Efficient datamining model for prediction of chronic kidney disease using wrapper methods

Ramaswamyreddy A, Shiva Prasad S, K V Rangarao, A Saranya
Department of Computer Science and Engineering, VFSTR University, India

Article Info

Article history:

Received Dec 12, 2018
Revised Mar 20, 2019
Accepted Apr 12, 2019

Keywords:

Bagging
Boosting
Filter methods
Wrapper methods

ABSTRACT

In the present generation, majority of the people are highly affected by kidney diseases. Among them, chronic kidney is the most common life threatening disease which can be prevented by early detection.Histological grade in chronic kidney disease provides clinically important prognostic information. Therefore, machine learning techniques are applied on the information collected from previously diagnosed patients in order to discover the knowledge and patterns for making precise predictions.A large number of features exist in the raw data in which some may cause low information and error; hence feature selection techniques can be used to retrieve useful subset of features and to improve the computation performance. In this manuscript we use a set of Filter, Wrapper methods followed by Bagging and Boosting models with parameter tuning technique to classify chronic kidney disease.Capability of Bagging and Boosting classifiers are compared and the best ensemble classifier which attains high stability with better promising results is identified.

Corresponding Author:

Ramaswamyreddy A,
Department of Computer Science and Engineering,
VFSTR University, Vadlamudi, India.
Email: ramaswamyredyymail@gmail.com

1. INTRODUCTION

Kidneys are the important functional unit in the human body which is situated below the spinal cord of the body. Kidneys works like water filter how water filter purifies the water similarly it purifies the blood and removes the waste materials from the body, so kidneys can be called as Filtration unit. The other functionality of kidneys is Regulates the blood pressure and controls the sugar level inside the body. Failure of the kidney occurs when it cannot be able to purifies the blood and cannot be able to eradicate the waste contents form the body, then the body is fully filled up with high toxins which leads to the death of a person.

Kidney diseases can be categorized into two types (i) Acute Kidney Disease (ii) Chronic Kidney Disease. Acute kidney Disease is usually caused by an event that leads to kidney malfunction, such as dehydration, blood loss from major surgery or injury, or the use of medicines. Chronic kidney disease (CKD) is usually caused by a long-term disease, such as high blood pressure or diabetes that slowly damages the kidneys and reduces their function over time. Chronic kidney disease may not show any symptoms if little kidney function remains so it is difficult to predict whether a person is suffering from disease or not. Along with chronic kidney disease other problems may arises such as anemia, hypophosphatemia which shows high complications in kidney failure.

The objective of this paper is to predict the chronic kidney disease with the help of ensemble classifiers. This paper is organized into six sections. Section 2 shows the overview of Literature work, Feature selection techniques are explained in the Section 3, Section 4 illustrates the classification techniques, Proposed Methodology is presented in the Section 5 and finally Section 6 demonstrates the Result analysis and conclusion part.

2. LITERATURE WORK

Researchers have proposed various methods on predicting the chronic kidney disease problem and some of these works are specified here. Rubini et al [1] specified the classifiers like Multilayer perceptron (MLP),logistic regression, radial basis function network and they analyzed that Multilayer perceptron (MLP) has given the best accuracy. Baby et al [2] implemented the six classification algorithms in predicting the chronic kidney disease. The classifiers are kstar, J48, Naive Bayes, Random Forest, AD Trees and concluded that Naive Bayes is giving highest accuracy.

Dhayanand et al [3] used SVM and Naive Bayes classifiers in predicting the renal disease. They compared the two classifiers and identified that SVM is the best one in renal disease prediction. Jena et al [4] suggested the classification algorithms like Support Vector Machine, Naive Bayes, conjunctive rule, J48, Multilayer perceptron, decision table. These classifiers are implemented on weka tool and found that Multilayer perceptron has shown the highest accuracy.

