

Predict Chronic Kidney Disease using Data Mining Algorithms in WEKA Toolkit

M.Sivakumar¹, R.Shruthi², E.Vaibhava³ and R.Sangeetha⁴

¹AP/CSE

²⁻⁴UG student/CSE

dharsinishru@gmail.com

Abstract—This paper explains the chronic kidney disease prediction with the theory of various model evaluation mechanisms in data mining. From the given dataset, various statistical measures such as Accuracy, Error, Precision, Recall, F1 error, Elapsed Time used to analyze performances are described here. All of the above-mentioned parameters are compared in the tabular form for both the data mining classifiers, KNN and SVM are generated by using MATLAB itself. It is a two-step process, in first step KNN algorithm uses training dataset to build a classifier, and then in second step process, it uses this classifier to predict the class label of unlabeled data instance to perform the prediction of chronic kidney disease by accessing Hadoop in itself.

Index Terms— Data Mining, KNN, Weka toolkit, Chronic Kidney Disease.

I. INTRODUCTION

In today's modern world malady is the most lethal one. So diagnosing patients correctly on timely basis is the most challenging task for the medical fraternity. Unfortunately most of the systems rarely use the huge clinical data where vital information is hidden. Data mining is used for extracting information from huge sets of data. It is used to predict missing or unavailable numerical data values rather than class labels. Prediction can also be used for identification of distribution trends based on available data. Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. Data mining helps in identifying customer patterns and trends that lead to improved quality of service. Data mining have many algorithms such as Naïve Bayes, KNN, and SVM. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

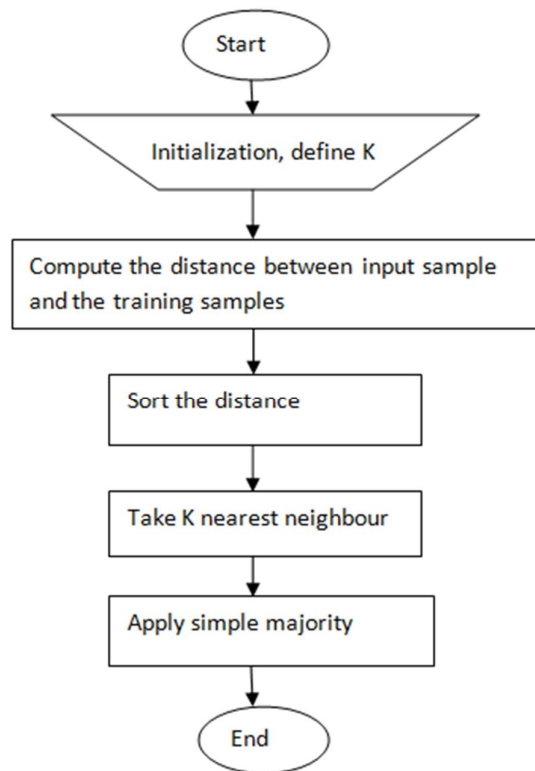
Chronic kidney disease (CKD) is a type of kidney disease in which there is gradual loss of kidney function over a period of month or years. Causes of chronic kidney disease include diabetes, high blood pressure etc. Chronic kidney disease affected 753 million people globally in 2016, including 417 million females and 336 million males. It increases the risk of cardiovascular disease, and people with CKD often have other risk factors for heart disease, such as high blood lipids. The most common signs and symptoms of CKD include anemia, blood in urine, decreased mental alertness, edema, fatigue, hypertension, nausea, muscle cramps etc. It is important for kidney disease to be diagnosed and treated early for serious damage to be prevented.

In this paper we use Data mining techniques called KNN and SVM on our database for the analysis. There is no current cure for chronic kidney disease. However, some therapies can help control the signs and symptoms, reduce the risk of complications, and slow the progression of the disease. Some conditions increase the risk of chronic kidney disease (such as diabetes). Controlling the condition can significantly reduce the chances of developing kidney failure.

