Research of Chronic Kidney Disease based on Data Mining Techniques

M. Thiyagaraj, G. Suseendran

Abstract--- Kidney disease is one of the real general medical issues these days. Ceaseless ailments prompt to horribleness and mortality in India and furthermore in the low pay and center nation. The interminable infections on record is 60% of death all through the around the world. 80% of unending malady passing overall additionally happen in low and center pay nations. In India, most likely the quantity of passing is because of the ceaseless ailment observed to be 5.21 million in 2008 and is by all accounts brought to 7.63 million up in 2020 roughly 66.7%. Information mining is the procedure of extraction is the concealed data from the given expansive dataset. Different information mining strategies, for example, bunching, characterization, affiliation investigation, relapse, outline, time arrangement examination and succession investigation were utilized to anticipate kidney maladies. The strategies that were presented so far had minor downsides in the nature of pre handling or at some other stages. In this paper, the different information mining methods are reviewed to foresee kidney sicknesses and real issues are quickly clarified.

Keywords--- Chronic Kidney Disease (CKD), Risk Factors of CKD, Challenging Issues of CKD and Data Mining and Machine Learning (ML) Algorithms.

I. INTRODUCTION

The tremendous progress of the measure of organic information accessible has brought up a terrible issue of being characterized, overseen adequately and to be changed from crude information to important data. The rise of this huge measure of raises doubt about the ideal models of advanced calculation. It looks for an answer towards receiving significant outcomes in return, keeping a particular thinking for fundamental calculations. Machine Learning clearly stands to catch a noteworthy part of the issue and in this manner represents the most recent advance in the field of bioinformatics, computational science and use of machine learning techniques on noticeable issues in human science and conduct. The calculations and numerical strategies permit us to go past an insignificant delineation of the information and make offers legitimate outcomes as scientifically testable models. The thoughts of regulated and unsupervised learning make this procedure simple and understandable. By streamlining deliberation that foundations a model, we can get factual forecasts of a framework.

As people's day to day life become more and more modernized and extended life span in the society, Chronic Kidney Disease (CKD) also found common and result in degradation in the functionalities of kidney function. Once any person gets CKD, they may suffer from the disease which may decrease their working capability as well as living quality. It is also rapidly results in other chronic diseases such as high blood pressure, anemia, and the weak bones due to the poor nutritional health and become nerve damage. In the meantime, the kidney disease is maximises in the patient risk of contracting heart and blood oriented diseases. Chronic kidney disease even causes other chronic disease such as diabetes, high blood pressure and other disorders. High risk groups are classified as person with diabetes, hypertension, and hereditary. It is possible to get rid of chronic kidney disease through early detection and proper treatment once the progress of the disease is observed it may greatly leads to kidney failure.

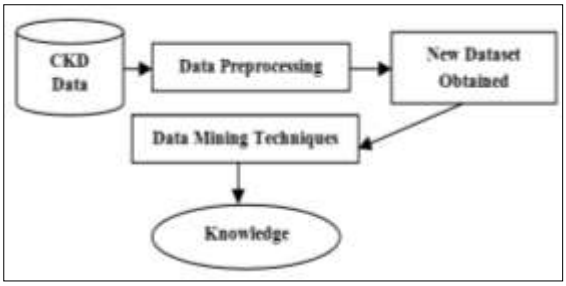

Figure 1: Overview of Research Work

Figure 1 above specifies the overview of the research work. CKD dataset is taken from the UCI Repository. Data preprocessing is performed to fill the missing values, noise removal, data cleaning, etc. At last, a new dataset will be obtained. Data Mining techniques are applied on it to evaluate the performance in case of predicting whether the person has kidney diseases or not. Kidney disease is a growing problem. Kidney damage occurs slowly over many years this is often due to the diabetes or high blood pressure is also called chronic kidney disease. The sudden change in the kidney function because of illness, injury and a certain medications this is called acute kidney injury. It can cause a person with normal kidneys or already exists for someone his problems (kidney Basics, 2015). The factors for developing kidney disease are: Diabetes, High blood pressure, cardiovascular disease, and family history of kidney failure (Pavithra,N and R.Shanmugavadivu, 2012)

Chronic kidney diseases which occur 60% of deaths in the world. 80% of chronic disease deaths occur in worldwide range of low and middle income countries (Ilangovan Veerappan, Georgi Abraham, 2012). The National Kidney foundation determines the different stages of chronic kidney disease is on the kidney damage and glomerular filtration rate (GFR), measure at the level of kidney function. Five stages in the chronic kidney disease. The health care dataset contains missing values.