Sedighi et al [5] developed a decision support system for chronic kidney disease prediction and they applied the feature selection techniques on kidney data. They observed that the classifier with feature selection technique has shown better performance. Chetty et al [6] demonstrated the role of feature selection in detecting the chronic kidney disease. They used the Wrapper method with Best first search for selecting the most relevant attributes and concluded that the classifiers on reduced data set have shown the best performance.

Kunwar et al [7] used the classifiers like Naive Bayes, Artificial Neural Network (ANN) in detecting the chronic disease. These classifiers are implemented on Rapid miner tool and concluded that Naive Bayes is giving the best performance. Tazin et al [8] used the classification algorithms like SVM, Decision Tree, k-Nearest Neighbor, Naive Bayes algorithm in chronic kidney disease detection. They used the ranking algorithm to select the most relevant attributes and compared the classifiers on Top 10,15,20,25 attributes. They observed that the classifiers on top 15 attributes has shown the better performance.Boukenze et al [9] used k-Nearest Neighbor, Decision Tree, Support Vector Machine , Multilayer perceptron, Bayesian Network classification algorithms for chronic disease prediction and they identified that Decision tree has given the best performance.

In [6, 8] Researchers has used Feature selection techniques for selecting the significant attributes and observed that the performance of the model is increased when significant attributes are considered. So in this paper we use Filter and Wrapper methods of feature selection techniques to select the significant attributes.

Researchers has applied various classification algorithms in chronic kidney disease prediction but there is no stability in terms of accuracy it is due to, while training model some instances may cover and some may be not, suppose if the test data consists of the instances which may not be covered the model while training. Then the classifier may predicts incorrectly. To overcome this problem, models need to be stabilized for that we use Bagging and Boosting classifiers which are also called ensemble classifiers.

A classification algorithm consists of parameters, randomly if we assign values to the parameters for a given classifier. Depends on the value, sometimes the performance of a classifier may be high sometimes it may be low. The performance of a classifier is fluctuating due to randomly assigning the parameter values, so to maintain the stability we need to give optimal values for a given parameter. To find the optimal values for a given parameter we use parameter tuning technique.

3. FEATURE SELECTION

Feature selection is the procedure of finding the subset of most significant features used in model construction. Alternative names for Feature selectionare variable selection, attribute selection, variable subset selection. The advantages of feature selection techniques are able to train the algorithm very faster, improves the performance of the model, avoids curse of dimensionality and it reduces the model complexity and makes the model very simple which can be easier to understand. Feature selection algorithms are classified into Filter methods, Wrapper methods and Embedded methods.

Filter Methods: Filter methods select the attributes independent of any algorithm.It uses statistical methods to assign scores to each feature for their correlation with the outcome variable. Examples of Filter methods are Information gain, Gain ratio, Consistency, OneR, Chi-Squared Test etc.

Pseudo code

```
Input: Feature set $S$
1. Identify candidate subset $s \subseteq S$
2. While! Stop criterion ()
2.1. Assess utility function $J$ using $s$.
2.2. Adopt subset $s$.
3. Return $s$.
```

Wrapper Methods:Wrapper methods use Best First, Stochastic and Heuristic Search techniques for selecting the subset of features. A model is trained using different combinations of feature subsets and score is assigned to features based on the model accuracy and the conclusions drawn from the previous model then a decision is considered whether to add or remove features from the subset. Examples of wrapper methods areForward Feature Selection, BackwardFeature Selection andRecursive Feature Elimination.

Pseudo code

```
Input: Feature set $S$
1. Identify candidate subset $s \subseteq S$
2. While! Stop criterion ()
2.1. Assess the error of a classifier using $s$.
2.2. Adopt subset $s$.
3. Return $s$.
```

Embedded Method: This method is a hybrid approach; it combines both Filter and Wrapper methods for selecting optimal feature subset.

