ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Website: www.ijircce.com*

Vol. 7, Issue 4, April 2019

# Evaluate and Classification based Scheme for the Heart Disease Prediction

Kumari Samriti[1], Er. Monika Pathania[2]

M. Tech Student, Department of Computer Engineering, Bells Institute of Management & Technology, Shimla, India[1]

Assistant Professor, Department of Computer Engineering, Bells Institute of Management & Technology, Shimla, India[2]

**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 the base paper, neural network technique is applied for the prediction analysis. In the technique whole data is divided into testing and training part. The test data is classified into two classes' means first class is of data which have regional disease and second which does not have regional disease. In this research work, further improved will be proposed in this existing method using back propagation algorithm. The proposed improvement increase accuracy of classification and reduce execution time.

## I. INTRODUCTION TO DATA MINING

The process of extraction of interesting knowledge and patterns to analyze data is known as data mining. In data mining there are various data mining tools available which are used to analyze different types of data. Decision making, market basket analysis, production control, customer retention, scientific discovers and education systems are some of the applications that use data mining in order to analyze the collected information [1]. The multimedia, object relational, relational and data ware houses are some of the databases for which data mining has been studied.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

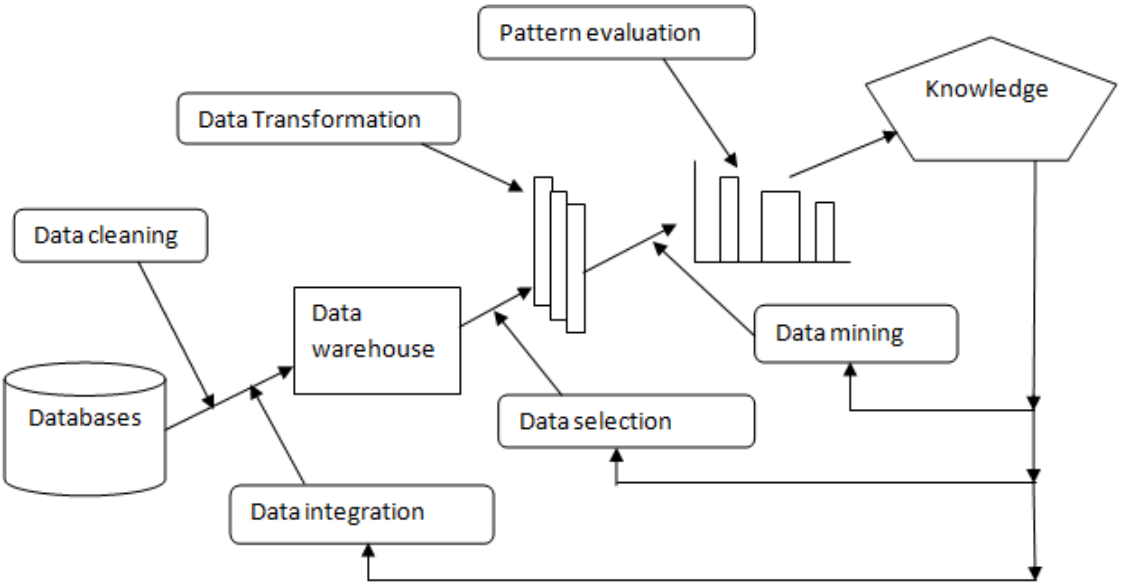



**Fig. 1.1 Data Mining Process**

**1.2 Clustering in Data Mining**

Image processing, market research, data analysis and pattern recognition, are some of the applications that use cluster analysis [2]. The customer categorized group and purchasing patterns done by clustering can be used by marketer to discover their customer's interest. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. In a city, similar houses and lands area can be identified by employing clustering in geology. To discover new theories, information clustering can be used that classify all documents available on Web.

