ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 11 , November 2017

J48 Decision Tree and Novel Genetic Algorithm Framework for Chronic Kidney Disorder (CKD)

P. Suguna, Dr. S. Prema

M. Phil Research Scholar, Department of Computer Science (PG), K.S.Rangasamy College of Arts and Science (Autonomous), Tiruchengode, Tamilnadu, India

Associate Professor, Department of Computer Science (PG), K.S.Rangasamy College of Arts and Science (Autonomous), Tiruchengode, Tamilnadu, India

ABSTRACT: Chronic kidney diseases have become a major public health problem. Chronic diseases are a leading cause of morbidity and mortality in India. Chronic kidney diseases account for 60% of all deaths worldwide. Eighty percentage of chronic disease deaths worldwide occur in low- and middle-income countries. The National Kidney foundation determines the different stages of chronic kidney disease based on the presence of kidney damage and glomerular filtration rate (GFR), which is measure a level of kidney function. The contributions of this paper are: preprocessing noisy data, preparatory to applying machine learning algorithms, enhancing the automated detection of glomerular filtration rate (GFR) variability, a serious problem for patients with CKD and predicting patient CKD feature levels in order to preemptively detect and avoid potential health problems. These contributions expand the scope of optimizing the performance of the CKD diagnosis and potentially enable new clinical applications for CKD management. The diagnosis of CKD is very important now days using various types of techniques. Here, there are various techniques, their classification and implementation using various types of software tools and techniques. Proposed method seems to outperform the available classifiers in the literature. Addressing the CKD seems to be most important aspect as of witnessed from these reports. The proposed strategy and adopted classifiers seems to perform well for the CKD prediction. This strategy hikes and optimizes the overall performance. Among the experimented classifiers the Genetic Algorithm seems to vary apt for the problem. Hence, it is clearly evident that Genetic algorithm classifier seems to be optimal for the CKD prediction problem

KEYWORDS: Data mining; chronic diseases; Genetic Algorithm; glomerular filtration rate; classification

I. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD). Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. In data mining a Genetic algorithm can be used either to optimize parameters for other kind of data mining algorithms or to discover knowledge by itself. The advantage of Genetic algorithm becomes more obvious when the search space of a task is large. Genetic algorithms are a probabilistic search and evolutionary optimization approach which is inspired by Darwin's theory about evolution. And this technique used in computing to find exact or approximate solution to optimization and search problems. Existing Data mining approaches have not combined sequential algorithms and decision making systems

- The Dataset collected in real time in medical systems requires a higher level of preprocessing.
- A balance between classification and Decision Tree is needed for integration.
- A generalized Data mining framework that can be generalized across different systems does not exist.
- Health care domains need the support of the techniques of data mining for quick decision making.

II. RELATED WORK

Manish Kumar (Feb, 2016) explained Chronic Kidney Disease prediction is one of the most central problems in medical decision making because it is one of the leading cause of death. So, automated tool for early prediction of this disease will be useful to cure. In this study, the experiments were conducted for the prediction task of Chronic Kidney Disease obtained from UCI Machine Learning repository using the six machine learning algorithms, namely: Random Forest (RF) classifiers, Sequential Minimal Optimization (SMO), NaiveBayes, Radial Basis Function (RBF) and Multilayer Perceptron Classifier (MLPC) and Simple Logistic (SLG). S.Ramya et al (jan 2016) produces the work to reduce the diagnosis time and to improve the diagnosis accuracy using classification algorithms. The proposed work deals with classification of different stages in chronic kidney disease according to its severity. The experiment is performed on different algorithms like Back Propagation Neural Network, Radial Basis Function and Random Forest. Dr. S. Vijayarani et al (March, 2015) proposed a research work that is used to predict kidney diseases by using Support Vector Machine (SVM) and Artificial Neural Network (ANN). The aim of this work is to compare the performance of these two algorithms on the basis of its accuracy and execution time. From the experimental results it is observed that the performance of the ANN is better than the other algorithm. Comparisons of Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms are done based on the performance factors classification accuracy and execution time. R.Sujatha et al (June 2016) taken data set used for the analysis is the chronic kidney disease from UCI data repository. As mentioned earlier the need of the hour work is to think about the health. The data set is rich with the vital things to be considered for classifying with nominal and numerical values for the attributes. The various attributes are age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anaemia and finally the class that tells about chronic kidney disease or not.