4. CLASSIFICATION ALGORITHMS

The classification algorithms we use in the paper are TreeBag, AdaBoost, Gradient Boosting (GBM) and Random Forest.Under bagging technique TreeBag, Random Forest classifiers are used. AdaBoost, GBM comes under boosting technique.

4.1. TreeBag

TreeBag is an ensemble machine learning meta algorithm is used for both classification and regression models and it is an example of bagging technique.TreeBag uses decision trees classifiers, these decision trees are built on multiple sample data where samples are taken from random with replacement from the original data and the classifiers are aggregated to form an ensemble classifier.In order to predict the class label of test data, the test data is given as input to the ensemble classifier. Based on the outputs of the individual classifier, ensemble classifier takes the majority and assigns the class label to the test data.

Pseudo code

```
1. Repeat 'k' times.
1.1. Create a random training set.
1.2. Train a classifier on random training set.
2. To test, run each trained classifier.
3. Each classifier votes on the output then takes majority.
```

4.2. AdaBoost

AdaBoost is an example of boosting technique and it is an ensemble machine learning meta algorithm used for both classification and regression problems. AdaBoost produces strong classifier by combining weak classifiers; the strong classifier covers the instances which are not covered by individual weak classifiers so that the performance of model is produced. AdaBoost uses decision tree algorithms for classification and the procedure for AdaBoost classifier is given in the pseudo code. The decision tree in AdaBoost consists of one level tree called Decision stumps.

Pseudo code

```
1. The data set consists of (x1, y1), (x2, y2), (Xn, yn) instances, weights are assigned to all instances. $\forall i : D_0(i) \leftarrow 1/N$ where $N$ is the total number of instances.
2. A random sample $D_k$ is taken from the Data (D).
3. Train the classifier $h_k$ on $D_k$ (sample data).
4. Test the classifier $h_k$ on over all Data (D) and calculate the Error rate ($E_k$) on $h_k$.
5. Calculate the voting power of classifier on kth sample. $\alpha_k = \frac{1}{2}\log\left(\frac{1-E_k}{E_k}\right)$ here $\alpha_k$ is the voting power.
```

6. Update the weights of the instances which are incorrectly classified by classifier $h_{k..}$ $D_k(i) \leftarrow \frac{D_{k-1}(i)e^{-\alpha_k y_i h_k(x_i)}}{Z_k}$, Z_k is the normalization function such that $\sum_{i=1}^{N} D_k(i) = 1$.
7. Repeat step 2 to step 6 for *k* times, where *k* is the no of samples.
8. Finally an ensemble boosted classifier is obtained $H(x) = sign(\sum_{i=1}^{k} \alpha_i h_i(x))$ here *H(x)* is an ensemble classifier.

4.3. Gradient Boosting (GBM)

GBM is an ensemble machine learning meta algorithm which combines the weak classifiers and produces a strong classifier. It builds the model on the sample data and test the model on the original data then updates the weights of the instances which are incorrectly classified by the model. This process repeats till the maximum number of models are reached or the model achieves the highest accuracy.

Pseudo code

```
1. Initialize $f^{\wedge}{}_{\theta}$ with a constant.
2. For k=1 to M do
2.1. Compute the negative gradient $g_k(x)$.
2.2 Fit a model $h(x,\theta_k)$.
2.3 Choose the best gradient step-size $\rho_k$. $\rho_k = \arg\min_{\rho} \psi\left[y_{i,} f \wedge_{k-1}(x_i) + \rho h(x_{i,}\theta_k\right]$.
2.4 Update the function estimate. $f_k^{\wedge} \leftarrow f_{k-1}^{\wedge} + \rho_k h(x,\theta_k)$
3. End for.
```

4.4. Random Forest

Random Forest is an ensemble machine learning meta algorithm which is used for both classification and regression problems .Random Forest is similar to Bagging technique but with a slight change. In bagging classifiers are built on all the features of training set but in Random Forest classifiers are built on random features of training set. Random Forest follows a general rule it is "The number of random features (m) taken from all the features (p) of the training set (D) is m=square root(p)".

