ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques

[1] Tabassum S, [2] Mamatha Bai B G, [3] Jharna Majumdar
[1] PG Student, [2] Asst. Professor,
[3] Dean R&D, Professor & Head, Principal Investigator of VGST Project
[1][2][3] Department of M. Tech CSE, Nitte Meenakshi Institute of Technology, Bengaluru, India

***Abstract* - Data Mining in Healthcare has become a present trend for obtaining accurate results of medical diagnosis, Chronic Kidney Disease (CKD) has become an international fitness problem and is a place of concern. It is a situation where kidneys turn out to be damaged and cannot filter toxic wastes within the frame. By using Data Mining Techniques, researchers have the scope to predict the Chronic Kidney Disease. This helps doctors to diagnose and suggest the treatment at an early stage. It also helps the patients to know about their health condition at an earlier stage and follow necessary diet and prescriptions.**

***Key Words*— Big Data, Data Mining Techniques, Expectation Maximization, Artificial Neural Network, C4.5 Algorithm.**

I. INTRODUCTION

Big Data [1] deals with the tremendous measure of information being handled through the Data Mining environment. Three Vs are the characteristics of Big Data as shown in the Fig 1:

Fig 1: Three V's of Big Data

- Volume - The volume referring to the tremendous measure of the information which is generated each second that are bigger than the standard social database.
- Variety – The Variety deals with the kind of attributes of the dataset like numerical, continuous, categorical, ratio and ordinal.
- Velocity – The Velocity represents the analysis of the algorithm for the computation of different type of attributes of process data.

Data Analytics as shown in the Fig 2 generally includes gathering information from various sources, manipulate it in a way that it winds up plainly accessible to be devoured by examiners lastly convey information items valuable to the association business. The way toward changing over a lot of unstructured crude information, recovered from various sources to an information item valuable for associations shapes the core of Big Data Analytics Big Data Healthcare [2] in Healthcare using Data Mining techniques is mainly useful in medical discipline where no availability is there for the proof of favouring a selected treatment alternative is located. A Huge volume of complicated information is created constituting patients data,

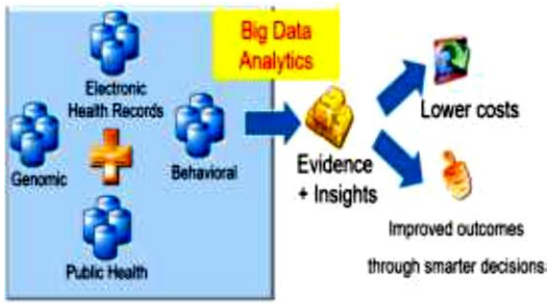

Fig 2: Big Data Analytics in Healthcare

disease record, hospitals bills, medical equipments, insurance claims, treatment price and so on. That requires evaluation and processing for extracting useful knowledge from this data. Data mining constitutes of several methodologies and algorithms which is performed on this

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

processed information. Recent advancements in healthcare are helpful for the specialists for making suitable selections and improving the quality of living of the diseased person. Patients with comparable health problems may be grouped collectively and effective remedy plans can be recommended based totally on patient's records, bodily examination, prognosis and former remedy styles.

II. LITERATURE SURVEY

Harshit Kumar et. al [3] ,discussed the overview of Big Data in Healthcare, the features of big data, and the stakeholders of the data is presented. The increase in digitization of information in healthcare industry has commenced producing information that fits in the definition of large data with the aid of all the attributes and definitions. The analytics of those digital statistics will offer multidimensional blessings in scientific practices, disorder surveillance, population health management and control in healthcare industry

Ms. AsthaAmeta et. al [4], it gives the solutions to diagnose the disease by analysing the data through different classification techniques. This helps to give quick and proper treatment to the patients. The major challenge is to find the best classification technique algorithm on the basis of type accuracy and execution time overall performance elements. The algorithm with better accuracy and minimal execution time is selected as the great algorithm. In this classification, different rate of accuracy charge is proven by means of every classifier. ANN has the maximum classification accuracy and ANN algorithm is considered as a better set of rules for classification technique.