A. K-Nearest Neighbor (KNN)

In pattern recognition, the K-Nearest Neighbor algorithm (K-NN) is a non parametric method used for classification and regression. In both cases, the input consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning. In KNN Classification, the output is a class membership. Classification is done by a majority vote of neighbors. If $K = 1$, then the class is single nearest neighbor. In a common weighting scheme, individual neighbor is assigned to a weight of $1/d$ if d is the distance to the neighbor. The shortest distance between any two neighbors is always a straight line and the distance is known as Euclidean distance. The limitation of the K-NN algorithm is it's sensitive to the local configuration of the data. The process of transforming the input data to a set of features is known as Feature extraction. In Feature space, extraction is taken place on raw data before applying K-NN algorithm.

The accuracy value of CKD prediction for KNN classifier optimal subset dataset (selected 8 attributes from 25) dataset by wrapper Sub Set value and Best First search feature selections are most satisfactory it has the maximum value accuracy. KNN without used feature selection has accuracy 98.5% in prediction CKD. KNN with Info Gain Attribute value with ranker (selected 15 Attribute) has list accuracy rate 98%. KNN with Info Gain Attribute value with ranker (selected 20 Attribute) has accuracy 98.75% it improved to from 98.5% to 98.75% the normal data.



II. RELATED WORKS

Data used for this will be extracted from the Medline plus in unstructured format and by using open source tools it will be converted to structured format. The Clustering of data according to requirements and then by applying association on it for the required output i.e. prediction of diseases and suggestion of appropriate medication [1].

The classification techniques are used to predict the risk level of each person based on pulse rate and the accuracy is high when using more number of attributes and provides the classified report for the heart disease through probability and minimized the error. In this approach each class has the maximum risk values is the major issue [2].

While detecting a disease number of tests should be required from the patient. Data mining technique has been used to reduce the number of tests, thereby it improves the performance. This paper analyzes data mining techniques which can be used for predicting different types of diseases. It mainly concentrates on predicting heart disease, Diabetes and Breast cancer etc [3].

By using Bayesian and KNN algorithm the database for diabetes patients has been created and analyzed them by taking various attributes of diabetes for prediction of diabetes disease and changes it into a reasonable structure for further use through the gathered data of diabetes to create a smart network to help the physicians. It will give accurate results whether the patient is diabetic or not with the help of the knowledge base of the larger dataset [4].

This paper enables clients to get moment direction on their medical problems through a social intelligent health care system. It allows users to share their symptoms and issues, also uncovering new biomedical and healthcare related knowledge to support clinical decision making. Biomedical sensors will sense body parameters real-time and server system will give predictions to tackle the conditions. The usage of paper work is high is the only drawback [5].

By using classification modeling techniques it predicts the likelihood of patients getting heart disease and finding previously unknown patterns and trends in the database to build predictive models also it predicts the wide range of conditions that affect the heart and the blood vessels in the body. It further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction. It does not work well on large datasets [6].

By using data mining techniques it measures the accuracy using a standard dataset as well as a collected data set based on the performance by extracting necessary and hidden predictive data from a huge database to address the issue in the prediction of heart disease. The major drawback is, it cannot be easily predicted and it is a difficult task that demands expertise and higher knowledge for prediction. It provides the optimized solution among the all other classifiers [7].

Generally, DNA has become an extremely useful tool for predicting disease and it does not trigger by a single genome but by a combination of genomes together. In this paper, the knowledge is derived by determining the probabilistic levels of a disease if the genes are mutated. The major issue in this paper is, it predicts the statistics by eliminating the redundant data and taking into account only those genes in close correlation [8].

The fuzzy algorithm has a positive effect on the outcome of diabetes disease and liver disorder prediction models; it leads to a remarkable increase in classification accuracy. The results obtained with the dataset is analyzed and compared with the existing methods. The major issue in this paper is, we have to tune this model to make more robust to work on other datasets [9].

By using different techniques to predict the prevalence of cardiovascular diseases and heart diseases in particular time by implementing a variety of algorithms and the select number of factors required for efficient prediction having high mortality rate and risk to cause various disabilities [10].