Manuscript received September 16, 2019.

M. Thiyagaraj, Ph.D., Research Scholar, Department of Information and Technology, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai. T.N, India. (e-mail: mthyagaraj09@gmail.com)

G. Suseendran, Assistant Professor, Department of Information and Technology, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai. T.N, India. (e-mail:suseendar_1234@yahoo.co.in)

To address this problem, pre processing techniques will be used in healthcare datasets. These missing values can degrade the performance of abnormality detection. Several methods were proposed to fill up these missing values. An existing classification framework used a data preprocessing method where data cleaning method was used to fill in the missing values and to correct the erroneous ones. A recalculation process is performed on the chronic Kidney disease (CKD) stages and the values were recalculated and filled in for unknown values. Though this method is efficient, the influence of expert in the field of healthcare dataset values for CKD is needed. So to avoid this need and improve the preprocessing as a layman, an ensemble-learning based sparse-data modeling framework is proposed. The dataset transformation is performed to convert the nominal features to numerical features of the data set to make it support the generation of subsets. That is positive responses, such as Yes, Good, Present, are defined as 1, negative responses such as No, Poor and not present, are defined as 0. The proposed work generates subsets of the data, and then it trains models using subsets and then combines the candidate models into an ensemble learner for making predictions and also to generate a score for the importance of missing values in each feature in the dataset. Figure 2 illustrates the classification of kidney disease. It is of two type namely acute kidney disease and chronic kidney disease. Here in our research work we concentrate on factors supporting chronic kidney diseases respectively.

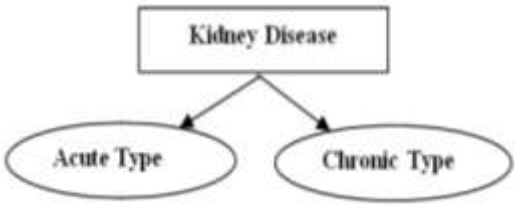

Figure 2: Classification of Kidney Disease

II. LITERATURE REVIEW

To measure the different methods of CKD the influence factors are used as the inputs for network modeling. Next, the detection of CKD classification and risk evaluation are evaluated. The three different neural network models are employed for detection. Kidney disease influences the kidneys damage and fails to filter blood. The damage cause wastes to buildup in our body. Kidney damage occurs slowly over years is due to the diabetes or high blood pressure. Anyone feels a sudden changes in kidney function due to illness, injury, has taken certain medications is called acute kidney injury. This can occur in a person with normal kidneys or in anyone who already suffers from kidney problems. It seems to be a growing problem. The 10% of population in worldwide were affected by this kidney disease.

Due to CKD millions of people die because no proper treatment. Accordingly 2010 Global Burden of Disease study, chronic kidney disease ranked 27th position in the list of causing a total number of deaths in worldwide in the year of 1990, but increases to 18th in the year 2010. Risk factors in the kidney disease are smoking, diabetes (types I and II),high cholesterol, obesity, autoimmune disease, obstructive kidney disease, bladder obstruction caused by Benign Prostatic Hyperplasia, atherosclerosis, cirrhosis and liver failure, narrowing of the artery, kidney cancer, bladder cancer, kidney stones, kidney infection (Kirubha and S.Manju Priya, 2016). Kidney diseases are predicted and compared by the help of Support Vector Machine and Artificial Neural Network algorithm based on the accuracy and execution time. Result shows that ANN outperforms with reduced execution time (Vijayarani, S and Dhayanand S, 2015). CKD dataset is taken from UCI repository and performance are evaluated using the algorithms such as Naive Bayes, Multilayer Perceptron, SVM, J48, Conjunctive rule and Decision tree. It shows that multilayer perceptron algorithm gives better classification accuracy rate in prediction of chronic kidney diseases (Lamboder Jena and Narendra Ku. Kamila, 2015). K-Means Clustering Algorithm with a single mean vector of centroids have been formulated to classify the clusters of varying probability of likeliness suffers from CKD. The results are obtained from a real case dataset (UCI Repository) to show the probability of disease causing factors (Abhinandan Dubey, 2015). Machine learning algorithms like AD Trees, J48, KStar, Naive Bayes, Random Forest algorithms are used to predict the kidney disease. The performance result of the Naive Bayes shows better accuracy rate compared to other algorithms (Swathi Baby, P and Panduranga Vital, 2015).

