RESEARCHERID
THOMSON REUTERS

[Prerna, 5(9): September 2018] ISSN 2348 – 8034
DOI- 10.5281/zenodo.1408013 Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

NOVEL APPROACH FOR PREDICTION OF HEART DISEASES IN DATA MINING

Prerna*[1]& Er. Ram Singh[2]

[*1]Research Scholar of Computer Science Engineering, Punjabi University Patiala (Punjab)
[*2]Assistant Professor of Computer Science Engineering, Punjabi University Patiala (Punjab)

ABSTRACT

The data mining is the technique to analyze the complex data. The prediction analysis is the technique which is applied to predict the data according to the input dataset. In the recent times, various techniques have been applied for the prediction analysis. In this work, the k-means clustering algorithm and SVM (support vector machine) classifier based prediction analysis technique is used for clustering and classification of the input data. In order to increase the accuracy of prediction analysis, the back propagation algorithm is proposed to be applied with the k-means clustering algorithm to cluster the data. The proposed algorithm performance is tested in the heart disease dataset which is taken from the UCI repository. The 14 attributes are selected from the dataset for the classification. Specifically, machine learning researchers have used Cleveland database, particularly at all times. The proposed work will also be compared with the existing scheme (using arithmetic mean) in terms of accuracy, fault detection rate and execution time.

Keywords*: Data mining, Classification, Clustering, K-means, SVM, Back propagation.*

I. INTRODUCTION

The process of extraction of interesting knowledge and patterns to analyze data is known as data mining. To ensure the accuracy and correctness of data by altering it within the given storage resource is known as data cleaning process. Within several software and data storage systems, data cleaning can be achieved in several ways. The Cardiovascular / Heart Disease is a class of disease that affects the heart and blood vessel many of which are related to a process is called atherosclerosis. Mainly, the data sets and protocols those are relevant to specific data storage system are focused and reviewed here [1]. The process through which data can be converted from one format to another is known as data transformation. Mainly, a new destination system is generated from the source system's format. In case if the data that is stored within the database or data warehouse is accurate and consistent, then the data is known as integral. The state, process or function of data can be known through data integrity and it is also as data quality [2]. The first step is data cleaning which is used to remove noise and irrelevant data. Second, a step is data integration which is used to combine multiple data sources. In third step, data are retrieved from the database that comes under the step of data selection. When aggregations and summary operations are applied, the transformation or consolidation of data is done such that appropriate data can be generated within the fourth step. For the extraction of data patterns, data mining is an important process which applies different intelligent methods then knowledge-based interesting patterns is identified using pattern evaluation. With the help of knowledge representations and visualization approaches, the users are presented with the mined knowledge within the final step. The data mining process extract a large amount of data in order to acquire knowledge which is termed as a misnomer. In today's world, a large amount of data is collected on a daily basis to analyze which is complex in nature. The data which is in the data warehouse is present in the raw form thus data mining is the procedure of mining knowledge from raw data [3]. It became overwhelming as the massive collections of data stored and it was difficult to handle such huge information. Therefore, to solve these issues database management system (DBMS) and structured databases are created. Whenever it is required to retrieve the particular information from a large amount of data an important role is played by efficient database management systems. It became easier to gather all sorts of information due to increase usage of database management systems. Data mining has great success in many applications and is primarily used today by many companies such as communication, financial, retail and marketing