Viktor Medvedev et. al [5], this paper depict, a best level perspective of procedures and innovation utilized for Big Data grouping is provided. The bunching is one of the fundamental insights digging inconveniences particularly for substantial data examination, where expansive volume records ought to be gathered. Enormous insights convey new requesting circumstances to information mining in light of the fact that colossal volumes and particular sorts should be considered. The regular methods and gear for records preparing and assessment are not ready to oversee such measures of actualities, despite the fact that effective PC groups are utilized. To break down enormous information, numerous new actualities mining and machine becoming more acquainted with calculations notwithstanding innovation had been produced. Thus, enormous records do no longer just yield new data sorts and carport components, however moreover new techniques of examination.

Dr. S. Vijayaram et. al [6] , the creator focused on expectation of four sorts of kidney ailments, Kidney disorders are foreseen the utilization of insights mining calculations which incorporates Artificial Neural Network (ANN) and Support Vector Machine (SVM). Order methods are utilized for the arrangements of four styles of kidney infections. Examinations of SVM algorithm and ANN algorithm calculations are executed construct absolutely with respect to the execution factors class precision and execution time. From the results, it could be inferred that the ANN accomplishes increased sort execution, yields outcomes which may be right, thus it is thought about as pleasant classifier while in examination with SVM classifier calculation. Possibly, SVM classifier orders the realities with negligible execution time. In predetermination, ANN set of guidelines is better than limit the execution time

J Chitra Devi et. al [7], the creator depicts that C4.5 grouping calculation manages numerical properties and additionally all out characteristics. Information mining is the broad zone wherein the dataset beneath tests are examined and designs are removed. Dataset is the social event of characteristics and its cost for various cases. Properties speak to the qualities of the insights and occasions are the estimations of the measurements. Each dataset has an ascribe that should be recognized. This trait is named on the grounds that the class name. Result can be a decision tree display, arrangements mining, realities class and so on. Different comprehensively utilized methods in information mining are Association rules mining, Classification, Clustering, Temporal mining, spatial mining and so on. Among the strategies, Classification assumes a basic part in information mining. Neural people group based class, Decision tree, Likelihood assessment, Naïve Bayes order are few class methods employed. In this paper, choice tree based absolutely arrangement is mulled over. A whole dataset is required in ID3 while C4.5 would paintings be able to with the dataset with lacking realities esteems. For decision tree generation, the C4.5 classifier erases the occurrences whose characteristic esteems are deficient. C4.Five classifier can handle each numeric and express

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

quality. In the event that particular quality has n unmistakable esteems, at that point the split on the characteristic will brings about n brilliant hubs. On the off chance that n is enormous; the choice tree will have high many-sided quality. In this way markdown in choice tree length is required. Truck calculation normally creates the parallel decision tree, yet makes utilization of the change on particular trademark to cut up the hub. This paper reasons that the paired decision tree classification construct absolutely with respect to C4.5 classifier is far cutting edge access to the general determination tree built.

III. METHODOLOGY

The clinical data consist of patient's records which are has been considered for the analysis and that dataset is taken from UCI Machine Learning Repository [8]. Totally 25 attributes are there in the dataset. Numerical and Nominal values of attributes are considered. The attributes of Chronic Kidney Disease are as shown in Table 1 and In the Proposed System, using Chronic Kidney Disease dataset and Data Mining techniques like Classification and Clustering, the related work will analyze and predict the Chronic Kidney Disease and its severity. Clustering algorithm like Expectation Maximization [EM] algorithm is used where the dataset is fed as input and the corresponding output is fed as input to the Classification Algorithms like Artificial Neural Network [ANN] and C4.5 Algorithm. The clustered output is analyzed by the Classification algorithms and respective results are obtained. Finally, considering all the results obtained, the accuracy check is done to analyze which algorithm is best suitable for prediction of CKD. The overview Methodology and flow of the Proposed System is as shown in the Fig 3

(i) Objective

The objective of the proposed work is to cluster similar kind of affected patient's records and to further classify the affected patient's records as per the variations and severity of the disease being affected.

(ii) Data Mining Techniques

Clustering is the method of grouping the information; this process is known as clusters. Grouping is accomplished by finding similar characteristics in the actual information. Thus clusters will be outlined as cluster of like parts.