III. PROLEM DESCRIPTION

The researchers in the medical field identify and predict the diseases besides proffering effective care for patients with the aid of data mining techniques. One of the major techniques used in Chronic Kidney Disorder disease classification is the classification and clustering tasks. The data mining techniques have been utilized by a wide variety of works in the literature to diagnose kidney oriented diseases with the following dataset: datasets from UCI machine Learning Repository. Information associated with the disease, prevailing in the form of electronic attributes which taken from electronic dataset, images and more. Several data mining techniques were utilized by several researchers to present prediction and diagnosis approaches for Chronic Kidney Disorder diseases. The analysis of different data mining techniques that can be employed in automated Kidney Disorder prediction systems. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective Kidney disease diagnosis. But the previous classifiers are failed to produce the maximum level of accuracy and fast detection of disease. The existing classification algorithms are suffered from the need of large datasets for accurate diagnosis. The decision trees were used for diagnosis, but the problem is the selection of attributes for fast classification.

IV. METHODOLOGY

A. MACHINE LEARNING TECHNIQUES

Machine learning deals with the erection and study of systems that learns from data. A training dataset with its corresponding feature vectors and labels are fed into the machine learners. Prediction is made for the test dataset and expected class is determined (Figure 1). In CKD prediction, samples are mapped either CKD or NO CKD.

This experiment have done with the help of open source data mining tools in window environment using net beans software. In this experiment, we have used java code and libraries which are available in WEKA to implement the J48 classifier. MATLAB is used to generate the genetic algorithm implementation.

Machine Learning performance efficiency is evaluated with the metrics such as Precision, Recall and F-Measure. The total samples are divided into True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN).

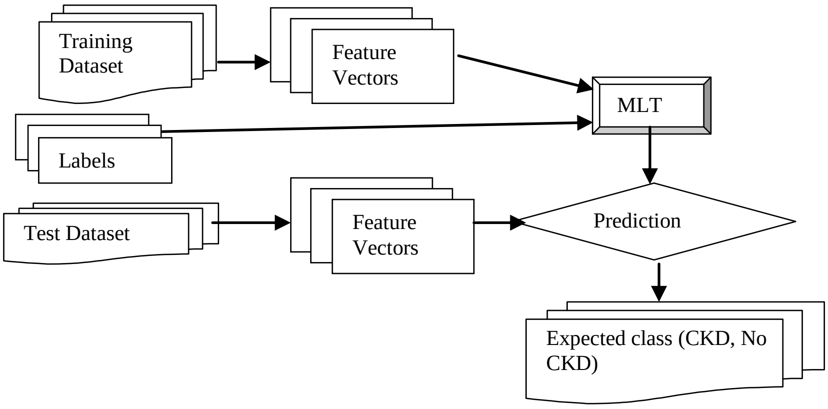


Figure 1: Machine Learning Techniques Working Scenario

Consider Positive (identified) and Negative (rejected), then

- True Positive: Number of correctly identified samples
- False Positive: Number of incorrectly identified samples
- True Negative: Number of correctly rejected samples
- False Negative: Number of incorrectly rejected samples

The evaluation metrics with their appropriate formulas are enlisted in Table 1. Experiments conducted in the work deploy these metrics.