5. PROPOSED METHODOLOGY

In this section a methodology is proposed for classifying the chronic kidney disease.The steps involved in the Model Framework are:

Step 1: Dataset

Chronic Kidney dataset is taken from the UCI repository, this dataset consists of 25 attribute and 400 instances. This dataset is prepared form the previously diagnosed patients.

Step 2:Pre-processing

The dataset consists of missing values; this may cause trouble in analysis. There are several techniques to handle the missing values, some of them are ignores the missing values, replaces the missing values .The power of analysis is degraded when missing values are ignored; so to strengthen the analysis replace technique is chosen. We use KNN imputation algorithm to replace the missing values with the most frequent nearer observations.

Step 3: Feature selection

Once the data is pre-processed, feature selection technique is applied on the data set to get the significant attributes. We considered Filter and Wrapper methods of feature selection techniques. Under Filter methods we have chosen **Information gain, Gain ratio, Chi-squared test and consistency measure.** Under Wrapper method **RFE** measure is considered. In each measure of the Feature selection technique, we will consider the top 15 attributes this is because in [6] researcher applied ranking algorithm on the chronic data and observed the performance of the classifiers on top 5,10,15,20 and 25 attributes and observed that the models are giving promising results on top 15 attributes.

Step 4: Classification algorithms with Parameter Tuning

A parameter tuning is the process of finding the best parameter for an algorithm to boost its performance. There is two ways for parameter tuning: (i) Grid search parameter tuning (ii) Random search parameter tuning.In Grid search approach, user will specify the different values for a given parameter then model evaluates each value in the grid and give the best value among the given values. In Random search approach, parametric values are sampled from a random distribution for fixed number of iterations, and then model evaluates each parametric combination and give best parameter combination values. We will consider Random search parameter tuning technique, because in Grid search approach user should supply values which may not give optimal parameters.

Step 5: Performance analysis

TreeBag, AdaBoost, GBM and Random Forest classifiers are compared on each measure of the feature selection technique and the classifier which shows better promising results is identified.

6. RESULT ANALYSIS

To identify the best ensemble classifier for chronic kidney disease prediction we performed experiments in R.

6.1. Tree Bag

TreeBag classifier is based on bagging technique. Performance of the TreeBag classifier is compared on each measure of filter and wrapper methods which can be viewed in Table 1.

Table 1. TreeBag Classifier Performance on Filter and Wrapper Methods

Feature selection	Kappa	Accuracy (%)
Filter method		
Information gain	0.9349	96.97
Gain ratio	0.9349	96.97
Chi-squared test	0.9349	96.97
Consistency	0.9564	96.97
Wrapper method		
RFE	0.9564	97.98

6.2. Ada Boost

AdaBoost classifier is based on boosting technique and Random search parameter tuning is applied on AdaBoost parameters (i) nIter (number of iterations) and (ii) Method can be viewed in Table 2.

Table 2. Performance of the AdaBoost Classifier on Each Measure of Filter and Wrapper Methods

Feature selection	Parameters		Kappa	Accuracy (%)
Filter methods	**nIter**	**Method**		
Information gain	50	AdaBoost.M1	0.9349	96.97
Gain ratio	50	Real AdaBoost	0.9564	97.98
Chi-squared test	50	AdaBoost.M1	0.9349	96.97
Consistency	50	AdaBoost.M1	0.9349	96.97
Wrapper method				
RFE	50	Real AdaBoost	0.9568	97.98

6.3. GBM

GBM a Gradient Boosting Machine learning algorithm is based on boosting technique and Random search tuning technique is applied with the GBM parameters n.trees (number of trees), Interaction.depth, shrinkage and n.minobsinnode (minimum number of observations in terminal node of a tree). Parameters Shrinkage and Interaction. Depth default sets o the values of 0.1 and 10 as shown in Table 3.

