International Journal of Engineering & Technology, 7 (4.36) (2018) 1161-1167

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Machine Learning-Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique

Zixian Wang[1*], Jae Won Chung[2], Xilin Jiang[3], Yantong Cui[4], Muning Wang[5], Anqi Zheng[5]

[1]Cornell University, USA.
[2]Hankuk University of Foreign Studies, Korea.
[3]University of Illinois Urbana-Champaign, USA.
[4]Anshan No.1 H.S., China.
[5]High School Affiliated to Liaoning Normal University, China.
**Corresponding author E-mail:ibroseme88900@gmail.com*

Abstract

Technological development, including machine learning, has a huge impact on health through an effective analysis of various chronic diseases for more accurate diagnosis and successful treatment. Kidney disease is a major chronic disease associated with aging, hypertension, and diabetes, affecting people 60 and over. Its major cause is the malfunctioning of the kidney in disposing toxins from the blood. This study analyzes chronic kidney disease using machine learning techniques based on a chronic kidney disease (CKD) dataset from the UCI machine learning data warehouse. CKD is detected using the Apriori association technique for 400 instances of chronic kidney patients with 10-fold-cross-validation testing, and the results are compared across a number of classification algorithms including ZeroR, OneR, naive Bayes, J48, and IBk (k-nearest-neighbor). The dataset is preprocessed by completing and normalizing missing data. The most relevant features are selected from the dataset for improved accuracy and reduced training time. The results for selected features of the dataset indicate 99% detection accuracy for CKD based on Apriori. The identified technique is further tested using four patient data samples to predict their CKD.

Keyword: *Machine Learning, Classification Technique, Prediction System*

1. Introduction

Kidney disease is considered a major problem for people 60 and above. The major cause is the degeneration of the kidney that reduces the rate of glomerular filtration. This problem, when lasting more than three months, is generally considered as chronic kidney disease (CKD) [1]. CKD is ranked as the 10th major cause of death in the world. Hypertension, diabetes, and aging are considered leading causes of CKD, in addition to other factors such as high blood pressure, coronary artery disease, and anemia. If the problem can be detected in early stages, then it is considered feasible to save kidney function for the longer survival of the patient. Early diagnosis of CKD can facilitate its treatment and help avoid costly treatment procedures such as dialysis and transplants.

With machine learning techniques, it is possible to analyze lab records and other information on patients for the early detection of CKD [23]. Low-level data can be transformed into high-level knowledge through the knowledge discovery in databases (KDD) [2]. This transformation can help practitioners better understand CKD patterns for its early diagnosis.

This study analyzes CKD using machine learning techniques using a CKD dataset from the UCI machine learning data warehouse. CKD is detected using the Apriori association technique for 400 instances of chronic kidney patients with 10-fold-cross-validation testing, and the results are compared across a number of classification algorithms including ZeroR, OneR, naive Bayes, J48, and IBk (k-nearest neighbor). The dataset is preprocessed by completing and normalizing missing data. The most relevant features are selected from the dataset to improve accuracy and reduce training time for machine learning techniques. A set of experiments is conducted using various WEKA-implemented machine learning techniques to detect CMD based on the CKD dataset from the UCI machine [21]. The results are compared for detection accuracy across different machine learning techniques.

The rest of this paper is organized as follows: Section 2 describes machine learning techniques. Section 3 presents the state of the art in the field for CKD detection. Section 4 presents the proposed work for investigating machine learning techniques. Section 5 reports the results, and Section 6 concludes with some avenues for future research.