Classification is done for the data mining that assigns to the objects in a group to the target categories or to the classes. Classification is the method of predicting the final results will be based on the given input dataset. The set of rules attempts to discover relationships among the attributes that might make it feasible to expect the outcome. The aim of classification technique is to appropriately be expecting the target class for every case within the records. Example: Whether the affected person is affected from Chronic Kidney Disease or not.

Table 1: Attributes of Chronic Kidney Disease

SL. NO.	ATTRIBUTES
1	Age
2	Blood pressure
3	Specific gravity
4	Albumin
5	Sugar
6	Red blood cells
7	Pus cell
8	Pus cell clumps
9	Bacteria
10	Blood glucose random
11	Blood urea
12	Serum creatinine
13	Sodium
14	Potassium
15	Hemoglobin
16	Packed cell volume
17	White blood cell count
18	Red blood cell count
19	Hypertension
20	Diabetes mellitus
21	Coronary artery disease
22	Appetite
23	Pedal edema
24	Anemia
25	Class

A. Expectation-Maximization [EM] Algorithm

EM algorithm is the type of clustering techniques which will be come under model based method which depend upon probability distribution. EM algorithm is type unsupervised learning, only with a set of inputs you are training your machine learning task; it is called unsupervised

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

learning. Expectation–Maximization (EM) algorithm is defined as an iterative method to find out the maximum probability or map posterior (MAP), which estimates the various Parameters in statistical models. Here the model depends on latent variables which are unobserved.

Algorithm: Expectation-Maximization [EM]

Input: Chronic Kidney Disease dataset as input.

Output: Form the 5 clusters.

The following steps are followed to form the clusters using Expectation Maximization algorithm:

Step1: Expectation (E): which creates a function for the expectation of the log-likelihood evaluated using the current
estimate for the parameters, and
X = E [log L (θ|Z)]…… 1
Where, Θ = initial guess for parameter, and Z = missing value
.
Step2: Maximization (M): This computes parameters maximizing the anticipated log-likelihood observed at the E
step.
$\theta^{t} \leftarrow$argmax X…… 2
Use the computed value to obtain better estimates for Θ.
Step3: Iterate E step and M step until converge

Fig 3: Methodology of Proposed system

B. Artificial Neural Network [ANN] Algorithm
Artificial Neural Network (ANN) might be an order procedure that is normally useful to determine information mining applications. A system that is regulated necessities the specific fancied yield for contribution as unsupervised systems needn't bother with the required yield for each information. Counterfeit neural system square measure scientific strategies square measure formed on the start of prevalent learning forms inside the human. The learning procedure is performed by evening out the net on the premise on the introduce of relations that exist between segments Network framed in this way is prepared for the obscure information and it will respond in view of gained learning.
The center work of simulated neural systems is expectation.
One in everything about all around preferred algorithmic program of neural system is back proliferation algorithmic program.

Algorithm: Artificial Neural Network [ANN]

Input: the output of EM algorithm as input to the ANN algorithm.

Output: calculate accuracy, precision and recall

- **True Positive:** The number of persons who are actually suffered from CKD among those who are diagnosed CKD.
- **True Negative:** The number of persons who are healthy among those who are diagnosed CKD.
- **Accuracy:** Accuracy refers to the ability of classifier. Number of correct predictions and total of all cases of predicted.
- **Precision and Recall:** Exactness or quality is a measure of precision, whereas completeness or quantity is a measure of recall

C. C4.5 Algorithm
C4.5 algorithm is a type of classification algorithm. C4.5 is an enhancement of ID3 algorithm. It handles both discrete values and continuous values. And it is used to generate the decision tree. Improvement of C4.5 algorithm made over ID3 algorithm

➢ Handling each continuous and distinct attributes threshold value has been created then it split the attributes according to the threshold value above and those value which are less than or the same to it.

➢ Handling the dataset with missing attribute values also – C4.5 allows for missing value and the attribute value is marked as?, the missing attribute values part unit is simply not utilized in the calculations of gain and entropy .

➢ Handling attributes at variance prices

Algorithm: C4.5

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

Input: The output of the EM algorithm as input to the C4.5 algorithm.