Evolution of big data in healthcare field is evaluated which support Vector Machine, Decision Tree and Bayesian Network machine learning algorithms (Basma Boukenze et al, 2016). Chronic Kidney Disease dataset is used to predict patients with chronic kidney failure and normal person. C4.5 algorithm provides better results with less execution time and accuracy rate.

Performances are judged by Basic concepts of Decision Tree, Bayes classification, Rule based classification, Back Propagation, Support Vector Machine and K-Nearest Neighbour algorithms.

To address the same problem, classifiers namely Multilayer Perceptron, Random forest, Naive Bayes, SVM, K-Nearest Neighbour and Radial Basis Function are also involved (Pushpa M. Patil, 2016). Three different neural network models have been implemented for chronic kidney disease prediction which includes back propagation neural network, generalized feed forward neural network and modular neural network. This research shows that all these models influences genetic algorithm in to their respective neural factor. All three models give better accuracy more than 85%. Compared to other models, back propagation neural network has the highest accuracy (Ruey Key Chiu and Renee Yu-Jing, 2011).

Chronic kidney disease are predicted using six different data mining algorithms like Random Forest Classifiers, Sequential Minimal Optimization, Naive Bayes, Radial Basis Function, Multilayer perceptron classifier and Simple Logistic. Totally, 400 records are used for training set to perform prediction. Among these, Random Forest outperforms well (Manish Kumar, 2016).

III. RISK FACTORS OF CKD

- **Susceptibility factors:** It results in increase susceptibility to kidney damage. These symptoms are seen in aged people, hereditary facts; reduce kidney mass, less weight baby, low income and educational levels.

The following table 1 shows the stages of CKD with GFR rate and action plan that must be taken eventually at each stage.

Table 1: Classification of Chronic Kidney Disease and Evaluation Plans (Ruey Kei Chiu, Fu Jen, 2011)

S.No	Stages of CKD	Glomerular Filtration Rate (GFR)	Action Plans
1	Kidney damage with normal GFR	90 or above	Diagnosis and treatment of comorbid conditions, disease progression, reduction in risk factors for a cardiovascular disease
2	Kidney damage with mild decrease	60 to 89	Estimation in disease progression
3	Moderate decrease	30 to 59	The Evaluation and treatment of disease complications
4	Severe reduction	15 to 29	Preparation in kidney replacement treatment (dialysis and transplantation)
5	Kidney failures	Less than 15	Kidney replacement therapy

- **Initiation Factors:** It specifies the factors directly involves in the kidney damage. The symptoms were diabetes high blood pressure, autoimmune diseases, mellitus, systemic infections, urinary stones, obstruction of lower urinary tract, urinary tract infections and drug toxicity.
- **Progression Factors:** It leads to worse fact of kidney damage and rapid decline functionalities once the damage gets started. The symptoms are found to be high level proteinuria, high BP, less glycemic control due to diabetes and smoking.
- **End stage Factors:** Increasing in the morbidity and mortality due to kidney failure results in lower dialysis dose, anemia, temporary vascular access, less serum albumin level and the late referral dialysis.

IV. CHALLENGES IN CHRONIC KIDNEY DISEASE MANAGEMENT IN INDIA

The average in global occurrence value is treating the End Stage Renal Disease (ESRD), dialysis and transplant patients is found to be 280, 215 and 65 patients per million (ppm) respectively. In India, the average occurrence values for the treated ESRD (not diagnosed), dialysis and transplant patients were 70, 60 and 10 ppm. Global increasing in number is 7% in every year. It is estimated that only 10 to 20% of ESRD patients in India undergoes the long term renal replacement therapy. It was estimated that in India 1 year, 3,500 new renal transplant, 3,000 new Continuous Ambulatory Peritoneal Dialysis initiation and 15,000 new Maintenance Hemo Dialysis patients (Agarwal, S.K and Srivastava, R.K, 2009).