2. Machine Learning Techniques

Machine learning is a data analysis method automating the analytical model-building process. It is based on the idea that a system can learn from data, identifying key patterns for better decision making with minimal human intervention. Machine learning techniques can be unsupervised or supervised [3]. Supervised algorithms require machine learning skills to provide some input and desired output as well as feedback on prediction accuracy during algorithm training. Data scientists determine the variables to be analyzed and used by a model for predictions. When the training is complete, the algorithm applies what it learns outcome data, using an iterative approach called deep learning for conclusions.

to new data. Unsupervised algorithms need no training for desired

This study uses supervised machine learning techniques including ZeroR, OneR, IBk, J48, naive Bayes, and k-nearest neighbor to classify the CKD dataset disease detection [23]. These techniques are described as follows:

1. Apriori association technique: This technique uses a bottom-up approach to mine recurrent items in a given dataset, and its general workings can be described as follows:

I_k : Candidate itemset of size k
F_k: Frequent itemset of size k

F_1 = {frequent items};
for (k = 1; F_k != ∅; k++)
do begin
I_{k+1} = candidates generated from F_K;
for each transaction *t* in the database *do*
increment count of all candidates in I_{k+1} that are contained in *t*
F_{k+1} = candidates in I_{k+1} with min_support
end
return $\cup_k F_k$;

2. ZeroR technique: This is the simplest classification method serving as the baseline performance for other classifiers. It ignores all predictors, relying only on target values. During the classification, it produces a frequency table corresponding to target values to select the most frequent values.
3. OneR technique: This technique generates "One Rule for One Predictor" for various predictors in the dataset to select the one with the minimum combined error for the predictor [4], and its workings can be explained as follows:

Count= value of target
While (Number_of_Predictor! =NULL)
{
For every value of the predictor
{
If (Target value appears)
++Count;
Calculate most frequent class
Create a Rule corresponding to that class as per value calculated in the above step for the predictor
}
Calculate total error of the rules
}
The predictor is finalized based upon the smallest total error.

4. IBk technique: This is an instance-based classification method using k-neighborhood nodes that produces output with mean absolute error, confusion matrix, and relative absolute error, among others. It classifies a given dataset based on the Euclidean distance, which determines the class of an unknown sample and is computed using the value of k as follows [5]:

$$d(y,z) = \sqrt{\Sigma_{i=1}^{k}(y_i - z_i)^2} \quad (1)$$

5. J48 technique: This is a decision tree-based technique that creates a tree by continuously breaking down of a given dataset into smaller subsets. The resultant tree includes leaf (terminating) nodes and decision nodes. The top node is called the root node and is the most significant in making a major contribution to predictions. This technique can address numerical as well as categorical data.
6. Naive Bayes technique: This technique works based on the Bayes theorem along with the liberation convention among predictors and is helpful for classifying large datasets because its noncomplicated iterative parameters. It can be represented as follows [6, 22]:

$$P(c|y) = \frac{P(y|c)P(c)}{P(y)}, \quad (2)$$

$P(c \mid y) = P(y_1 \mid c) \times P(y_2 \mid c) \times P(y_3 \mid c) \times P(y_n \mid c) \times P(c)$ (3)
where the posterior probability for the *predictor* and *class* is $P(c \mid y)$; the prior probability of *class* is P(c); the probability of predictor for a class is $P(y \mid c)$; and the prior probability of *predictor* is P(y).

3. Literature Survey

Many studies explore and analyze chronic diseases using various techniques for early diagnosis. Patil [13] surveys various data mining techniques for their detection accuracy, including logistic regression, multilayer perception, ANN, decision table, radical basic function, naive Bayes, k-nearest neighbor, and sequential minimal optimization. Depending on the type of dataset, such techniques show differences in the level of accuracy, and there is no single rule for the best result.

Dulhare and Ayesha [1] use the naive Bayes classifier with OneR as the attribute selector to predict CKD using a dataset from the UCI digital repository with 25 attributes, where 11 are numeric, 13 are nominal, and 1, a class attribute. They reduced the attribute number by 80% through OneR for a 12.5% increase in detection accuracy.

Gopika and Vanitha [2] employ a clustering technique for accurate CKD detection and reduced diagnosis time. They use fuzzy c-means, k-means, and k-medoids techniques and show 87% accuracy using the fuzzy c-means clustering technique for a dataset from the UCI machine repository.