Table 1: Evaluation Metrics

Confusion Matrix		Actual outcome (Observation)		
		Positive	Negative	
Test outcome (Expectation)	Positive	*True Positive* (TP)	*False Positive* (FP)	Positive Predictive Value (PPV) (or) Precision (α) = TP/(TP+FP)
	Negative	*False Negative* (FN)	*True Negative* (TN)	Negative Predictive Value (NPV) = TN/(TN+FN)
		Sensitivity or Recall (β) = TP/(TP+FN)	Specificity (or) True Negative Value (TNV) = TN/(TN+FP)	Accuracy (ACC) = (TP+TN)/(TP+FP+TN+FN)
				F-Score = 2.(α. β)/(α+ β)

B. Genetic Algorithm based Classification Tree

GA is used to evolve the decision trees for the closely related target concept disregard of the irrelevancy. CKD classification has been done with the GA, to evolve accurate and as well as simple decision trees. Complex decision tree consumes time and space complexity, which decreases the performance. The working scenario is portrayed in Table 2.The genetic algorithm is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution. The genetic algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution.

Table 2: Classical Algorithm vs. Genetic Algorithm

Classical Algorithm	Genetic Algorithm
• Generates a single point at each iteration. The sequence of points approaches an optimal solution.	• Generates a population of points at each iteration. The best point in the population approaches an optimal solution.
• Selects the next point in the sequence by a deterministic computation.	• Selects the next population by computation which uses random number generators.

The main aspects of the genetic algorithm are discussed as follows:

Fitness Functions

1. The *fitness function* is the function needs to optimize. For standard optimization algorithms, this is known as the objective function. The toolbox software tries to find the minimum of the fitness function.

2. Fitness functions as a file or anonymous function, and passes it as a function handle input argument to the main genetic algorithm function.

Individuals

3. An *individual* is any point to which can apply the fitness function. The value of the fitness function for an individual is its score. For example, if the fitness function is

$f(x1,x2,x3)=(2x1+1)2+(3x2+4)2+(x3-2)2,$

4. The vector (2, -3, 1), whose length is the number of variables in the problem, is an individual. The score of the individual (2, –3, 1) is *f*(2, –3, 1) = 51.

5. An individual is sometimes referred to as a *genome* and the vector entries of an individual as *genes*.

Populations and Generations

6. A *population* is an array of individuals. For example, if the size of the population is 100 and the number of variables in the fitness function is 3, represent the population by a 100-by-3 matrix. The same individual can appear more than once in the population. For example, the individual (2, -3, 1) can appear in more than one row of the array.

7. At each iteration, the genetic algorithm performs a series of computations on the current population to produce a new population. Each successive population is called a new *generation*.

Diversity

8. *Diversity* refers to the average distance between individuals in a population. A population has high diversity if the average distance is large; otherwise it has low diversity.

9. Diversity is essential to the genetic algorithm because it enables the algorithm to search a larger region of the space.

Fitness Values and Best Fitness Values

10. The *fitness value* of an individual is the value of the fitness function for that individual. Because the toolbox software finds the minimum of the fitness function, the *best* fitness value for a population is the smallest fitness value for any individual in the population.

11. Parents and Children

To create the next generation, the genetic algorithm selects certain individuals in the current population, called *parents*, and uses them to create individuals in the next generation, called *children*. Typically, the algorithm is more likely to select parents that have better fitness values.

V. RESULTS AND DISCUSSION

The following outline summarizes how the genetic algorithm works:
The algorithm begins by creating a random initial population. The algorithm then creates a sequence of new populations. At each step, the algorithm uses the individuals in the current generation to create the next population. To create the new population, the algorithm performs the following steps:

1. Scores each member of the current population by computing its fitness value.
2. Scales the raw fitness scores to convert them into a more usable range of values.
3. Selects members, called parents, based on their fitness.
4. Some of the individuals in the current population that have lower fitness are chosen as elite. These elite individuals are passed to the next population.
5. Produces children from the parents. Children are produced either by making random changes to a single parent—mutation—or by combining the vector entries of a pair of parents—crossover.
6. Replaces the current population with the children to form the next generation.

The algorithm stops when one of the stopping criteria is met.
Initial Population
The algorithm begins by creating a random initial population, as shown in the following figure 2.