V. DATA MINING TASK PRIMITIVES

The mining assignment was determined in inquiry used to include the information mining framework. An information mining question suggests the information mining assignment primitives. The client intuitively speak with the information mining framework to produce the mining procedure, or to inspect the discoveries from various edges or profundities. The predefined information can be mined utilizing the database or the arrangement of information the client is intrigued. It incorporates the database characteristics or in information distribution center measurements of intrigue. Learning can be mined utilizing portrayal, separation, affiliation or connection examination, order, expectation, bunching, exception investigation, or advancement examination. The learning is utilized to revelation prepare for assessing the examples was found. Chains of importance idea were turned out to be well known type of foundation information that permit mined information in a numerous levels of deliberation. This intriguing quality measures and limits for example assessment manage the mining procedure to assess the found examples. Learning may have distinctive intriguing quality measures like support and certainty for affiliation rules.

VI. ADVANTAGES AND DISADVANTAGES OF DATA MINING IN MEDICAL FIELD

Advantages

- Data mining in healthcare detects fraud and abuse.
- Help physicians to identify effective treatments and best practices.
- Patients exploit better and greater affordable healthcare services.
- Increases in the speed of working with large datasets and rapid report generation, faster analysis, improved operational efficiency and reduced operating cost.
- Data Mining can extract predictive knowledge from large databases.

Disadvantages

- Heterogeneous medical complex physician's interpretation with poor mathematical classification.
- Ethical, Legal and Social Issues.
- Data Ownership issues.
- Privacy and Security related to Human Data Administration.
- It Involves privacy issues and security issues and
- Misuse or incorrect information.

VII. DATA MINING ALGORITHMS AND TECHNIQUES

A data mining algorithm was well defined procedure to takes data input and generates the output. It involves patterns in form of model.

It encompasses several algorithms and techniques like Classification, Clustering, Prediction, Association Rules, Neural Networks etc., to perform knowledge discovery from databases.

a. Classification

Classification is the simplest and one of the popular data mining techniques. Where objects are divided and assigned to the other groups called classes. Each object has to be dispersed exactly to one class and not more than one and never to no classes at all. The classification algorithms are,

Naive Bayes Algorithm

The Naive Bayes algorithm combines prior probability and conditional probabilities for a single formula, To calculate the probability of each and possible classifications to turn. The classification with the largest value from a given a set k mutually exclusive and exhaustive classifications with prior probabilities and n attribute followed by values of instance. Posterior probability of class is occur to the specified instance was shown to proportional to prior probabilities along with respective values. The assumption that the attributes were independent, to the value of expression is calculated using the product by calculating the product for each value from 1 to k the classification to largest value is selected.

Nearest Neighbor

It is the kind of distance based algorithm used mainly that all the attribute values is continuous, it can be modified according to categorical attributes. To estimate the classification unnoticed instance by using the classification of the instance or instances that are closest. Even more instances the training set applies the same principle to classify the k nearest neighbours or most nearest one known as k- Nearest Neighbour.

Decision Tree based Algorithm

To solve the classification problems where the tree is constructed model to the classification process. As soon as tree is built, it is applied to each tuple database to generate classification.

b. Clustering

It examines the data to find a groups of items that similar to each other as partition-based, hierarchical and agglomerative methods.

c. Partition-based method

This method examines the overall grouping of the input database objects into k partitions as cluster. Clusters are created by optimizing a parameter to calculate cluster distance based on similarity measure. It enables similar objects placed within a cluster and unrelated objects are grouped under different clusters. It involves two types of clustering namely,

- **k-means clustering:** This is a centroid-based approach that takes the number of partitions k as input and creates k clusters of the input database consist of n objects or records by optimizing the rule of clustering. The resulting clusters with high similarity kept as intra-cluster and low similarity as inter-cluster. In this algorithm primarily a random set of k objects are chosen as the cluster centers to compute the mean value representing the cluster mean or centre. Each object can be involved in the cluster to exhibits high similarity. Updation of new cluster is made till there is no change found in the cluster structure.
- **k-medoids clustering:** It eliminates the sensitivity using medoid as a measure for similarity by choosing the mean value of objects in a cluster, the most centrally located object within a cluster is chosen. This position may be changed while maintaining the distance-based similarity measure. Randomly k objects are chosen as medoid and remaining objects are assigned to the cluster based on the medoid distance and higher iterations reveal the cluster mediods.

d. Hierarchical Clustering Algorithm

Hierarchical clustering methods focus on the principle of decomposing databases either in a top-down or bottom-up fashion. It is divided into divisive and agglomerative techniques.