Charleonnan et al. [5] employ decision tree, logistic regression, support vector machine (SVM), and k-nearest neighbor (KNN) as classifiers for CKD detection using a dataset with 2 classes, 400 instances, and 24 attributes. They use a CKD dataset from the UCI machine learning repository. The results show the SVM technique as the better detection technique for detection accuracy and sensitivity.

Ramya and Radha [7] examine kidney function failure using classification algorithms. According to case severity, they classify according to different stages of kidney disease using random forest, radial basic function, and back-propagation neural network. They evaluate different techniques for performance metrics including specificity, kappa, sensitivity, and accuracy and use a dataset from the Coimbatore state for about 1000 patients with 15 attributes, concluding the radical basic function to be the best classifier with 85.3% detection accuracy.

Padmanaban and Parthiban [8] explore early kidney disease detection using machine learning techniques and a dataset of 600 instances to validate decision tree and naive Bayes, achieving 91% detection accuracy through the decision tree and reporting 95% sensitivity and 94% specificity with the decision tree for CKD detection.

Iqbal et al. [9] use texture analysis techniques to analyze ultrasound images of kidney disease to distinguish between normal and kidney disease patients. They use mathematical operations such as the Fourier analysis to calculate the root mean square (RMS), homogeneity, average values, and a gray-level correlation matrix (GLCM) through MATLAB. They consider ultrasound images of 32 patients and distinguish between normal and kidney disease patients using RMS and cortex region values as 0.3 and 0.0049, respectively.

Kayaalp et al. [10] employ hybrid classification techniques to analyze kidney disease using a dataset from the UCI machine learning repository with information on about 400 patients. They employ support vector machine and k-nn classifier and conduct feature selection using the relief and gain ratio algorithm for the most relevant feature in the dataset. They conclude that the k-nn algorithm provides better performance for selected features in

comparison to other algorithms in terms of f-measure, precision, and contrast matrix.

Wibawa et al. [11] use feature selection and boosted classifier to diagnose CKD and employ AdaBoost for ensemble learning and correlation-based feature selection (CFS). They use naive Bayes, k-nearest neighbor (KNN), support vector machine for CKD detection and conclude AdaBoost and CFS as the most promising classifiers in addition KNN and naive Bayes classifiers in CKD detection. They achieve 0.98 f-measure rate, 0.98 recall rate, and 0.981 accuracy rate.

Wickramasinghe et al. [12] control CKD using an appropriate diet plan and recommend diet plans to different patients through their classification method. They recommend a diet plan based on the patient's blood potassium level. They use multicast neural network, multiclass decision jungle, multiclass logistic regression, multiclass decision forest and achieve 99.17% accuracy through the multiclass decision forest algorithm.

Previous research suggests that machine learning provides important insights into data and can help classify data into different classes. The findings indicate that machine learning techniques can produce accurate classification results if used in conjunction with feature selection techniques. Therefore, retaining the benefits of classification results for machine learning techniques, this study employs a set of the most popular machine learning techniques in combination with feature selection technique to classify normal and kidney disease patients.

4. The Proposed Work

The main objective of this study is to investigate machine learning techniques in combination with feature selection techniques for effective CKD detection in terms of detection accuracy. The study proposes various prediction models using classification algorithms with different techniques offered by the WEKA tool and compares them for correctly classified instances. The identified classification technique can provide predicted values for early CKD diagnosis.