organizations. For the specific customer segments, a retailer can use records of the customer purchase in order developed product and promotions using data mining. It plays a critical role when it is impossible to enumerate all application. On the basis of the given output prediction for the certain outcomes can be done by the classification process [4]. The outcome is predicted by the processed algorithm as the different set of attributes are present in the training set and in the respective outcome which is called as goal or prediction attribute. The relationship between the attributes is discovered by the algorithm that will be helpful in the prediction of the outcome. A data set is provided by the algorithm known as the prediction set in which the same set of attributes are present but the prediction attribute is absent that is not well known. The input is analyzed by the algorithm that is helpful in the prediction process. In order to provide a solution to the data mining classification issues, Genetic Programming is utilized widely by the applications. Optimal results are produced by the GP with global search issues such as classification [5]. Several 'peaks' are present in the search space for classification this cause local search algorithms also known as simulated annealing that performs badly. Stochastic search algorithms are present in the GP on the basis of the abstraction process of Darwinian evolution. Neural networks are the interconnectivity between the processing elements also called units, nodes, or neurons. These networks are designed after the cognitive processes of the brain. These networks are used to predict new outcomes from the previous observations. In order to produce an output function all the present neurons within the network work together. Ant Colony algorithms are the naturally inspired technique and by the behavior of ants as they help in finding the optimal path from the colony to food. They use the good paths within a graph in order to find optimal ways [6]. The group membership for data instances can be predicted with the help classification technique within the data mining. In order to predict the data for example classification can be utilized by the applications on a specific day to identify the weather which can be either "sunny", "rainy" or "cloudy". Two steps are followed by this process. A set of predefined classes can be presented through the construction of a model. Each sample that is predicted to be belonging to a predefined class is determined by class label attribute. In order to generate the training set model, several tuples are required. In different forms, these tuples are represented. The second step used in the classification is model usage. In order to classify the future and unknown objects, model usage is widely used as this model estimates the accuracy of the model [7]. The classified result from the model is used to compare with the known label of the test sample. A test set is not dependent on a training set. Few of the applications in which cluster analysis is applied include pattern recognition approach, image processing, and data analysis. The customer categorized group and purchasing patterns done by clustering can be used by the marketer to discover their customer's interest. In order to cluster the tasks being performed in low dimensional data sets, the k-means clustering algorithm is applied. K is utilized as a parameter here and the k clusters are generated by partitioning n objects [8]. A binary classifier through which the margin is increased is known as SVM classifier. This algorithm helps in performing classification in which all the data points present in individual class are separated by the best hyperplane. The best hyperplane of SVM can be presented on the basis of the highest margin present in the two classes. The simple probabilistic classifier that depends on Bayes' theorem is known as the Naive Bayes classifier which strong independence naïve assumption. This algorithm is also known as the independent feature model.

II. LITERATURE REVIEW

Tülay Karayilan, et.al **(2017)** proposed heart disease is the fatal disease from which a large number of population is currently suffering as its detection and prevention are major and required to diagnose at the early stage. This disease causes the maximum numbers of causalities [9]. In the traditional methods there are various limitations as analyzed by doing experiments, therefore enhanced methods have been proposed in this paper. On the basis of machine learning, medical diagnosis system for the prediction of heart disease has been developed. For the prediction of the heart disease, a Back propagation algorithm has been proposed for the artificial neural network. Input used has the clinical features in which all the networks were trained using a back propagation algorithm. This is done for the neural network in order to determine the condition of the patient whether the patient is suffering from heart disease or not.

Ms. Tejaswini U. Mane, et.a**l (2017)**presented the survey performed by the world health organization in the worldwide for the heart disease in which every year more than 12 million deaths occur due to this fatal disease, therefore maximum casualties are caused due to which detection of this disease is necessary. The improvement in

the clustering K-means and decision tree algorithm in case of the hybrid approach is done by using ID3 for the classification purpose [10]. Heart disease can be diagnosed at the early stage using various parameters such as gender, age, chest pain, blood pressure, blood sugar and so on. As per performed experiments, it is concluded that a proposed technique provide optimal results for the prediction of the heart disease as compared to other techniques as it improves the treatment process and provides better clinical decision making.

M. A. Jabbar, et.al **(2016)** presented that coronary heart disease is the most fatal heart disease as a large amount of the deaths occur due to this disease in the worldwide. The author in this paper discussed the use of data mining techniques in the medical system [11]. These techniques provide the idea to the doctors whether the patient is suffering from any heart disease or not. Hidden Naïve Bayes is the extended version of the traditional Naïve Bayes method in the data mining. The conditional independence assumption of the traditional method, in the data mining, is relaxed by using this model. For the classification and prediction of heart disease, Hidden Naïve Bayes has been utilized in accordance with the proposed model. On the basis of the performed experiments, it is concluded that Hidden Naïve Bayes (HNB) is superior to naïve Bayes in terms of optimal accuracy.

Theresa Princy, et.al **(2016)** discussed various data mining techniques have been utilized to detect the rate of the heart disease. For the effective and efficient diagnosis of the heart disease various Data mining techniques and classifiers have been utilized so far, discussed in this paper. All the obtained results compared to provide effective technique [12]. Various technologies and the different number of attributes has been utilized by many authors for their study. On the basis of a number of attributes taken different accuracy was provided by the different technologies. The risk rate of heart disease was detected with the help of KNN and ID3 algorithm and it also provides the accuracy level for a different number of attributes. It is concluded from the observation that using new algorithms the numbers of attributes could be reduced that increase the accuracy for the detection of the heart disease.