In order to perform clustering, various algorithms have been proposed over time. Following are some broader categories into which the clustering methods have been categorized:

a. **Partitioning Methods:-** The gathering of samples that are of high similarity in order to generate clusters of similar objects is the basic functioning of this method. Here, the samples that are dissimilar are grouped under different clusters from similar ones. These methods completely rely on the distance of the samples [3].
b. **Hierarchical Methods:-** A given dataset of objects are decomposed hierarchically within this technique. There are two types in which this method is classified on the basis of type of decomposition involved. They are agglomerative and divisive based methods. A bottom up technique in which the formation of separate group is the first step performed is known as agglomerative technique. Further, the groups that are near to each other are merged together.
c. **Density Based Methods:-** The distance amongst the objects is taken as a base in order to separate the objects into clusters in most of the technique. However, these methods can only be helpful while identifying the spherical shaped clusters. It is difficult to obtain arbitrary shaped clusters within these techniques.
d. **Grid Based Methods:-** A grid structure is generated by quantizing the object space into finite number of cells which is known as grid based method. This method has high speed and does not depend on the number of data objects available.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Website: www.ijircce.com*

Vol. 7, Issue 4, April 2019

### 1.3 K-Means Clustering Algorithm

The basic partitioning based method which is used by various clustering tasks that are performed within the low dimensional data sets is known as k-means clustering algorithm. K is utilized as a parameter here and the k clusters are generated by partitioning n objects. It ensures that the similar types of objects are grouped within one cluster and dissimilar objects are placed in separate clusters [4]. The cluster centres are identified here with the help of this algorithm. It is made sure that there is a reduction of the sum of the squared distances of each data point to the nearest centre of cluster. Initially, the k objects are chosen on random basis.

### 1.4 SVM classifier

SVM stands for support vector machine. It is a binary classifier that maximizes the margin. The best hyper plane which separates all the data points of an individual class can be identified through the classification provided by SVM. The largest margin between the two classes describes the best hyper plane for an SVM [6].

## II. LITERATURE SURVEY

**Min Chen, et.al (2017)** proposed in this paper [7], a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm. 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.

**Akhilesh Kumar Yadav, et.al (2013)** presented in this paper [8], that different analytic tool has been used to extract information from large datasets such as in medical field where a huge data is available. The SGPGI real data set has been used that are always linked with different challenges. The classification becomes inefficient due to noise, high dimensional and missing values. Due to the different challenges have to face while performing data analytics clustering is used in replace of it.

**Sanjay Chakrabotry et.al, (2014)** stated in this paper [9], that powerful tool clustering is used as different forecasting tools. The weather forecasting has been performed using proposed incremental K-mean clustering generic methodology. The purpose behind this paper is to analyze air pollution for it they have used dataset of west Bengal. The clusters peak mean values are used to develop a weather category list and K-means clustering is applied on the dataset of air pollution. The weather category has been defined in different clusters and a new data is checked by incremental K means to group it into existing clusters.

**Chew Li S. et.al, (2013)** presented in this paper [10] particular university student results has been recorded to keep a track using Student Performance Analysis System (SPAS). The design and analysis has been performed to predict student's performance using proposed project on their results data. The data mining technique generated rules that are used by proposed system to gives enhanced results in predicting student performance. The student's grades are used to classy existing student using classification by data mining technique.

**Qasem A. et.al, (2013)** presented in this paper [11] that data analysis prediction is considered as import subject for forecasting stock return. The data analysis future can be predicted through past investigation. The past historical knowledge of experiments has been used by stock market investors to predict better timing to buy or sell stocks. There are different available data mining techniques out of all a decision tree classifier has been used by authors in this work.

**K.Rajalakshmi et.al, (2015)** presented in this paper [12] a study related to medical fast growing field authors. In this field every single day a large amount of data has been generated and to handle this much of large amount of data is not an easy task. So, this data need to be handled properly for it different technologies need to be used after that a data need to be mined to turn it into useful pattern.