By using Data mining techniques it performs early prediction of heart diseases and absolute relative error has been found and it offers a number of techniques to detect hidden patterns from the data that are difficult to trace with traditional statistical methods. The extraction of a large dataset and its analysis from different perspectives to retrieve useful information about data can be done. It includes prevention of hospital errors, early sensing, and cure of diseases, preventable hospital deaths etc., the algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. Absolute relative error is minimum as compared to all other algorithms. The major issue is, some of the algorithm used in this paper produces a maximum error [11].

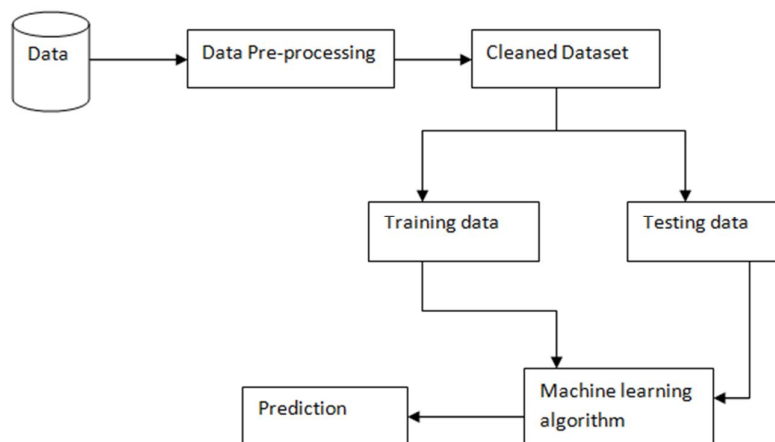
Generally, DNA has become an extremely useful tool for predicting disease and it does not trigger by a single genome but by a combination of genomes together. In this paper, the knowledge is derived by determining the probabilistic levels of a disease if the genes are mutated. The genetic information in a genome is held within genes, and the complete set of this information is useful for identifying and predicting diseases. Pathway analysis is used to extracting and explaining the high throughput molecular measurements for validation to computing the results. Net Beans IDE is used to run all the modules to provide the desired records from the huge amount of data based on the user input. By using this approach we have visualized pathways for each gene responsible for a disease and the statistics of occurrence of a disease is predicted [12].

III. SYSTEM METHODOLOGY

A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.

The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. Less used names for this kind of data sets are data corpus and data stock. An example of this type is the data sets collected by space agencies performing experiments with instruments aboard space probes. Data sets that are so large that traditional data processing applications are inadequate to deal with them are known as big data.

The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. Less used names for this kind of data sets are data corpus and data stock. An example of this type is the data sets collected by space agencies performing experiments with instruments aboard space probes. Data sets that are so large that traditional data processing applications are inadequate to deal with them are known as big data.



The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. Less used names for this kind of data sets are data corpus and data stock. An example of this type is the data sets collected by space agencies performing experiments with instruments aboard space probes. Data sets that are so large that traditional data processing applications are inadequate to deal with them are known as big data.

The values may be numbers, such as real numbers or integers, for example representing a person's height in centimeters, but may also be nominal data (i.e., not consisting of numerical values), for example representing a person's ethnicity. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values are normally all of the same kind. However, there may also be missing values, which must be indicated in some way.

In statistics, data sets usually come from actual observations obtained by sampling a statistical population, and each row corresponds to the observations on one element of that population. Data sets may further be generated by algorithms for the purpose of testing certain kinds of software. Some modern statistical analysis software such as SPSS still presents their data in the classical data set fashion. If data is missing or suspicious an imputation method may be used to complete a data set.

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values impossible data combinations missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. Often, data preprocessing is the most important phase of a machine learning project, especially in computational biology.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data preprocessing includes cleaning, Instance selection, normalization, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set.

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities.. Some data cleansing solutions will clean data by cross checking with a validated data set. A common data cleansing practice is data enhancement, where data is made more complete by adding related information. Standardization of data is a means of changing a reference data set to a new standard. In machine learning, the study and construction of algorithms that can learn from and make predictions on data is a common task. Such algorithms work by making data-driven predictions or decisions, through building a mathematical model from input data.