4.1. The Proposed Framework

The proposed framework for developing prediction machine learning models and their comparison are depicted in Fig. 1. The main objective of the present research is to propose a machine learning technique to predict CKD using associative and classification algorithms. The proposed technique generates classification association rules (CARs) to determine techniques with a high percentage of correctly classified instances, and identified classifiers can facilitate early CKD diagnosis. A comparative analysis of the proposed technique is performed using other state-of-the-art techniques. Fig. 1 briefly details various stages:

i. Dataset selection stage: The dataset is selected to predict CKD for data analysis and effective knowledge. Enough data are required to implement a machine learning technique for a selected dataset. In this set of experiments, CKD data are obtained from the UCI machine learning repository.

ii. Preprocessing and transformation stage: The dataset is prepared in attribute-relation file format with 16 attributes. The dataset is converted into a binomial format to implement associative techniques. In addition, missing records, duplicate records, and unnecessary fields are removed for a standard data format.

iii. Feature selection stage: The most promising features of the CKD dataset are selected using the WEKA tool for better results. Feature evaluators and search methods are used for this purpose. The correlation-based feature selection subset evaluator is used as the feature evaluator, and the greedy stepwise search method is used. The selected features include blood pressure, red blood cells, pus cell, serum creatinine, haemoglobin, hypertension, diabetes mellitus, appetite, and pedal oedema for better results from the CKD dataset.

iv. Selection of associative rules: The Apriori association algorithm is implemented, and 10 best rules are selected to prepare the training dataset to implement different classification algorithms.

v. Implementation of classification algorithms: The five classifiers are trained using the dataset selected based on association rules including k-nearest neighbor, naive-Bayes, ZeroR, OneR, and J48.

vi. Performance evaluation stage: k-nearest neighbor, naïve Bayes, ZeroR, OneR, and J48 are trained and tested using the identified CKD dataset, and the performance of each classifier is measured for correctly classified instances of the identified dataset.

vii. Disease prediction system: The identified best classifier helps to form an intelligent CKD prediction system (ICKDPS) for the accurate prediction of other chronic diseases such as heart disease.

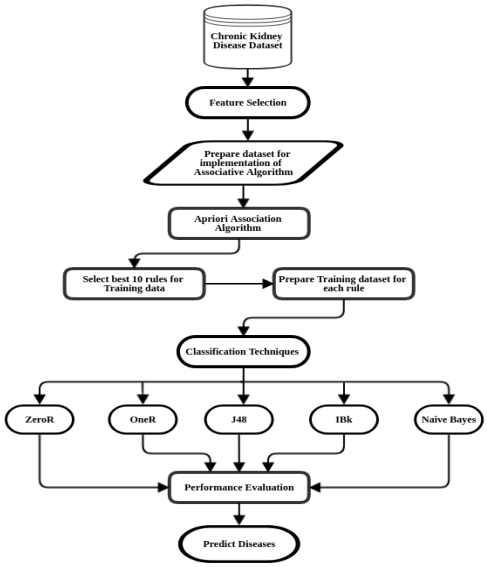

Fig. 1: Flow of proposed work

4.2. Benchmark Dataset

A dataset with a total of 400 instances with 16 selected attributes is used. The dataset is obtained from the Machine Learning Repository [14]. The attribute "class" is a measurable field with the value "ckd" and indicates an individual with CKD, and "nonckd" indicates an individual with no CKD. Table 1 shows the attributes, descriptions, and values for the CKD dataset. The dataset has 250 "ckd" and 150 "nonckd" instances.

Table 1: Description of attributes

Attributes	Description
Age	Range [2 -90] In the year
Blood pressure	Range [50 - 180] In mm Hg
Red Blood Cell	having two nominal value "normal" or "abnormal"
Pus Cell	having two nominal value "normal" or "abnormal"
Bacteria	having two nominal value Bacteria is "present" and "not present"
Serum Creatinine	Numerical value in mgs/dl
Haemoglobin	The numerical value in gms
Hypertension	having two nominal value "yes" and "no"
Diabetes Mellitus	having two nominal value "yes" and "no"
CoronaryArtery Disease	having two nominal value "yes" and "no"
Appetite	having two nominal value Appetite is "good" and "poor"
Pedal Edema	having two nominal value Pedal Edema is "yes" and