- **Agglomerative clustering:** Each individual object is treated as a cluster. Then, clusters are merged according to the similarity of objects until a single cluster is formed. Finest case of clustering is made when all objects in the database influences same cluster type or when the terminate condition is reached.
- **Divisive:** It operates in the reverse direction or bottom-up and considers the whole database as a single cluster. Clusters splits based on similarity and dissimilarity measures, until each object formed as individual cluster or a terminating condition is reached (Margaret H. Dunham, 2006).

e. Prediction

This technique predicts or guesses the data values for attributes as missing or dislocate.

- **Linear Regression-Based Prediction**

Linear regression is the simplest among the different types of regression where data are modeled by straight lines. Data being modeled by straight lines could be interpreted as the growth of the data being in a linear form or following a straight line. Regression coefficients are estimated using the least squares method.

f. Association Rules

This rule is basically in expression form $X \rightarrow Y$ where X and Y are item-sets. Association rule mining is controlled to support and Confidence. Support gains statistical signification of a rule and confidence forms the degree of certainty for detecting the associations.

g. Apriori Algorithm

Apriori is a type of candidate generation algorithm proceeds in a level-wise order. The apriori algorithm follows the join and prune steps.

Join step constructs new candidate sets and prune step helps in filtering out candidate item-sets based on the anti-monotonic property.

h. Neural Networks

It consists of interconnected group in the artificial neurons to processes information is computation by using associated weights. These weights are updated or adjusted during prediction of input records. In ANN the adaptive system are changes the structure based on external or internal information that flows through the network in the learning phase. Modern neural network posse's non-linear statistical data modeling tool were used to model complex relationships between inputs and outputs. The most prestigious model used any application models to facilitate artificial intelligence**.**

i. Propagation

The approach used for processing is called propagation. Tuple of values are given as input to the neural network, in every node in input layer. Then the summation and activation functions are applied for each node, output value is created in each of output arc the node in turn sent to the subsequent nodes until a tuple of output values are reached from the nodes of the output layer.

j. Back Propagation

The learning technique is adjusts to weights in the neural networks by propagating the weight changes backward from sink to the source nodes. Back propagation is purely a generalized delta rule approach known as a feed-forward back propagation network. It constitutes a supervising learning Multi Layer Perceptron model. A training set of input patterns was applied to the networks to compute the output pattern, any error occurs, the difference between actual and desired output patterns is taken by adjusting its weight. Since the real uniqueness or effectiveness of the network exists in the values of the weights between neurons, it is necessary to adjust the weights among them.

k. Radial Basis Function Network

A radial basis function type of functions was the values decreases or increases from the central point at proper the distance. The Gaussian activation functions in RBF network typically a neural network with three layers. The input layer is used to input the data, gaussian activation function is specified at the hidden layer and linear activation function is involved in output layer.

l. Perceptrons

Simplest neural network called perceptron. It is single neuron with multiple inputs, one output developed to promote activation. Perceptron can be classified into two classes in unipolar activation function, where output 1 can be used to classify into one class, while an output 0 can be used to pass in the other class. The simplest feed forward can neural network is called a multilayer perceptron that forms the network of perceptrons.

VIII. DISCUSSIONS & RESULTS

Data Mining gains greater importance in extracting hidden information from large set of database. Though technique is intelligently extracted hidden information data mining techniques is not seems to be straight forward. To apply the data mining techniques it is significant to understand the nature of data. The most important step involves preprocessing of data, if the pre-processing is not carried out correctly then the entire decision making process may go worse. It involves data cleaning, data Integration and transformation of data into understandable format and reduction in size to improve the quality of pattern found.