The data used to build the final model usually comes from multiple datasets. In particular, three data sets are commonly used in different stages of the creation of the model .The model is initially fit on a training dataset, that is a set of examples used to fit the parameters (e.g. weights of connections between neurons in artificial neural networks) of the model. The model (e.g. a neural net or a naive Bayes classifier) is trained on the training dataset using a supervised learning method (e.g. gradient descent or stochastic gradient descent).

In practice, the training dataset often consist of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), which is commonly denoted as the target (or label). The current model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation.

Test data is data which has been specifically identified for use in tests, typically of a computer program. Some data may be used in a confirmatory way, typically to verify that a given set of input to a given function produces some expected result. Other data may be used in order to challenge the ability of the program to respond to unusual, extreme, exceptional, or unexpected input.

Test data may be produced in a focused or systematic way (as is typically the case in domain testing), or by using other, less-focused approaches (as is typically the case in high-volume randomized automated tests). Test data may be produced by the tester, or by a program or function that aids the tester. Test data may be recorded for ore-use, or used once and then forgotten.

Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown. Prediction derives the relationship between a thing you know and a thing you need to predict for future reference. For example, prediction models in data mining are used by a marketing manager who predict that how much amount a particular customer will spend during a sale, so that upcoming sale amount can be planned accordingly. The prediction in data mining is known as Numeric Prediction. Generally regression analysis is used for prediction.

Machine learning is a way of identifying patterns in data and using them to automatically make predictions or decisions. In this data science course, you will learn basic concepts and elements of machine learning. The two main methods of machine learning you will focus on are regression and classification.

Machine learning tasks are classified into several broad categories. In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each image would have a label (the output) designating whether it contained the object. In special cases, the input may be only partially available, or restricted to special feedback. Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels.

Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. Examples of a continuous value are the temperature, length, or price of an object.

In unsupervised learning, the algorithm builds a mathematical model from a set of data which contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of "features", or inputs, in a set of data.

IV. EXPERIMENTAL AND RESULTS

The obtained experimental results are shown as two data mining algorithms which are KNN (K-Nearest Neighbor) and SVM (Support Vector Machine). Feature evaluation or extraction and classification are done on the basis of the dataset that is being used. All the outcomes are discussed in the graphical sequence. The decision is taken on the basis of few variables with higher significance, which is selected in the first layer of classification model using k-nearest neighbor classifier. In this paper, the application of the multi-layered classification can be described from the following figure.

By deal with various factors that affects prediction performance (such as class imbalance, incomplete learning etc.) for improving the prediction accuracy and finally identifying the exact cause (through checking very high similarity by generating human interpretable rules through KNN algorithm.

TABLE I: SELECTED FEATURES FOR CHRONIC KIDNEY DISEASE USED IN EXPERIMENT

S. No	Attribute	Description
1	Age(nominal)	Age in years
2	Blood Pressure(nominal)	Bp in mm
3	Specific gravity(nominal)	Sg
4	Albumin(nominal)	Al
5	Sugar(nominal)	Su
6	Red blood cells(nominal)	rbc
7	Pus cells(nominal)	Pc
8	Pus cell clumps(nominal)	pcc
9	Bacteria(nominal)	ba
10	Blood glucose random(nominal)	bgr in mgs
11	Blood urea(nominal)	Bu in mgs
12	Serum creatinine(nominal)	Sc in mgs
13	Sodium(nominal)	Sod in mEQ
14	Potassium(nominal)	Pot in mEQ
15	Haemoglobin(nominal)	Hemo in gms
16	Packed cell volume(nominal)	Pcv
17	White blood cell count(nominal)	Wc in cells
18	Red blood cell count(nominal)	Rc in millions
19	Hypertension(nominal)	htn-(yes,no)
20	Diabetes mellitus(nominal)	dm-(yes,no)
21	Coronary artery disease(nominal)	Cad-(yes,no)
22	Appetite(nominal)	Appet-(good,poor)
23	Pedal edema(nominal)	Pe-(yes,no)
24	Anaemia(nominal)	Ane-(yes,no)
25	Class(nominal)	class