S. Rajathi, et.al **(2016)** proposed a technique in order to enhance the performance of the k-Nearest Neighbor (kNN) algorithm is the integration of Ant Colony Optimization technique. With the help of this method prediction of the heart disease becomes easy. Heart disease is considered as one of major disease that causes major causalities in the worldwide [13]. In this technique, there are two different phases. kNN algorithm was utilized in the initial phase for the classification of the test data. For the optimized solutions, the ACO technique was utilized as it initializes the population and searches to get the desired result. In order to present a dataset, Acute Rheumatic Fever (ARF) disease has been utilized that is related to a data set. kNNACO algorithm which is an integrated technique is proposed in this paper that is experimented and accuracy is evaluated in terms of accuracy and error rate performance.

Jagdeep Singh, et.al **(2016)** proposed health care services provide various medical facilities as well as protection against various diseases. Many frameworks have been developed in this paper for the prediction of the heat disease at the early stage using heart dataset. All these datasets is based on the associative classification techniques and this dataset is implemented on the dataset of Cleveland heart diseases in order to check various data mining techniques [14]. This Cleveland heart disease is a machine learning repository in the University of California Irvine (UCI). For the diagnosis of the heart disease, there are various parameters such as gender, age, chest pain, blood pressure, blood sugar and many more. As per the performed experiments, it is concluded that a hybrid technique has been utilized for the classification of associative rules (CARs) that provide the optimal accuracy.

Min Chen, Yixue Hao, et.al **(2017)** proposed a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm[15]. The data was gathered from a hospital which included within it both structured as well as unstructured types of data. In order to make predictions related to the chronic disease that had been spread within several regions, various machine learning algorithms were streamlined here. 94.8% of prediction accuracy was achieved here along with the higher convergence speed in comparison to other similar enhanced algorithms

ISSN 2348 – 8034
Impact Factor- 5.070

III. DATA SOURCE

In this paper, we use the heart disease data from machine learning repository of UCI [12]. We have total 303 instances of which 164 instances belonged to the healthy and 139 instances belonged to the heart disease. 14 clinical features have been recorded for each instance.

Table 1: there are 14 attributes used in this system

NO.	Name	Description
1	Age	Age in year
2	Sex	1=male, 0=female
3	cp	Chest pain type(1 = typical angina, 2 =atypical angina, 3 = non-anginal pain, 4= asymptomatic)
4	Trestbps (mmHg)	Resting blood sugar(in mm Hg on admission to hospital)
5	chol	Serum cholesterol in mg/dl
6	fbs	Fasting blood sugar>120 mg/dl(1= true, 0=false)
7	restecg	Resting electrocardiographic results(0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy)
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment
12	Ca	Number of major vessels (0-3)colored by flourosopy
13	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
14	Num	Diagnosis of heart disease (Class)

IV. PROPOSED METHODOLOGY

This research work is based on the prediction analysis of heart diseases. The prediction analysis is the technique in which future possibilities can be predicted based on the current dataset. In this research work, a technique of SVM is applied previously for the prediction analysis. One of the simplest algorithms amongst all the learning machine algorithms is the SVM algorithm. Since there are no assumptions made on the underlying data distribution, a decision tree is known to be a non-parametric supervised learning algorithm. Here, on the basis of nearest training samples present within the feature space, the samples are classified. The feature vectors are stored along with the labels of training pictures within the training process. Towards the label of its k-nearest neighbors, the unlabelled question point is doled out during the classification process. Through majority share cote, on the basis of labels of its

neighbors, the object is characterized. The object is classified essentially as the class of the object that is nearest to it in the event when k=1. k is known to be an odd integer in the case when there are only two classes present. During the performance of multiclass categorization, there can be a tie in the case when k is an odd whole number.