Table 3. GBM Classifier Performance on Each Measure of Filter and Wrapper Methods

Feature selection	Parameters		Kappa	Accuracy (%)
Filter methods	**n.trees**	**Interaction. Depth**		
Information gain	150	1	0.9785	99
Gain ratio	100	2	0.9785	99
Chi-squared test	150	2	0.9785	99
Consistency	150	1	0.9785	99
Wrapper method				
RFE	50	2	0.9783	99

6.4. Random Forest

Random Forest is an ensemble classifier and random search technique is applied with the parameter mtry. Performance of random forest classifier on each methods can be viewed in Table 4.

Table 4. Performance of Random Forest Classifier on Each Measure of Filter and Wrapper Methods

Feature selection	parameter Mtry	Kappa	Accuracy (%)
Filter method			
Information gain	2	0.9783	99
Gain ratio	2	0.9783	99
Chi-squared test	2	0.9783	99
Consistency	2	0.9341	96.97
Wrapper method			
RFE	2	0.9568	97.98

6.1.1. Comparison of performance of TreeBag, AdaBoost, GBM and Random Forest classifiers on each measure of Filter methods and Wrapper methods

Ensemble classifiers TreeBag, AdaBoost, GBM and Random Forest are compared on each measure of Feature selection techniques which can be shown in the Figure 1,2,3,4 and 5.

Performance of the classifiers at Information Gain Measure

TreeBag, AdaBoost, GBM and RandomForest classifiers are compared at Information gain measure which is shown in the Figure 1 and observed that GBM, Random Forest has given highest accuracy of 99% when compared to other classifiers.

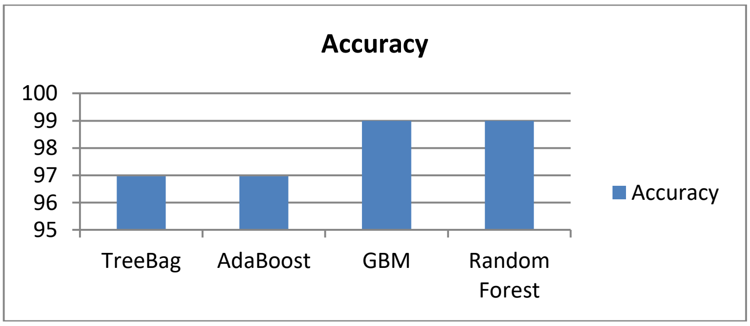

Figure 1. Comparison of classifiers at information gain measure

Performance of the classifiers at Gain ratio measure

TreeBag, AdaBoost, GBM and RandomForest classifiers are compared at Gain ratio measure which is shown in the Figure 2 and observed that GBM, Random Forest has given highest accuracy of 99% when compared to another classifiers.

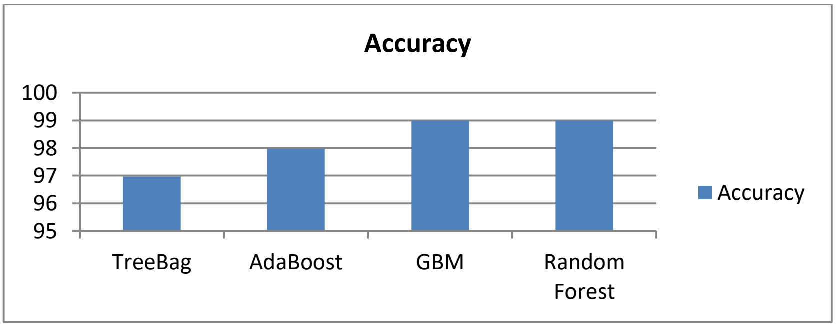

Figure 2. Comparison of classifiers at gain ratio measure

Performance of the classifiers at Chi-squared test

TreeBag, AdaBoost, GBM and RandomForest classifiers are compared at Chi-squared test measure which is shown in the Figure 3 and observed that GBM, Random Forest has given highest accuracy of 99% when compared to other classifiers.