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.58 seconds

=== Summary ===

Correctly Classified Instances	354	88.5	%
Incorrectly Classified Instances	46	11.5	%
Kappa statistic	0.7688		
Mean absolute error	0.1148		
Root mean squared error	0.2803		
Relative absolute error	24.4815	%	
Root relative squared error	57.905	%	
Total Number of Instances	400		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.816	0.000	1.000	0.816	0.899	0.790	0.986	0.990	ckd
	1.000	0.184	0.765	1.000	0.867	0.790	0.986	0.955	notckd
Weighted Avg.	0.885	0.069	0.912	0.885	0.887	0.790	0.986	0.977	

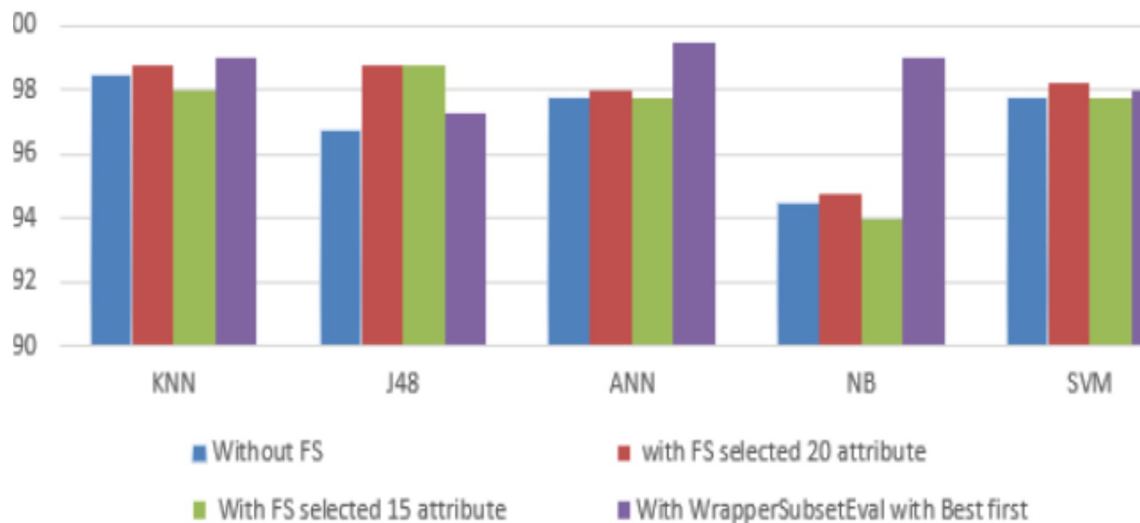
=== Confusion Matrix ===

```

a  b  <-- classified as
204 46 | a = ckd
  0 150 | b = notckd

```

Accuracy comparison of classifiers for Chronic Kidney disease datasets with Feature selection.



V. CONCLUSION

Results show that, first except of KNN which is sensitive to scale of data, performance of other classifiers are almost the same for original and normalized data set. Second, same results are obtainable using 8 or 9 features

instead of 24 features which validate the results correlation matrix which shows high correlativity between the features. For the analysis process, this model is very useful to decide their predictions on various cases.