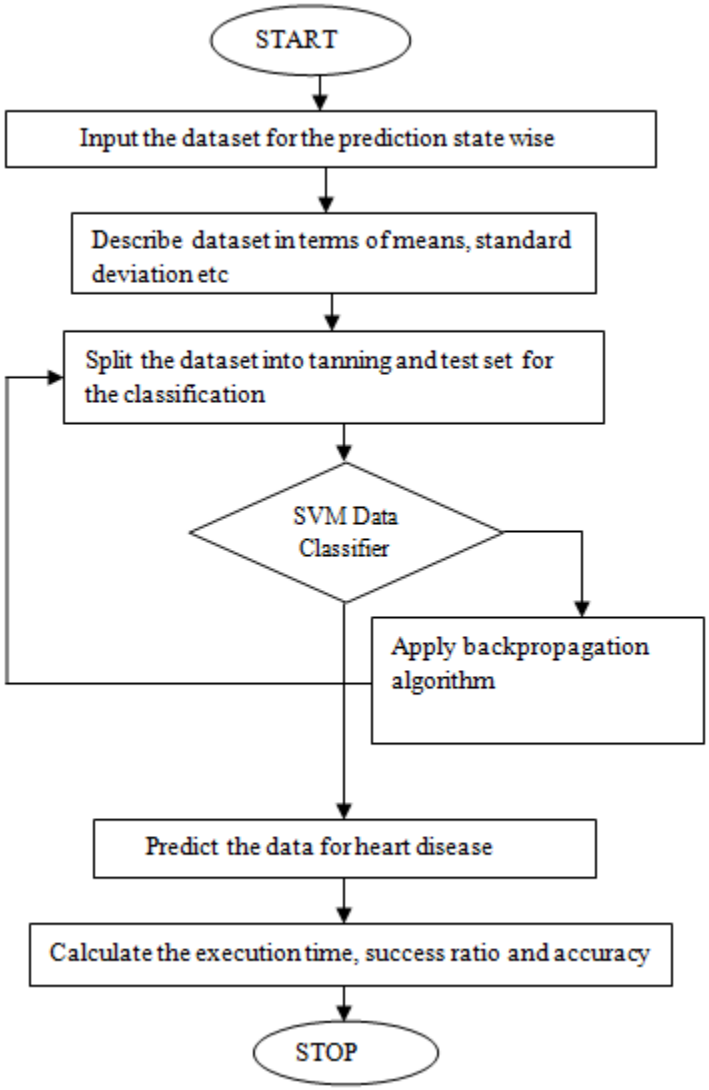


Figure 1: Proposed Methodology

V. TOOLS AND IMPLEMENTATION

A. Anaconda and Spyder Software:

A general purpose programming language which is utilized to process the text, numbers, images and other various data present within the modern computer operating systems is known as Python. There modules can be imported and exported very easily as well. The costs of both the standard library as well as the interpreter are zero in binary and source form which is a very important benefit of using this tool. A free open source distribution provided for Python and R programming languages is Anaconda. This tool helps in simplifying the packet management and deployment within the processing, analyzing and computing of large-scale data. Spyder tool is used for implementation of neural network and decision tree classifier using the Python.

[Prerna, 5(9): September 2018] ISSN 2348 – 8034
DOI- 10.5281/zenodo.1408013 Impact Factor- 5.070

B. Proposed Pseudo Code

```
Present the pattern to the network
        for each layer in the network
            for every node in the layer
                1. Calculate the weight sum of
the inputs to the node
                2. Add the threshold to the sum
                3. Calculate the activation for the
node
            end
        end
        for every node in the output layer
            calculate the error signal
        end
for all hidden layers
        for every node in the layer
            1. Calculate the node's signal error
            2. Update each node's weight in the
network
        end
    end
    Calculate the Error Function
while ((maximum  number of iterations <
than specified) AND
        (Error Function is > than specified))
```

VI. EXPERIMENTAL RESULTS

The proposed approach is implemented in Python and the results are analyzed by showing comparisons amongst proposed and existing approaches in terms of accuracy and execution time.

1. **Accuracy:** Accuracy is defined as the number of points correctly classified divided by a total number of points multiplied by 100, as shown in eqn. 1.

$$Accuracy = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100 \quad ---1$$

