripon SH. Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree. In: 2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019.; 2019.
6. Sisodia DS, Verma A. Prediction performance of individual and ensemble learners for chronic kidney disease. In2017 international conference on inventive computing and informatics (ICICI) 2017; p. 1027-31. IEEE.
7. Al-Hyari AY, Al-Taee AM, Al-Taee MA. Diagnosis and classification of chronic renal failure utilising intelligent data mining classifiers. Int J Inf Technol Web Eng 2014;9:1-12.
8. Sinha P, Sinha P. Comparative study of chronic kidney disease prediction using KNN and SVM. International Journal of Engineering Research and Technology 2015;4:608-12.
9. Hippisley-Cox J, Coupland C. Predicting the risk of chronic kidney disease in men and women in england and wales: Prospective derivation and external validation of the QKidney scores. BMC Fam Pract 2010;11:49.
10. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, *et al*. A predictive model for progression of chronic kidney disease to kidney failure. JAMA 2011;305:1553-9.
11. Embrechts MJ. Neural networks for data mining. In: Intelligent Engineering Systems Through Artificial Neural Networks. 7. ASME; 1997:741–6.
12. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. Data Min Concepts Tech 2012.
13. Amato F, López-Rodríguez A, Peña-Méndez E, Vaňhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. J Appl Biomed 2013;11:47-58.
14. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: An overview and application in radiology. Insights Imaging 2018;9:611-29.
15. Kolukisa B, Yavuz L, Soran A, Bakir-Gungor B, Tuncer D, Onen A, *et al*. Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm. Int J Biosci Biochem Bioinforma 2020;10:58-65.
16. Tseng C-J, Lu C-J, Chang C-C, Chen G-D. Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Comput Appl 2014;24:1311-6.
17. Qi Y. Random forest for bioinformatics. In: Ensemble machine learning. Springer US; 2012:307-23.
18. Song YY, Lu Y. Decision tree methods: Applications for classification and prediction. Shanghai Arch Psychiatry 2015;27:130-5.
19. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 1991;21:660-74.
20. Navaneeth B, Suchetha M. A dynamic pooling based convolutional neural network approach to detect chronic kidney disease. Biomed Signal Process Control 2020;62:102068.
21. Ekanayake IU, Herath D. Chronic kidney disease prediction using machine learning methods. In: MERCon 2020 - 6th International Multidisciplinary Moratuwa Engineering Research Conference, Proceedings. 2020:260–5.
22. Alaiad A, Najadat H, Mohsen B, Balhaf K. Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease. Journal of Information & Knowledge Management (JIKM), World Scientific Publishing Co. Pte. Ltd., 2020;19:1-17.
23. Aldhyani THH, Alshebami AS, Alzahrani MY. Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms. J Healthc Eng 2020;2020:4984967.
24. Jongbo OA, Adetunmbi AO, Ogunrinde RB, Badeji-Ajisafe B. Development of an ensemble approach to chronic kidney disease diagnosis. Sci African 2020;8:e00456.

25. Pinto A, Ferreira D, Neto C, Abelha A, Machado J. Data mining to predict early stage chronic kidney disease. In: Procedia Computer Science 177;2020:562-7.
26. Nurzahputra A, Muslim MA, Prasetiyo B. Optimization of C4.5 algorithm using meta learning in diagnosing of chronic kidney diseases. In: Journal of Physics: Conference Series. 1321; 2019.
27. Devika R, Avilala SV, Subramaniyaswamy V. Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest. In: Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019.; 2019:679-84.
28. Akben SB. Early stage chronic kidney disease diagnosis by applying data mining methods to urinalysis, blood analysis and disease history. IRBM 2018;39:353-8.
29. Balakrishna T, Narendra B, Reddy MH, Jayasri D. Diagnosis of chronic kidney disease using random forest classification technique. HELIX 2017;7:873-7.
30. Li JH, Luo JF, Jiang Y, Ma YJ, Ji YQ, Zhu GL, *et al*. Red blood cell lifespan shortening in patients with early-stage chronic kidney disease. Kidney Blood Press Res 2019;44:1158-65.
31. Raman M, Green D, Middleton RJ, Kalra PA. Comparing the impact of older age on outcome in chronic kidney disease of different etiologies: A prospective cohort study. J Nephrol 2018;31:931-9.