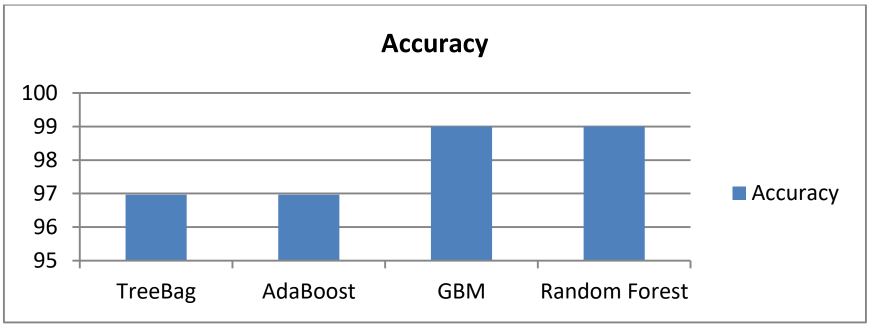

Figure 3. Comparison of classifiers at chi-squared test measure

Performance of the classifiers at Consistency measure

TreeBag, AdaBoost, GBM and RandomForest classifiers are compared at consistency measure which is shown in the Figure 4 and observed that GBM has given highest accuracy of 99% when compared to another classifiers.

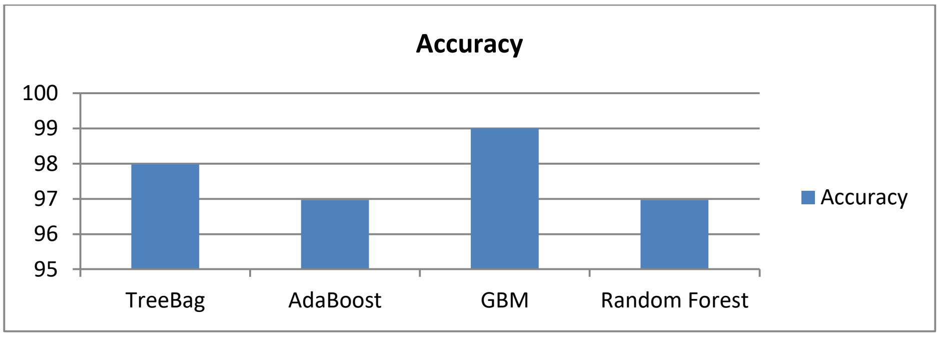

Figure 4. Comparison of classifiers at consistency measure

Performance of the classifiers at RFE measure

TreeBag, AdaBoost, GBM and RandomForest classifiers are compared at RFE measure which is shown in the Figure 5 and observed that GBM has given highest accuracy of 99% when compared to another classifiers.

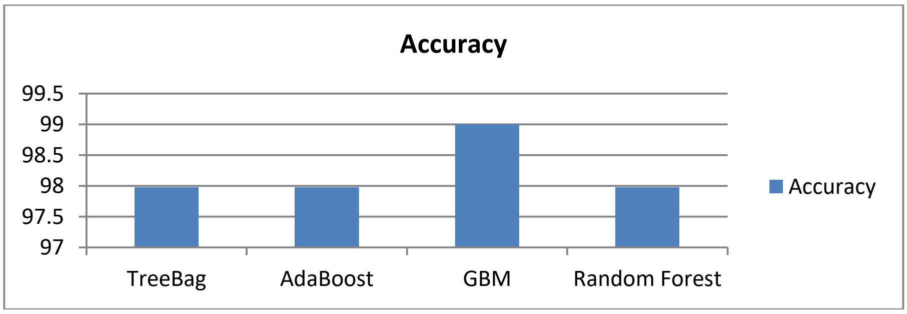

Figure 5. Comparison of classifiers at RFE measure

TreeBag, AdaBoost, GBM and Random Forest classifiers are compared at each measure of Filter and Wrapper methods. From the analysis it is identified that GBM is giving consistent accuracy of 99% at each measure.