ISSN 2348 – 8034
Impact Factor- 5.070

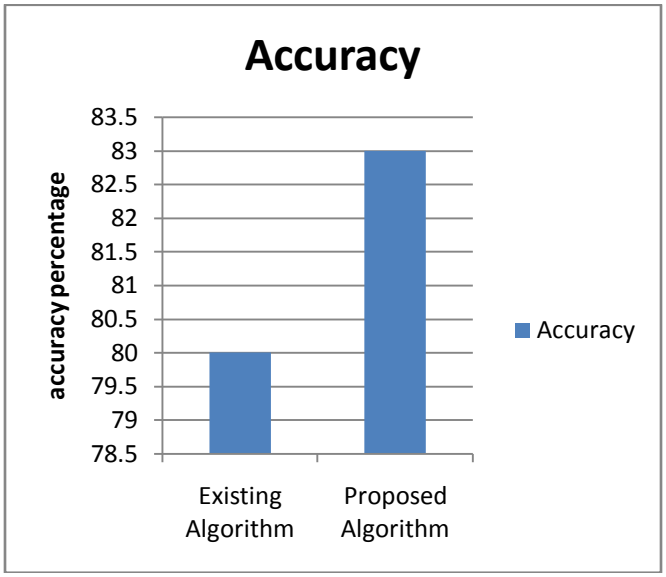


Figure 2: Accuracy Comparison

As shown in figure 2, the accurate comparison of the existing and proposed algorithm is shown. The accuracy of the proposed algorithm is high as compared to the existing algorithm.

2. **Execution Time**: Execution time is defined as the difference of end time when the algorithm stops performing and starts time when the algorithm starts performing as shown in eqn. 2.

   Execution time = End time of algorithm- the start of the algorithm --2

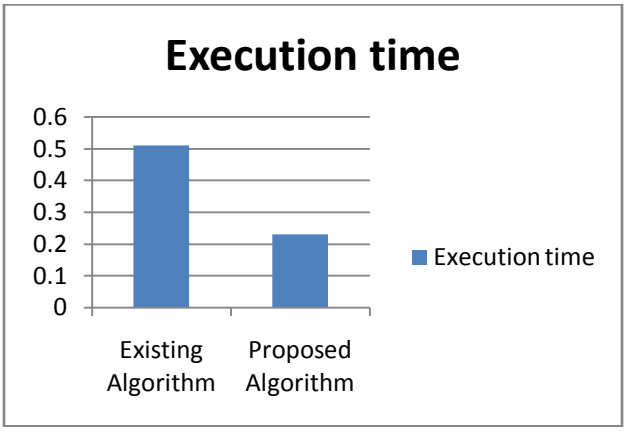


Figure 3: Execution time

As shown in figure 3, the execution time of the proposed and existing algorithm is shown. The execution time of the proposed algorithm is less as compared to the existing algorithm.

3. **Success Ratio**: Success rate is the fraction or percentage of success among a number of attempts

Success Ratio= $\frac{Number\ of\ points\ correctly\ classified}{Total\ Number\ of\ points}$

ISSN 2348 – 8034
Impact Factor- 5.070

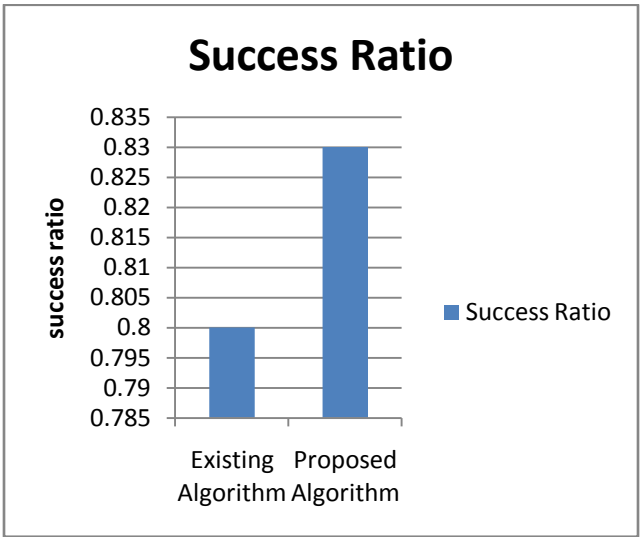
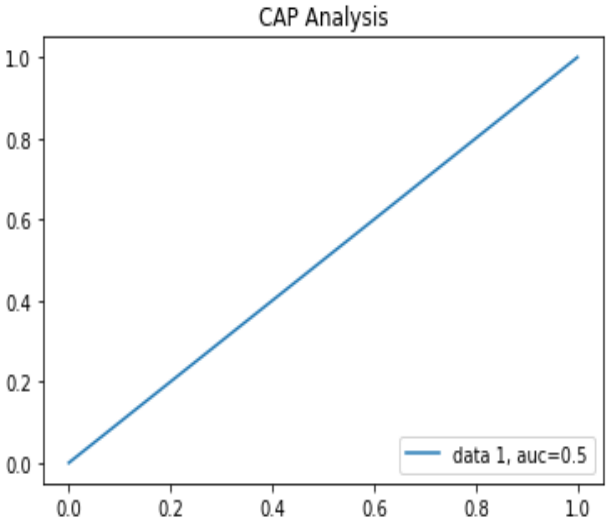