**Bala Sundar V et.al, (2012)** examined in this paper [13] real and artificial datasets that have been used to predict diagnosis of heart diseases with the help of a K-mean clustering technique results to check its accuracy. The clusters are partitioned into k number of clusters by clustering which is the part of cluster analysis and each cluster has its observations with nearest mean. The first step is random initialization of whole data then a cluster k is assigned to each cluster. In case of proposed technique k assigned clusters are further divided into k number of groups and a distance square of sum has been minimized.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

**Daljit Kaur et.al (2013)** explained in this paper explained [14] that data contained similar objects has been divided using clustering. A data of similar objects are in same group and in case dissimilar objects occur then it will be compared with other group's objects. In order to cluster a data K-means algorithm has gain lots of popularity but using it is expensive and even initial centroid selection is the factor that defines its final results quality.

## III. RESEARCH METHODOLOGY

The prediction analysis is the technology which can predict the future possibilities from the existing data. The multimodal disease risk prediction algorithm is implemented in the basepaper, which can predict regional diseases from the existing data. In the basepaper, technique of k-mean clustering is applied which can arithmetic mean of the whole dataset is calculated which will be the centered point. The Euclidian distance is calculated from the centered point to analyze data similarity. The data which is similar can be clustered in one cluster and another in the second cluster. The clustered data will be given as input for the classification in which SVM classifier is used to classify data. In this research work, improvement in the k-mean clustering will be applied in which to calculate centered point algorithm of back propagation will be applied. The back propagation algorithm will take attribute number and instance value as input and gave result in the form of relationship between the attributes. In the last step, the SVM classifier will be applied which can classify data into two classes. The first class will be of the instances which have regional disease and second class is of an instance which does not have regional disease.

### 3.1. Performance Parameters

The performance of the proposed method is compared in terms of certain parameters. The performance is analyzed in terms of accuracy, execution time and success ratio. These parameters are described below:-

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

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

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

$$\text{Execution time} = \text{End time of algorithm- start of the algorithm}$$

3. *Success Ratio:* Success Ratio is defined as the number of points correctly classified divided by total number of points, as shown in equation
4. $$\text{Success Ratio} = \frac{Number\ of\ points\ correctly\ classified}{Total\ Number\ of\ points}$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