Output: C4.5 algorithm generates decision tree classifiers. Based on C4.5 algorithm data classification representation construct the decision. To calculate gain ratio the formula are given below.

$E(S) = -\sum p_i \log p_i$

Where, i - 1..... Count of class labels P_i – Probability of occurrence of class label in dataset.

$$I(S, A) = \sum_i^n \frac{|Si|}{|S|} E(S_i)$$

$$Splits(S, A) = -\sum_i^n \frac{|Si|}{|S|} \log\left(\frac{|Si|}{|S|}\right)$$

$$Gain(S, A) = E(S) - I(S, A)$$

$$Gain\ Ratio(S,A) = \frac{Gain(S,A)}{Split(S,A)}$$

- Choosing dataset as an input to the rule for process.
- Then calculate the Normalized information gain for each one attribute.
- And the attribute which has the highest information gain that attribute will be selected as best attribute.
- Then create a decision tree and attribute which has highest information gain make it as root node of that decision tree
- Repeat the process and calculate the information gain for each attribute and add that attribute as children node

IV. EXPERIMENTAL RESULTS

We have used Data Mining technique like clustering and classification such as Expectation Maximization algorithm, Artificial Neural Network and C4.5. By using these Data Mining Techniques which helps doctors to diagnose and suggest the treatment. It also helps the patients to know about their health condition at an earlier stage. The output of the screen is as shown in the Fig 4:

Fig 4: Output Screen

The EM algorithm gives the result as it form 5 clusters like LOW CKD, MODERATE CKD, HIGH CKD, SEVERE CKD AND NOT CKD the cluster result is shown in the different type of graph As shown below in Fig 5, Fig 6, Fig 7:

- LOW CKD consists of 100 records
- MODERATE CKD consists of 200 records
- HIGH CKD consists of 300 records
- SEVERE CKD consists of 900 records
- NON CKD consists of 500 records

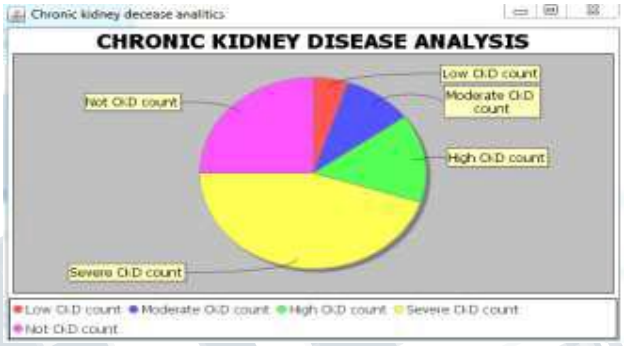

Fig 5: Bar Chart for the output of EM algorithm

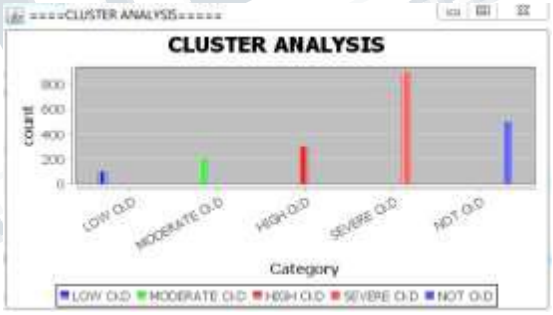

Fig 6: Pie Chart for the output of EM algorithm

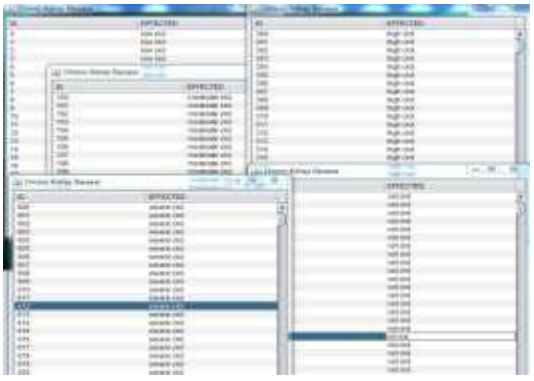

Fig 7: 5 cluster form

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

The result of ANN algorithm is shown in the below figures, Fig 8 show the outlook of ANN algorithm then and then enter the tuple number and submit the tuple number as shown in the Fig 9, Fig 10, then it classify the number of true positive and number of true negative as shown in the Fig 11 and calculate accuracy, precision, and recall with respect to the particular tuple which had been given and the result is as shown in the Fig 12:

Fig 8: outlook of ANN algorithm

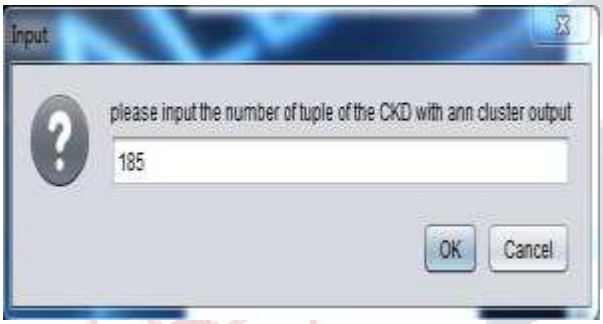

Fig 9: enter the tuple number

Fig 10: submit the tuple number

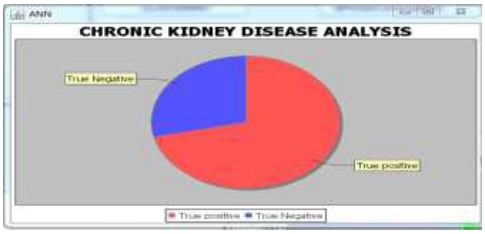

Fig 11: Output of ANN algorithm with respect to the classifiers

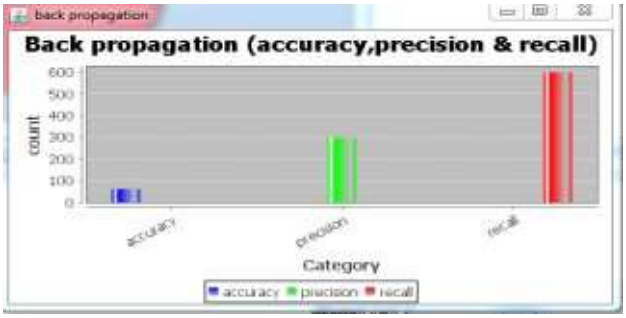

Fig 12: Output of ANN algorithm with respect to the accuracy, precision and recall

The result of C4.5 algorithm is shown in the Fig 13, it is represented in the form of decision tree structure.

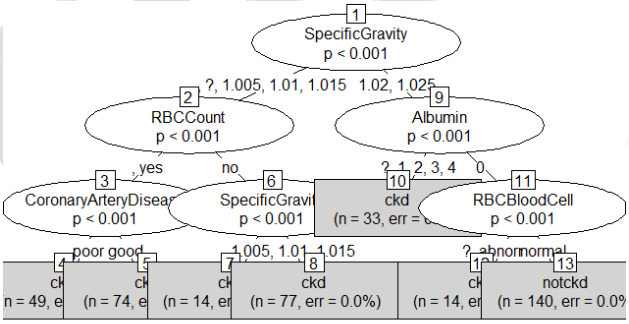

Fig 13: Decision Tree of C4.5 Algorithm

A. Accuracy Result

Accuracy was calculated for each data mining algorithm. It is observed that EM is a type of clustering algorithm; it got the accuracy result of 70%. Artificial Neural Network [ANN] and C4.5 are classification algorithm in which ANN got the accuracy result of 75% and C4.5 algorithm got the accuracy result of 96.75% . Comparison is made for the classification techniques; it indicates that the degree of accuracy of the C4.5 is more than ANN algorithm. Hence C4.5 is more accurate for prediction of Chronic Kidney Disease whether the patient is affected from disease or not.