Figure 4: Success Ratio

As shown in figure 4, the success ratio of proposed and existing algorithm is compared for performance analysis. The success ratio of proposed algorithm is high as compared to existing algorithm

4. **CAP Analysis**: -A canonical analysis of the principal coordinates for any resemblance matrix, including a permutation test. CAP takes into consideration the structure of the data. So, it is more likely to separate your different levels if there is no strong difference and is good to show the interaction between factors



s

Figure 5: CAP Analysis

As shown in figure 4, the CAP analysis is shown in this figure. On the axis of this cure the training dataset is given as input and on the y-axis, the test data is given as input. The blue line shows that CAP curve which represents the accuracy of the classifier.

Table 2: Table of Comparison

Parameter	Existing Algorithm	Proposed Algorithm
Accuracy	87 percent	90 percent
Execution Time	0.14 second	0.7 second
Success Ratio	0.87	0.90

As shown in table 1, the existing and proposed algorithms are compared in terms of accuracy, execution time and success ratio. The proposed algorithm performs well in terms of all parameters as compared to existing algorithm

VII. CONCLUSION

The relevant information is fetched from the rough dataset using data mining technique. The similar and dissimilar data is clustered after calculating a similarity between the input dataset. The SVM used to classify both similar and dissimilar data type in which central point is calculated by calculating an arithmetic mean of the dataset. The central point calculated Euclidian distance is used to calculate a similarity between different data points. According to the type of input dataset a clustered data is classified using an SVM classifier scheme. In this research work, the back propagation algorithm is applied with the SVM classifier to increase the accuracy of prediction. The proposed algorithm performs well in terms of accuracy and execution time. In the future, a proposed technique will be further improved to design a hybrid classifier for the heart disease prediction.

REFERENCES

1. *Jagdeep Singh, Amit Kamra, Harbhag Singh, "Prediction of Heart Diseases Using Associative Classification", IEEE, vol. 4, issue 1, pp. 23-48, 2016.*
2. *AbdelghaniBellaachia and ErhanGuven (2010), "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, vol. 6, issue 3, pp. 234-239, 2010.*
3. *Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S (2012), "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, vol. 7, 2012, issue 4, pp-56-62, 2012.*
4. *P. Gupta and B. Kaur, "Accuracy Enhancement of Heart Disease Diagnosis System Using Neural Network and Genetic Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 103, no. 13, pp. 11-15, 2014.*
5. *Kajal C. Agrawal and Meghana Nagori (2013), "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm", International Conf. on Advances in Computer Science and Electronics Engineering, vol. 23, issue 4, pp. 546-552, 2013.*
6. *D. Oliver, F. Daly, F. C. Martin, and M. E. McMurdo, s"Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review," Age and aging, vol. 33, no. 2, pp. 122–130, 2004.*
7. *Min Chen, YixueHao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", 2017, IEEE, vol. 15, issue 4, pp- 215-227, 2017.*

8. *Akhilesh Kumar Yadav, DivyaTomar and Sonali Agarwal (2014), "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, issue 16, pp.121-126, 2013.*
9. *TülayKarayilanTülayKarayilan, "Prediction of Heart Disease Using Neural Network", IEEE, vol. 14, issue 1, pp. 423-468, 2017.*
10. *Ms. Tejaswini U. Mane, "Smart heart disease prediction system using Improved K-Means and ID3 on Big Data", 2017 International Conference on Data Management, Analytics and Innovation (CDMA), vol. 8, issue 11, pp. 123-148, 2017.*
11. *M. A. Jabbar, Shirinasamreen, "Heart disease prediction system based on hidden naïve Bayes classifier", vol. 4, issue 11, pp. 23-48, 2016.*
12. *Theresa Prince. R, J. Thomas, "Human Heart Disease Prediction System using DataMining Techniques", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], vol. 4, issue 1, pp. 23-48, 2016.*
13. *S.Rajathi, Dr.G.Radhamani, "Prediction and Analysis of Rheumatic Heart Disease using kNN Classification with ACO", IEEE, vol. 4, issue 7, pp. 223-248, 2016.*
14. *Jagdeep Singh, Amit Kamra, Harbhag Singh, "Prediction of Heart Diseases Using Associative Classification", IEEE, vol. 7, issue 9, pp. 23-48, 2016.*
15. *Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", 2017, IEEE, vol. 15, pp- 215-227, 2017*