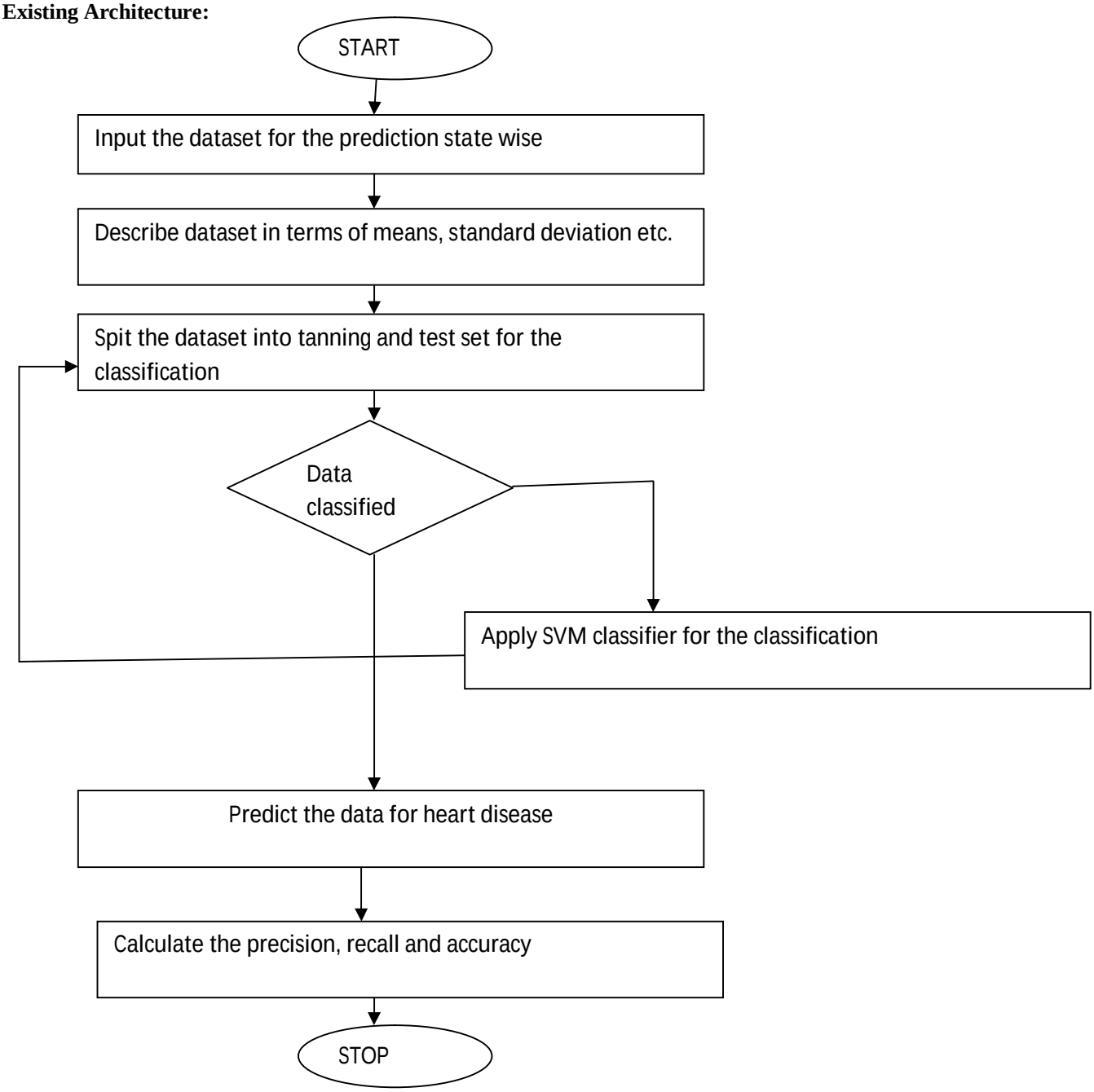
International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