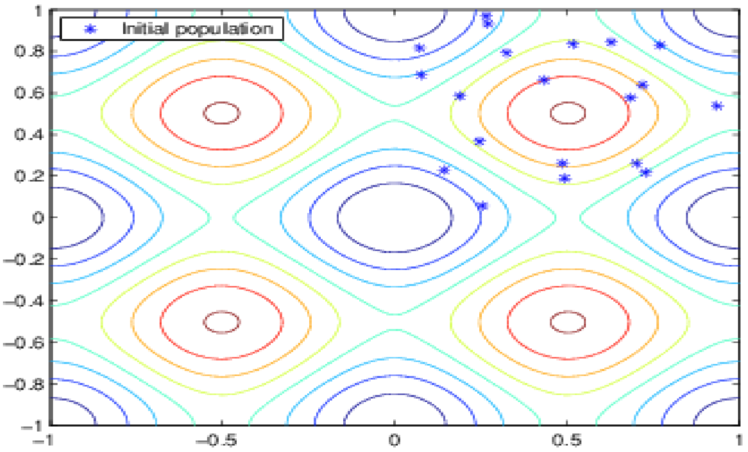


Figure 2: GA – Random initial population

This example, the initial population contains 20 individuals, which is the default value of Population size in the Population options. Note that all the individuals in the initial population lie in the upper-right quadrant of the picture, that is, their coordinates lie between 0 and 1, because the default value of Initial range in the Population options is [0;1].

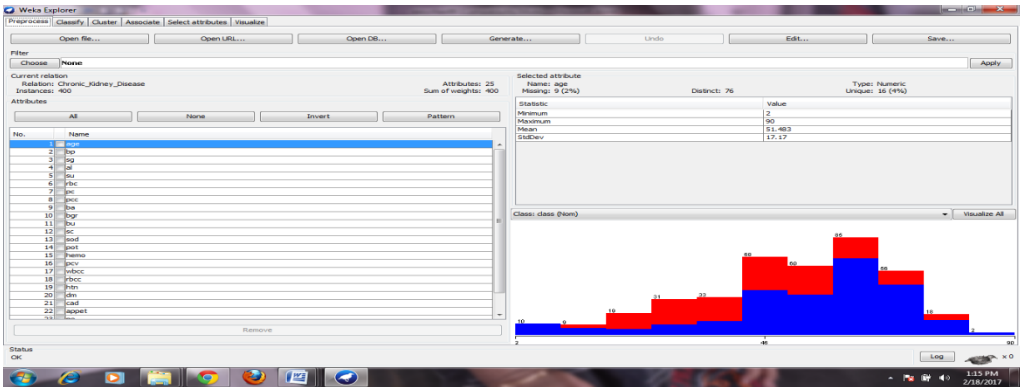
The J48 algorithm then recurs on the smaller sub-lists (Wikipedia 2013c). Figure 3 depicts the J48 based algorithm flow diagram for spamdexing. Standard 10 fold cross validation is used for the experiment and the tree produced by the J48 seems to have lower inference in which the assessment score is the only attribute considered. The remaining features and their purpose in the classification are not leveraged up to the mark.
Algorithm:

Step 1. Check for base cases
Step 2. For each attribute *a* , Find the normalized information gain from splitting on *a*
Step 3. Let *a_best* be the attribute with the highest normalized information gain
Step 4. Create a decision *node* that splits on *a_best*
Step 5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

The result of the J48 classifier is as follows:

```
J48 tree
-----------
Number of Leaves: 2
Size of the tree:    3
Time taken to build model: 0.27 seconds
```

Average Accuracy: **0.891**



Weka Explorer

Filter: None

Current relation
Relation: Chronic_Kidney_Disease
Instances: 400
Attributes: 25
Sum of weights: 400

Attributes

No.	Name
1	age
2	bp
3	sg
4	al
5	su
6	rbc
7	pc
8	pcc
9	ba
10	bgr
11	bu
12	sc
13	sod
14	pot
15	hemo
16	pcv
17	wbcc
18	rbcc
19	htn
20	dm
21	cad
22	appet