Fig 14 show the accuracy comparison between three algorithms and accuracy result in also shown in the Table 2:

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

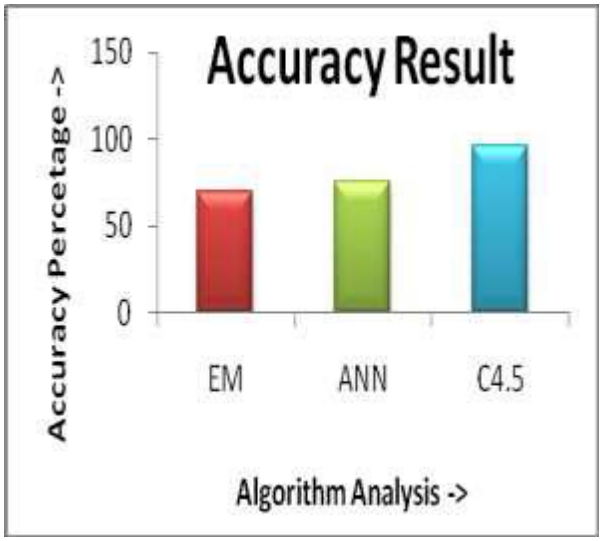

Fig 14: Accuracy Result of data mining techniques

Table 2: Accuracy result

ALGORITHM	ACCURACY
EM	70%
ANN	75%
C4.5	96.75

V. CONCLUSION AND FUTURE WORK

This research work is based on big data in healthcare which has been developed by using data mining techniques. For analysis and prediction of Chronic Kidney Disease (CKD) Data mining techniques has been used. Expectation Maximization [EM] is the clustering algorithm which is used to cluster similar type of person into one group. Artificial Neural Network [ANN] and C4.5 are classification algorithm which is used for prediction of the disease.
Accuracy was calculated for each data mining algorithm.
It is observed that EM is a type of clustering algorithm; it got the accuracy result of 70%. Artificial Neural Network [ANN] and C4.5 are classification algorithm in which ANN got the accuracy result of 75% and C4.5 algorithm got the accuracy result of 96.75% . Comparison is made for the classification techniques; it indicates that the degree of accuracy of the C4.5 is more than ANN algorithm. Hence C4.5 is more accurate for prediction of Chronic Kidney Disease whether the patient is affected from disease or not. In this study it shows that classification strategies are used for prediction of the disease whether the patient is affected from CKD or not. The clustering techniques are used to cluster the similar kind of affected person under one cluster. This technique allows the doctors to suggest the encouraged medicine and minimize the fee. The essential aim is to decrease the cost and offer better treatment. In this project consists of only one clustering algorithm and two classification algorithm. In future work we can implement different clustering and classification algorithm for different healthcare dataset and calculating the accuracy between different algorithms and find out which algorithm is more efficient.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Prof N.R Shetty, Advisor and Dr. H.C Nagaraj, Principal, Nitte Meenakshi Institute of Technology for giving constant encouragement and support to carry out research at NMIT. The authors extend their thanks to Vision Group on Science and Technology (VGST), Government of Karnataka to acknowledge our research and providing financial support to setup the infrastructure required to carry out the research.

REFERENCES

[1] T.Sanjay , C.M Sheela Rani and K.V Narayana, "A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology, January 2016.

[2] VeenitaKunwar and Khushboo Chandel,"Chronic Kidney Disease Analysis Using Data Mining Classification Techniques", 6th International Conference - Cloud System and Big Data Engineering (Confluence), 2016

[3]Harshit Kumar, Nishant Singh," Review paper on Big Data in healthcare informatics", International Research Journal of Engineering and Technology, Feb -2017

[4] Ms. AsthaAmeta,Ms. Kalpana Jain, "Data Mining Techniques for the Prediction of Kidney Diseases and Treatment: A Review", International Journal Of Engineering And Computer Science, Feb. 2017

[5] Viktor Medvedev, "Strategies for Big Data Clustering" IEEE 26th International Conference on Tools with Artificial Intelligence, 2014.

[6] Dr. S. Vijayaran, Mr.S.Dhayanand, "Kidney Disease prediction using SVM and ANN algorithms", International Journal of Computing and Business Research, March 2015.

[7] J Chitra Devi, "Binary Decision Tree Classification based onC4.5 and KNN Algorithm for Banking Application", International Journal of Computational Intelligence and Informatics, September 2014

[8]http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Di sease.