REFERENCES

- [1] Disha mahajan, Mrudhula palak, Shubhangi pankore, Saniya path an, Deepaabin, "Prediction System for Diseases and Suggestion of Appropriate Medicines", International Journal of Innovations & Advancement in Computer Science, pp.348-350, India, 2017.
- [2] R. Theresa prince, J. Thomas, "Human Heart Disease Prediction System Using Data Mining Techniques", 2016 International Conference On Circuit, Power and Computing Technologies[ICCPCT], pp.1157-60, India, 2016.
- [3] K. Gomathi Kamaraj, D.Shanmuga Priyaa, "Multi Disease Prediction Using Data Mining Techniques", Article can be accessed for this publication at <https://www.researchgate.net/publication/319851535>, pp.12-14, India, 2016.
- [4] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction using Data mining", 2017 International Conference On Innovations In information, Embedded and Communication Systems(ICIECS), pp.304-308, India, 2017.
- [5] Puja Sarage, trupti Sudrik, Kalyani Zodage, "Health Prediction System By Using Data Mining", International Journal for Research in Applied Science & Engineering technology(IJRASET), pp.44-46, India, 2017.
- [6] Aditya Methaila, Prince kansal, Himanshu Arya, Pankaj Kumar, "Early heart Disease Prediction Using data Mining techniques", Netaji Subhas Institute of Technology, pp.53-59, India, 2014.
- [7] Marjia Sultana, Afrin Haider, Mohammad Shorif Uddin, "Analysis of Data Mining Techniques for Heart disease Prediction, Department of computer science and Engineering, Jahangirnagar University, pp.149-153, India, 2016.
- [8] Syeeda Farah, Dr.Asha T, Cauvery B and Shivanand K S, "DNA Based Disease Prediction using pathway Analysis", 2017 IEEE 7th International Advance Computing Conference, pp.629-634, India, 2017.
- [9] Naganna Chetty, Kunwar Singh Vaisla, Nagamma Patil, "An Improved Method for Disease Prediction using Fuzzy Approach", 2015 Second International Conference on Advances in Computing and Communication Engineering, pp.568-571, India, 2015.
- [10] Hidayet Takci, "Improvement Of Heart Attack prediction By The Feature Selection Methods", Turkey Journal Of Electrical Engineering & Computer Science, pp.1611-235, Turkey, 2018.
- [11] Asem H. Shurab, Ashraf Y. A. Maghari, "Blood Diseases Detection Using Data Mining Techniques", 2017 8th International Conference on Information Technology (ICIT), pp.625-631, London, 2017.
- [12] Mudasir M Kirmani, "Cardiovascular Disease Prediction Using Data Mining Techniques", An International Journal Of Oriental Scientific Publishing, pp.520-528, India, 2017.
- [13] N.K Salma Banu, Suma Swamy, "Prediction Of Heart Disease At Early stage Using Data Mining And Big Data Analytics", 2016 International Conference On Computer And Optimization Techniques, pp.256-260, India, 2016.
- [14] Sambasiva Rao Voleti, Kiran Kumar Reddi, "Design Of An Optimal Method For Disease Prediction Using Data Mining Techniques", International journal Of Advanced Research In Computer science and Software Engineering, pp.328-30, India, 2016.
- [15] Sonu S R, Ravi Ranjan, Vivek Prakash, Saritha K, "Prediction of Parkinson's Disease Using Data Mining", International Conference On Energy , Communication, Data Analytics And Soft Computing(ICECDS), pp.1082-1085, India, 2016.
- [16] Dr. A. R. Pon Periasamy, K. Arutchelvan, "Data Mining Techniques in Multiple Cancer Prediction", International Journal of Advanced Research in Computer Science and Software Engineering, pp.472-475, India, 2017.
- [17] Jabbar MA*, "Prediction of heart disease using k-nearest neighbor and particle swarm optimization", Vardhaman College of Engineering, Hyderabad, pp.4154-4158, India, 2017.
- [18] Guneet Kaur, Ajay Sharma, "Predict Chronic Kidney Disease Using Data Mining Algorithms In Hadoop", International Journal of Advances in Electronics and Computer Science, pp.2835-2393, India, 2018.
- [19] Amanpreet Kaur, "A Comprehensive Approach to Predict Heart Diseases using Data Mining", International Journal of Innovations in Engineering and Technology (IJET), pp.1058-2319, India, 2017.
- [20] Haldurai Lingaraj , Rajmohan Devadass , Vidya Gopi , Kaliraj Palanisamy, "Prediction Of Diabetes Mellitus Using Data Mining Techniques", Journal of Bioinformatics & Cheminformatics, pp.54-60, India, 2015.