	"no"
Anaemia	having two nominal value Pedal Edema is "yes" and "no
Class	having the class value "ckd" represent Chronic Kidney Disease and "nonckd" represent Chronic Kidney Disease not present

5. Results & Discussion

A set of experiments is conducted using the identified benchmark dataset with different classification techniques implemented in WEKA. The results are compared for correctly classified instances. The evaluation of results is based on the following criteria:

1. Incorrectly classified instances, correctly classified instances, kappa statistic, and mean absolute error rate for different classifiers with and without the Apriori association algorithm using 10-fold-cross-validation testing are compared. The results are shown in Tables 2 and 3.
2. The results are compared for the accuracy of the CKD dataset from the UCI Machine Learning database, as shown in Table 4.
3. Four patient sample datasets are tested to predict CKD using the best classification technique, as shown in Table 5.

The results of the proposed framework are calculated using the WEKA tool. Table 2 compares different classifiers without the Apriori association algorithm with the 10-fold cross-validation testing option. Table 3 compares different classifiers for the Apriori association algorithm on the CKD dataset. The error rate plays no role in classification and is used for numeric prediction. The J48 algorithm uses the decision tree for classification, and Fig. 2 shows a sample decision tree created using J48.

Table 2: Comparison of Results for Classifiers on CKD Dataset

Classifiers	Correctly Classified Instances (%)	In-correctly Classified Instances (%)	Kappa statistic	Mean absolute error
ZeroR	62.5	37.5	0	0.4689
OneR	87.5	12.5	0.7468	0.125
J48	96	4	0.9153	0.0649
IBk	94.5	5.5	0.886	0.0491
Naïve bayes	96.5	3.5	0.9267	0.0397

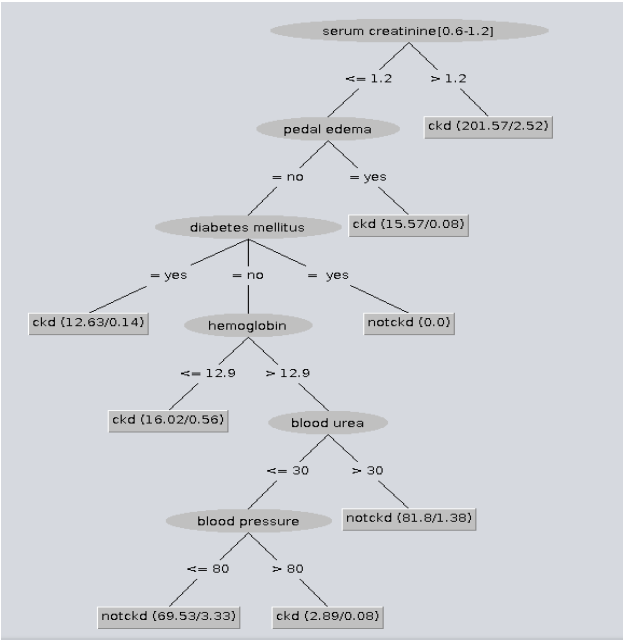

Fig. 2: Decision tree for classification

Table 3: Comparison of Results for Classifiers Using Apriori on CKD Dataset

Classifiers	Correctly Classified Instances (%)	In-correctly Classified Instances (%)	Kappa Statistic	Mean absolute error
ZeroR	56	44	0	0.4929
OneR	92	8	0.8316	0.08
J48	98.33	1.67	0.966	0.0186
IBk	99	1	0.9798	0.0109
Naïve Bayes	98.33	1.67	0.9663	0.014

Fig. 3 shows the results for ZeroR, OneR, J48, IBk, and naive Bayes with and without the Apriori associative algorithm for correctly classified instances. The figure shows the results for OneR, J48, IBk, and naive Bayes to improve through Apriori but not for ZeroR. The IBK (k-nearest neighbor) shows the best value of 99% accuracy with the Apriori associative algorithm.