Selected attribute
Name: age
Missing: 9 (2%)
Distinct: 76
Type: Numeric
Unique: 16 (4%)

Statistic	Value
Minimum	2
Maximum	90
Mean	51.483
StdDev	17.17

Class: class (Nom)

Status
OK

Figure 3: CKD Dataset Visualization

Results

```
Time taken to build model: 0.05 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        396              99      %
Incorrectly Classified Instances        4               1      %
Kappa statistic                      0.9786
Mean absolute error                     0.0225
Root mean squared error                   0.0807
Relative absolute error               4.7995 %
Root relative squared error            16.6603 %
Coverage of cases (0.95 level)          99.75   %
Mean rel. region size (0.95 level)       52.25   %
Total Number of Instances              400
```

Table 3. Performance Comparison for J48 classifier

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.996	0.02	0.988	0.996	0.992	0.999	ckd
0.98	0.004	0.993	0.98	0.987	0.999	notckd
Weighted Avg. 0.99	0.014	0.99	0.99	0.99	0.999	

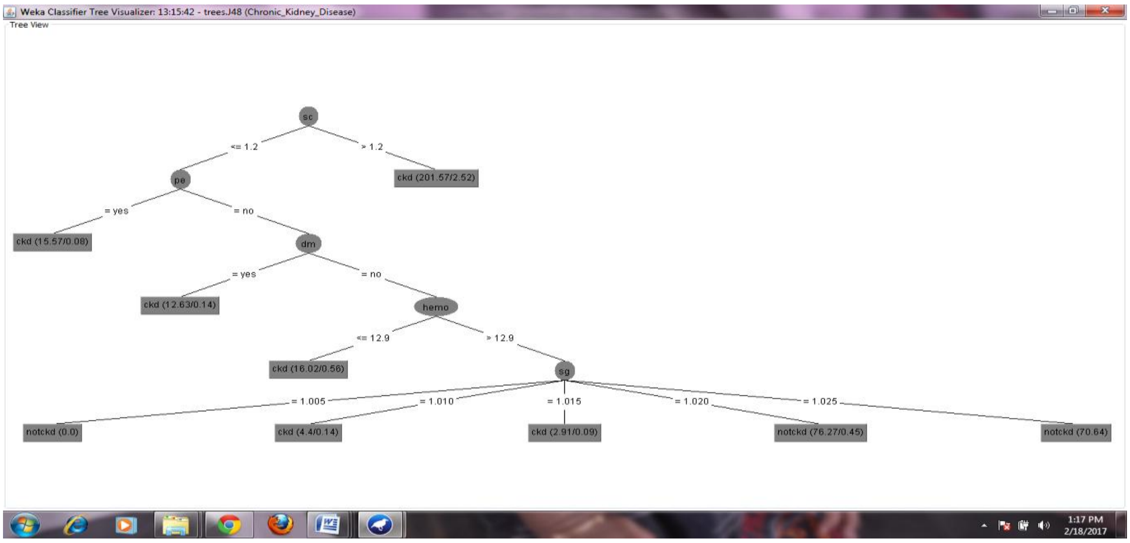


Figure 4: Results of the J48 Decision tree

J48 decision tree yield good result with less feature inference and GA based algorithm seems to create a better result, which considers many features and gets a clear inference deep through the data as in Figure 4. The average accuracy yielded from both classifiers show that GA is good when compared with J48 tree based algorithm. The maximum accuracy yielded through GA based classifier for initial population value 100 is 0.9375 and the least accuracy yielded is 0.6875. The score comparison in genetic algorithm in 100 iterations is given in Figure 5. Best and average score of the individuals is plotted along with the training and testing accuracy.