6.2.2. Comparison of GBM with normal classifiers

From the observations on prediction of chronic kidney disease using ensemble classifiers like TreeBag, AdaBoost, GBM and Random Forest. GBM classifier has shown best promising results which can be viewed in Table 5.

Table 5. Comparison of GBM with the Existed Classifiers

Approach	**Accuracy (%)**
Naïve Bayes	96
SVM	98.5
Decision Tree	98
KNN	97.5
GBM	**99**

GBM classifier is compared with normal classifiers and it is shown that GBM with more stability can able to predict the kidney disease with 99% accuracy.

7. CONCLUSION

In this Manuscript, prediction of chronic kidney disease has been done by using ensemble classifiers like TreeBag, AdaBoost, Gradient Boosting (GBM) and Random Forest to attain more stability in prediction. A Random search parameter tuning technique is applied along with the ensemble classifiers to get the optimal parameters and feature selection techniques are applied to retrieve the significant features. We have chosen the different measures of features selection techniques and we compared the ensemble classifiers at each measure and found that GBM is giving highest accuracy of 99% at all measures of feature selection techniques.we conducted the experiments on small amount of data (400 instances)but data is increasing day to day . In future, enormous amount of data will generates on chronic kidney disease so we would like to extend our work on handling such large scale data for disease prediction.

REFERENCES

[1] Rubini, L. Jerlin, and P. Eswaran. "Generating comparative analysis of early stage prediction of Chronic Kidney Disease." *International OPEN ACCESS Journal of Modern Engineering Research*. Vol. 5, no. 7 (2015): 49-55.

[2] Baby, P. Swathi, and T. Panduranga Vital. "Statistical analysis and predicting kidney diseases using machine learning algorithms." *International Journal of Engineering Research and Technology*, Vol. 4, no. 07, (2015): 206-210.

[3] Vijayarani, S., and S. Dhayanand. "Data mining classification algorithms for kidney disease prediction."*International Journal on Cybernetics and Informatics (IJCI)*, Vol. 4, No. 4, (2015): 13-25.

[4] Jena, Lambodar, and Narendra Ku Kamila. "Distributed data mining classification algorithms for prediction of chronic kidney disease."*International Journal of Emerging Research in Management and Technology*, Vol. 4, no. 11 (2015): 110-118.

[5] Sedighi, Zeinab, HosseinEbrahimpour-Komleh, and SeyedJalaleddinMousavirad. *"Featue selection effects on kidney desease analysis."* In Technology, Communication and Knowledge (ICTCK), 2015 International Congress on, pp. 455-459.IEEE, 2015.

[6] Chetty, Naganna, Kunwar Singh Vaisla, and Sithu D. Sudarsan. *"Role of attributes selection in classification of Chronic Kidney Disease patients."*In Computing, Communication and Security (ICCCS), 2015 International Conference on, pp. 1-6.IEEE, 2015.

[7] Kunwar, Veenita, KhushbooChandel, A. SaiSabitha, and AbhayBansal. *"Chronic Kidney Disease analysis using data mining classification techniques."* In Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference, pp. 300-305. IEEE, 2016.

[8] Tazin, Nusrat, ShahedAnzarusSabab, and MuhammedTawfiqChowdhury. *"Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique."* In Medical Engineering, Health Informatics and Technology (MediTec), 2016 International Conference on, pp. 1-6.IEEE, 2016.

[9] Boukenze, Basma, Abdelkrim Haqiq, and Hajar Mousannif. *“Predicting Chronic Kidney Failure Disease Using Data Mining Techniques.”* International Symposium on Ubiquitous Networking. Advances in Ubiquitous Networking 2, pp. 701-712.Springer Singapore, 2017.