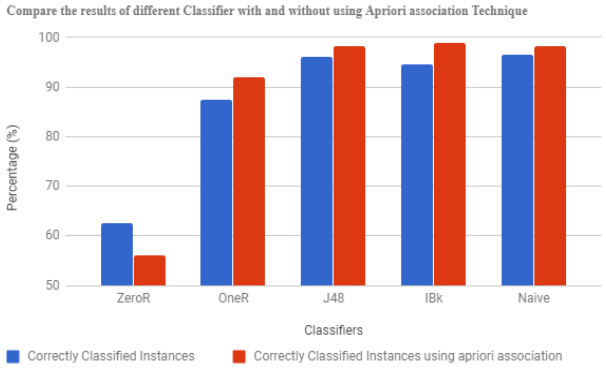

Fig. 3: Results with and without Apriori algorithm

Table 4: Comparison of Results from Previous Studies

Author	Tool Used	Techniques	Accuracy (%)
Ramya and Radha [7]	WEKA	K-Mean	86
Sinha & Sinha [15]	WEKA & Orange	SVM KNN	73 78
Khan and Westin [16]	WEKA	Naive Bayes J48 KNN	90.4 82.5 84.1
Abeer & Ahmad [17]	WEKA	SVM	93.14
Jena & Kamila [18]	WEKA	Naive Base SVM J48	95 62 99
Vijayarani and Dhayanand [19]	MATLAB	Naive Bayes SVM	70.96 76.32
Kumar [20]	WEKA	SMO Naive Bayes RBF MLPC	97 95 98 98
The proposed technique	WEKA	IBk with Aprior Algorithm	99

Table 4 compares the results from previous studies and finds the proposed method to show 99% accuracy using the IBk classifier with the Apriori association algorithm. Table 5 shows the four patient samples to test the proposed approach for the CKD risk level.

Table 5: Sample Data for Predicting CKD Risk Level

Attributes	Sample 1	Sample 2	Sample 3	Sample 4
Age	57	38	46	31
Blood pressure	152 {High}	75 {Low}	137 {High}	110 {Normal}
Red blood Cell	Abnormal	Normal	abnormal	Normal
Pus Cell	Abnormal	abnormal	Normal	Normal
Bacteria	Present	notpresent	present	notpresent
Serum Creatinine	3.8 {High}	0.5 {Low}	1.7 {High}	0.9 {Normal}
Hemoglobin	8 {low}	13.5	9.5	12.5

		{Normal}	{Low}	{Normal}
Hypertension	Yes	no	Yes	no
Diabetes Mellitus	Yes	Yes	No	No
Coronary Artery Disease	No	No	Yes	No
Appetite	Poor	Good	Good	Good
Pedal Edema	Yes	No	Yes	No
Anaemia	Yes	No	Yes	No
Prediction Result (Disease Risk Level)	High	Low	Medium	Normal

6. Conclusion and Future Scope

This study investigates various machine learning techniques, particularly classification and association techniques, to predict CKD. The study analyzes the effects of using feature selection techniques in combination with classification techniques. Classification techniques implemented in WEKA are used to benchmark the CKD dataset. The results are computed using 10-fold cross-validation with and without the feature selection technique. The results are compared for correctly classified instances, kappa statistic, and mean absolute value with and without the feature selection technique. The benchmark dataset is prepared using the Apriori association algorithm. The extracted data are further used to validate ZeroR, OneR, J48, IBk, and naive Bayes implemented in WEKA. The results note that the best result can be achieved using IBk with the Apriori associative algorithm for 99% accuracy. Future research should analyze different supervised and unsupervised machine learning techniques and feature selection techniques with additional performance metrics for better CKD prediction.

References

[1] Dulhare UN & Ayesha M, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier", *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, (2016), pp.1-5.