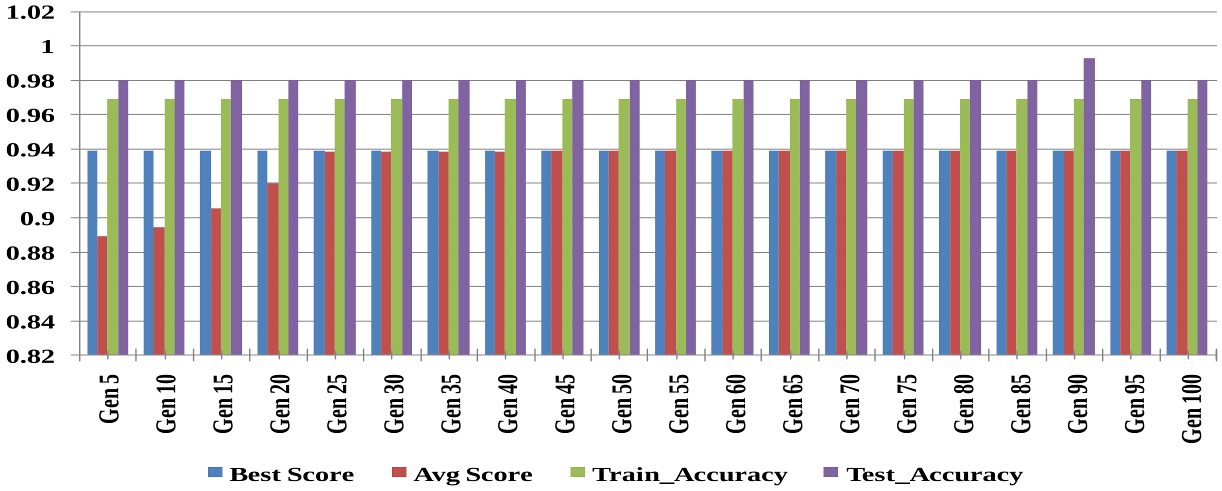


Figure 5: Performance Comparison of the GA in 100 Generations

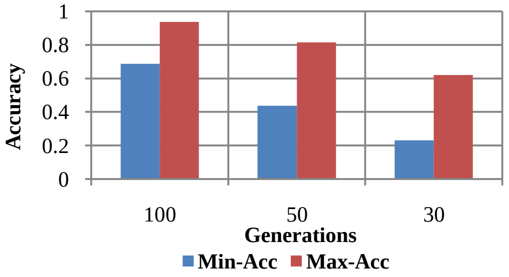


Figure 6: Minimum and Maximum Accuracy in Different Populations

GA decision tree with higher number of initial population tends to have good inference on features and accuracy on classification. The maximum and minimum accuracy obtained in three different populations are plotted in Figure 6. It is evident from figure that, if number of initial population is reduced the performance degrades. Experimental results show that GA based classifier seems to be a better discriminator with average accuracy of 0.912 for CKD classification and J48 classifier with average accuracy of 0.991. GA based decision tree gives optimal solution for this CKD classification with good feature inference on features. Prediction of chronic kidney disease is one of the essential topics in medical diagnosis.

The proposed work is to classify the different stages of chronic kidney disease according to its severity. The classification algorithms that have been considered for predicting chronic kidney disease are Genetic Algorithm and J48 decision tree. The models are evaluated with four different measures. From the experimental result, the genetic algorithm gives the better accuracy for predicting chronic kidney disease

VI. CONCLUSION AND FUTURE WORK

Addressing the CKD seems to be most important aspect as of witnessed from these reports. The proposed strategy and adopted classifiers seems to perform well for the CKD prediction. This strategy hikes and optimizes the overall performance. Among the experimented classifiers the Genetic Algorithm seems to vary apt for the problem. Hence, it is clearly evident that Genetic algorithm classifier seems to be optimal for the CKD prediction problem.

Machine learning deals with the erection and study of systems that learns from data. Future work is planned to build and improve upon the three contributions presented in this paper. Future plans include: collecting life event data via a cellphone interface for better data accuracy; expanding the size of the GMR variability dataset for improved classification; and investigating additional feature templates and modelling advances for better features to characterize the CKD prediction. There are also potential clinical applications of this work, including routine screening for excessive GMR variability and safety alarms to alert patients to predicted organ disfunction.