We have discussed about the different data mining techniques tasks that is used for the diagnosis the diseases and among these, classification the most frequently used algorithm as it has many advantages while comparing to other algorithm to yield better classification. The advantages and disadvantages of data mining techniques have been discussed above. Future research paper addresses the issue evolved by keeping preprocessing as a layman, an ensemble-learning based sparse-data modeling framework is proposed. The dataset transformation is performed to convert the nominal features to numerical features of the data set to make it support the generation of subsets. That is positive responses, such as yes, good, and present, were defined 1, negative responses such as no, poor and not present, defined as 0. The proposed work generates subsets of the data, and then it trains models using subsets and then combines the candidate models into an ensemble learner for making predictions and also to generate a score for the importance of missing values in each feature in the dataset.

With respect to estimation of genuine levels in CKD, by and by glomerular filtration rate is the most normally measuring variable utilized as a part of the medicinal services field to gauge kidney work. The doctor in medicinal services field is figure GFR from patient's blood creatinine, age, race, sex and different components relying on the kind of formal-perceived calculation equations is utilized. The GFR demonstrate how soundness of a patient's kidney with respect to the ailment issue. The doctor checks these measures decide the distinctive stage in kidney illness of a patient (Firman, 2009 and NKF, 2002).

IX. CONCLUSION

This paper aims to analyze the various data mining techniques in medical domain and some of the algorithms used to predict kidney diseases eventually. From the above survey, it is proven that results may vary for different stages of kidney disease diagnosis based on the tools and techniques used. Data mining provides better results in disease diagnosis when appropriate techniques used. Thus, data mining is the significant field for healthcare predictions.

REFERENCES

1. Abhinandan Dubey, A Classification of CKD Cases Using MultiVariate K-Means Clustering. International Journal of Scientific and Research Publications, 5(8), pp.1-5, 2015.
2. Agarwal, S.K and Srivastava, R.K, Chronic kidney disease in India: challenges and solutions. Nephron Clin Pract, 111(3), pp.197-203, 2009.

4. 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, 8(3), 2016.
5. Firman G, Definition and Stages of Chronic Kidney Disease (CKD), Released on Feb. 28, 2009.
6. Kidney Disease Basics | National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).
7. Kirubha, V, and S. Manju Priya, Survey on Data Mining Algorithms in Disease Prediction International Journal of Computer Trends and Technology, 38(3), pp.124-128, 2016.
8. Lambodar Jena and Narendra Ku.Kamila, Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease. International Journal of Emerging Research in Management &Technology, 4(11), pp.110-118, 2015.
9. Levey, A.S, Greene, T, Kusek, J.W, Beck, G.L, A simplified equation to predict glomerular filtration rate from serum creatinine. MDRD Study Group, J Am Soc Nephrol, 2000.
10. Manish Kumar, Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm. International Journal of Computer Science and Mobile Computing, 5(2), pp. 24-33, 2016.
11. Margaret H.Dunham, S.Sridhar, Data Mining Introductroy and Advanced Topics published by Dorling Kindersley (India)Pvt.Ltd., Pearson Education India, 2006.
12. National Kidney Foundation (NKF), Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification and Stratification. American Journal of Kidney Disease, 39, S1-S266, 2002.
13. Pavithra N. and R. Shanmugavadivu, Survey on Data mining Techniques used in Kidney related Diseases. International Journal of Modern Computer Science, 4(4), pp. 178-182, 2016.
14. Ruey Kei Chiu and Renee Yu-Jing, Constructing Models for Chronic Kidney Disease Detection and Risk Estimation. Proceedings of 22nd IEEE International Symposium on Intelligent Control, Singapore, pp. 166-171, 2011.
15. Swathi baby, P and T. Panduranga Vital, Statistical Analysis and Predicting Kidney Disease Using Machine Learning Algorithms. International Journal of Engineering Research and Technology, 4(07), pp. 206-210, 2015.
16. Veerappan, Ilangovan and Abraham Georgi, Chronic Kidney Disease: Current Status, Challenges and Management In India. Indian Journal of Public Health Research and Development, 6(1), pp. 1694-1702, 2014.
17. Vijayarani, S and S.Dhayanand, Kidney disease Prediction Using SVM and ANN Algorithms, International Journal of Computing and Business Research, 6(2), 2015.