[2] Gopika M, "Machine learning Approach of Chronic Kidney Disease Prediction using Clustering Technique", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol.6, No.7, (2017), pp.14488 14496.

[3] The relationship between machine learning and datamining, (2018).

[4] One R. An Introduction to Data Science, (2018).

[5] Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S & Ninchawee N, "Predictive analytics for chronic kidney disease using machine learning techniques", *IEEE International Conference on Management and Innovation Technology (MITicon)*, (2016), pp.MIT-80.

[6] Saltz JS & Stanton JM, *An introduction to data science*, SAGE Publications, (2017).

[7] Ramya S & Radha N, "Diagnosis of chronic kidney disease using machine learning algorithms", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol.4, No.1, (2016), pp.812-820.

[8] Padmanaban KA & Parthiban G, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease", *Indian Journal of Science and Technology*, Vol.9, No.29, (2016), pp.1-5.

[9] Iqbal F, Pallewatte AS & Wansapura JP, "Texture analysis of ultrasound images of chronic kidney disease. *IEEE Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, (2017), pp.1-5.

[10] Kayaalp F, Basarslan MS & Polat K, "A hybrid classification example in describing chronic kidney disease", *IEEE Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, (2018), pp.1-4.

[11] Wibawa MS, Maysanjaya IMD & Putra IMAW, "Boosted classifier and features selection for enhancing chronic kidney disease diagnose", *IEEE 5th International Conference on Cyber and IT Service Management (CITSM)*, (2017), pp.1-6.

[12] Wickramasinghe MPNM, Perera DM & Kahandawaarachchi KADCP, "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms", *IEEE Conference on Life Sciences (LSC)*, (2017), pp.300-303.

[13] Patil PM, "Review on Prediction of Chronic Kidney Disease using Data Mining Techniques", *International Journal of Computer Science and Mobile Computing*, Vol.5, No.5, (2016), pp.135-141.

[14] Jerlin L & Rubini DA, *Chronic Kidney Disease*, Retrieved from UCI Machine Learning Repository, (2018).

[15] Sinha P & Sinha P, "Comparative study of chronic kidney disease prediction using KNN and SVM", *International Journal of Engineering Research and Technology*, Vol.4, No.12, (2015), pp.608-612.

[16] Khan SH, Predictive models for chronic renal disease using decision trees, naïve bayes and case-based methods, (2010), pp.1-40.

[17] Al-Hyari AY, Ahmad MA & Majid AA, "Diagnosis and classification of chronic renal failure utilising intelligent data mining classifiers", *International Journal of Information Technology and Web Engineering (IJITWE)*, Vol.9, No.4, (2014), pp.1-12.

[18] Jena L & Narendra KK, "Distributed data mining classification algorithms for prediction of chronic-kidney-disease", *International Journal of Emerging Research in Management &Technology*, Vol.4, No.11, (2015), pp.110-118.

[19] Vijayarani S & Dhayanand S, "Data mining classification algorithms for kidney disease prediction", *International Journal on Cybernetics and Informatics (IJCI)*, (2015), pp.13-25.

[20] Kumar M, "Prediction of chronic kidney disease using random forest machine learning algorithm", *International Journal of Computer Science and Mobile Computing*, Vol.5, No.2, (2016), pp.24-33.

[21] Kumar G & Kumar K, "The use of multi-objective genetic algorithm based approach to create ensemble of ann for intrusion detection", *International Journal of Intelligence Science*, Vol.2, No.4, (2012), pp.115-127.

[22] Kumar G & Kumar K, "A novel evaluation function for feature selection based upon information theory. *IEEE 24th Canadian Conference on Electrical and Computer Engineering (CCECE)*, (2011), pp.000395-000399.

[23] Kumar G, Kumar K & Sachdeva M, "The use of artificial intelligence based techniques for intrusion detection: a review", *Artificial Intelligence Review*, Vol.34, No.4, (2010), pp.369-387.