VII. REFERENCES

[1] A Manish Kumar, "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm" International Journal of Computer Science and Mobile Computing, Vol. 5, Issue. 2, February 2016

[2] S.Ramya, Dr. N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.

[3] Jyoti Saini, R.C Gangwar and Mohit Marwaha, "A NOVEL DETECTION FOR KIDNEY DISEASE USING IMPROVED SUPPORT VECTOR MACHINE" International Journal of Latest Trends in Engineering and Technology Vol.(7)Issue(4), 2015.

[4] Parul Sinha, Poonam Sinha, " Comparative Study of Choric Kidney Disease Prediction using KNN and SVM" , International Journal of Engineering Research and Technology, Vol(4), Issue-12, 2015.

[5] Neha Sharma, Er.Rohit Kumar Verma, " Prediction of Kidney Disease by using Data Mining Tecniques" International Journal of Advanced Research in Computer Science and software Engineering, Vol 6, Issue 9, September 2016.

[6] Dr. S. Vijiayarani, Mr. S. Dhayanand, " KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS " International Journal of Computing and Business Research(IJCBR), Volume 6, Issue2, March 2015.

[7] Sai Presad Potharaju, M.Sreedevi, " Ensembled Rule Based Classification Algorithms for predicting Imbalanced Kidney Disease Data" Journal of Engineering Science and Technology Review 9(5) (2016).

[8] Lambodar Jena, Narendra Ku. Kamila ," Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease " International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-11).

[9] L.Jerlin Rubini, Dr.P.Eswaran, "Generating comparative analysis of early stage prediction of Chronic Kidney Disease" International Journal Of Modern Engineering Research (IJMER) ISSN: 2249–6645 ,Vol. 5 ,Iss. 7 ,2015.

[10] Morteza Khavanin Zadeh , Mohammad Rezapour , Mohammad Mehdi Sepehri, "Data Mining Performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients" International Journal of Hospital Research, 2012.

[11] Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq, "Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease", International Journal of Database Management Systems (IJDMS), Vol.8, No.3, June 2016.

[12] Pushpa M. Patil, "Review on Prediction of Chronic Kidney Disease using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol. 5, ISSN 2320–088X, Issue. 5, May 2016.

[13] Mohammad Rezapour, Morteza Khavanin and Mohammed Mehdi Sepehri, "Implementation of Predictive Data Mining Techniques for Identifying Risk Factors of Early AVF Failure in Hemodialysis Patients" Computational and Mathematical Methods in Medicine, Volume 3, 2013.

[14] R.Sujatha, Dr.Ezhilmaran, "PERFORMANCE ANALYSIS OF DATA MINING CLASSIFICATION TECHNIQUES FOR CHRONIC KIDNEY DISEASE" International Journal Of Pharmacy & Technology, ISSN: 0975-766X, Vol-6, 2016.

AUTHOR'S BIOGRAPHY

Dr. S. Prema, currently working as Associate Professor in Departmnet of Computer Science,K.S.R. College of Arts & Science has received Ph.D., from the Bharathiar University in 2015. She is a member of ACM CSTA, IACSIT, TIFR-CORE, WSEAS and WASET. She has published more than 50 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE,Springer,Elsevir and it's also available online. Her research paper entitled "An NLP based Approach for Facilitating Efficient Web Search Results using BSDS" received the **best paper award**. She has h-index value: 5, i-10 index: 3, Citations: 85 and her profile is listed in Marquis Who is Who in World, International Biography Center, London, UK, 2011. Her main research work focuses on Web Mining,Web personalization,Information Retrieval and Visualization. She has 13 years of teaching experience and 8 years of Research Experience. She is guiding 4 M.Phil and 1 Ph.D Scholars for doing their research. She has been involved in generating funds for R&D.

Ms.J.Suguna is pursuing M.Phil (Computer Science) in K.S.R. College of Arts & Science (Autonomous),Tamilnadu,India.She has attented 4 workshops and 3 Seminar related to Data mining tools.Her area of interest are data/web mining, Big data Analytics and Machine Learning.