**3.2. Existing Architecture:**



ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

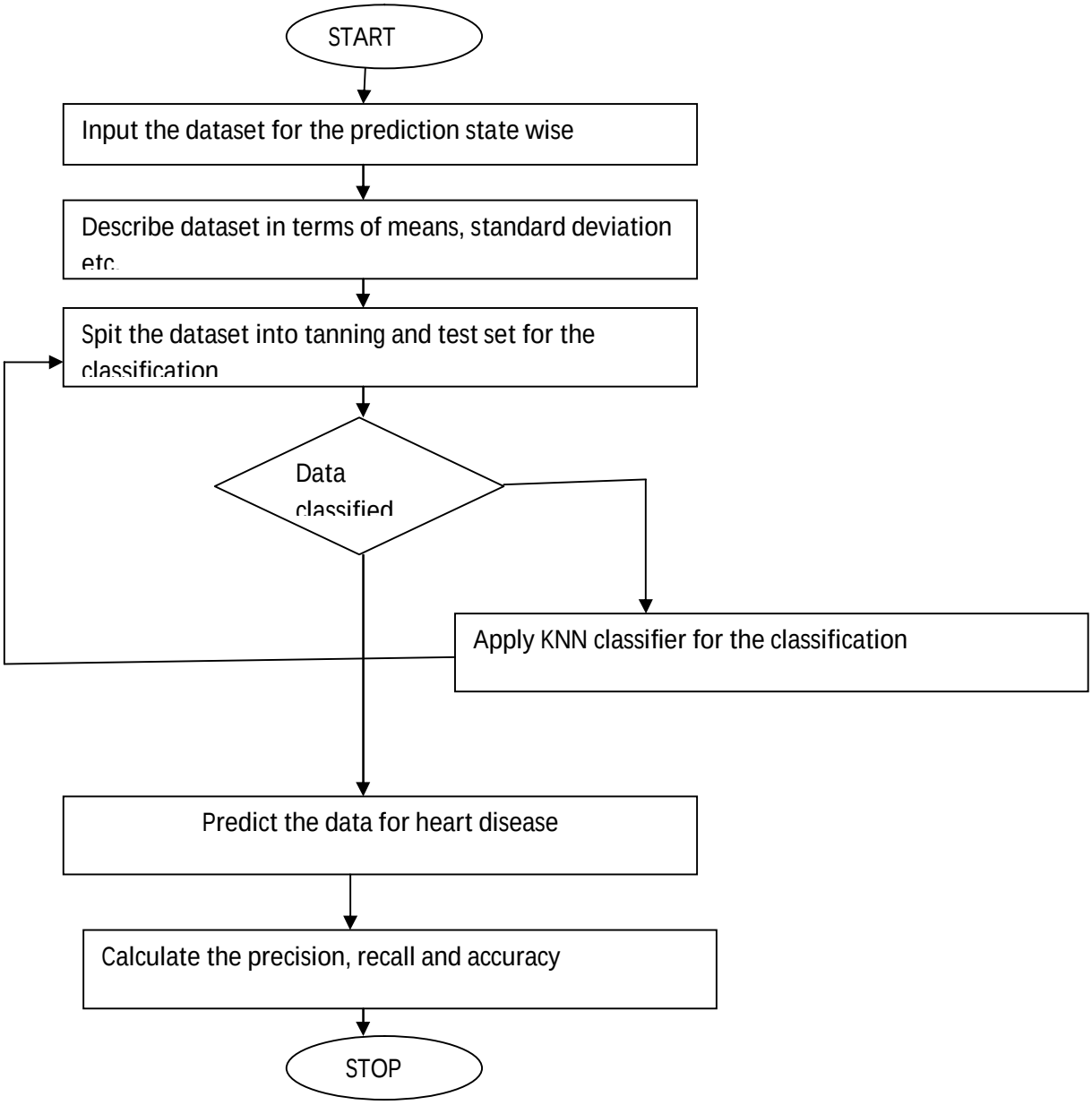
International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

**3.3 Proposed Architecture:**




## IV. CONCLUSION

The relevant information is fetched from rough dataset using data mining technique. The similar and dissimilar data is clustered after calculating a similarity between input dataset. The k-mean clustering is used to cluster 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 SVM classifier scheme in the last step. The clustering accuracy get

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

reduced when some of the data points remain uncluttered that has been concluded in this work. In this work technique will be proposed which calculated Euclidian distance in the iterative manner and increase clustering accuracy

## REFERENCES

[1] Abdelghani Bellaachia and Erhan Guven (2010), "*Predicting Breast Cancer Survivability Using Data Mining Techniques*", Washington DC 2005, vol. 6, 2010, pp. 234-239.
[2] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C (2010), "*Application of k-Means Clustering algorithm for prediction of Students' Academic Performance*", International Journal of Computer Science and Information Security, vol. 7, 2010, pp. 123-128.
[3] Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed (2012), "*Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity*", Middle-East Journal of Scientific Research, vol. 12, 2012, pp. 959-963.
[4] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S (2012), "*Reducing the Time Requirement of K-Means Algorithm*" PLoS ONE, vol. 7, 2012, pp-56-62.
[5] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed (2012), "*Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity*," Middle-East Journal of Scientific Research, vol. 5, 2012, pp. 959-963
[6] 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, 2013, pp. 546-552.
[7] 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, 2017, pp- 215-227
[8] Akhilesh Kumar Yadav, Divya Tomar and Sonali Agarwal (2014), "*Clustering of Lung Cancer Data Using Foggy K-Means*", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, 2013, pp.121-126.
[9] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), "*Weather Forecasting using Incremental K-means Clustering*", vol. 8, 2014, pp. 142-147.