Descriptive Data Analysis of Contagious Diseases using Statistical Parameters

Abhijeet Sudhakar[#1], Rajendra B. Patil[*2], Srivaramangai R[#3]

[#1]*Student, Thadomal Shahani Engineering College, Mumbai, India*
[#2] *Assistant Professor, S.K.Somaiya College of Arts, Science & Commerce, Mumbai, India*
[#3] *Assistant Professor, Department of Information Technology, University of Mumbai, Mumbai, India*

Abstract

Data Analytics techniques are popularly acceptable and widely preferred by researcher for analyzing the data. This includes descriptive analysis using statistical methods, classification techniques to segregate data, prediction to predict and various other techniques using data mining techniques. Data Analysis has its various applications widely applied in healthcare industry. Contagious diseases are a major concern for authorities; it runs rampant in places with less medical facilities. In order to curb them they need data of these diseases. To make the strategical decision and for necessary measures the authorities use different data mining techniques to discover the knowledge from the data. In this paper, the authors have analyzed the data using statistical methods and linear regression and time series methods. The data for contagious diseases for 5 years (2011-2015) is pooled from government website. As a result of this research strategical knowledge is generated that can be used for taking the necessary measures.

Keywords *— Data Analysis, Regression, Statistical techniques, Time Series, healthcare, contagious disease.*

I. INTRODUCTION

Today the world is changing very fast. Human lifestyle has dramatically changed. Technology has taken another leap, new concepts and techniques are introduced which has been very useful. Technology is used for many purposes like education, finance, management, health etc., as illustrated in figure 1. Health sector is important for human lives. If quality of health is improved, then quality of life is also improved. To achieve that, diseases have to be controlled. But we see rise of many contagious diseases like malaria, diphtheria, etc. which affect human lives. Contagious diseases are those which spread rapidly from one person to another either by direct contact, indirect contact or droplet contact. According to NCBI, 'Infectious or communicable disease can be defined as an illness caused by another living agent, or its products, that can be spread from one person to another. An emergency condition can be defined as a state of disarray that has occurred during or after a regional conflict, or a natural disaster (i.e.: flood, earthquake, hurricane, and drought)'. Their statistics also says 'Infectious disease during an emergency condition can raise the death rate 60 times in comparison to other causes including trauma. Greater than 40% of deaths in emergency conditions occur secondary to diarrheal illness with 80% of those involving children less than 2 years of age'. There is rapid rise of these diseases worldwide. For example, according to CDC (Centre for Disease Control and Prevention), in 2016 in US, there are 9272 new cases for TB(Tuberculosis), 53850 new cases of Salmonella, 36429 cases of Lyme disease and 375 new cases of Meningococcal disease. Now according to WHO there is the new threat of Ebola erupting in Africa. In 1995 17 million of 52 million deaths were caused by infectious diseases.

To improve health, we need to reduce the spread of these diseases. There is large amount of data on these diseases. But due to lack of analysis of this data, there is little progress in preventing and reducing the contagious diseases. Here data analytics helps to analyze the data for diseases and take actions according to results. When there is sufficient data about people affected by diseases, it is collected and used with various techniques to obtain results and take actions accordingly. This paper too emphasizes on studying the contagious diseases which affect small children (under 5 yrs.) and use techniques like linear regression to analyze the data.

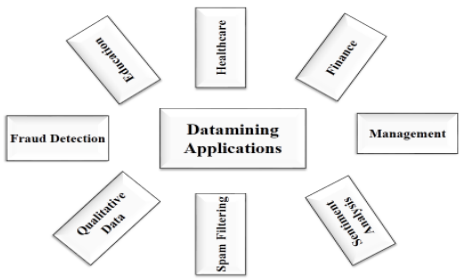

Figure 1: Applications of Data Analytics

II. RELATED WORK

Vijayashree and SrimanNarayanaIyengar [1], used data mining techniques to predict heart diseases where they have found that neural networks use supervised and unsupervised learning, decision tree algorithm uses ID3 algorithm, Naïve Bayesian classifier uses probability and genetic algorithm uses inheritance, mutation, selection and crossover for classification and prediction of heart disease. Results demonstrate that weighted associative classifier (WAC) gives more accurate results in predicting the heart disease. Umadevi and Snehapriya [2] have done a survey of various techniques and tools used for heart disease diagnosis. Their survey shows that Baysean classifier, decision tree, neural networks and support vector machines are widely used for heart disease predictions. Further they have compared the tools such as RapidMiner, Orange, Tanagra, Matlab, KNIME and their usage. Oswal and Shah [3], in their paper tell us about the various studies of data mining in health issues and its applications in health sector. They have analysed the various applications of data mining in healthcare sector especially in predictive analysis which leads to a proper medical diagnosis. It has been found by them that decision tree, Naïve Bayes and neural networks are the widely used techniques for applications like CAD, ALL (Acute lymphatic leukemia) and various biomedical applications. The research done by Shinde and Priyadarshi [4] also deals with heart disease. The diagnosis of heart disease has been done using Naïve Bayes algorithm with an accuracy of 88.33% and 86.66 on changing the testing data set. Khemphila and Boonjing[5] have used artificial neural networks (ANN) with feature selection for diagnosis of heart disease and found that feature selection is helping in increasing the accuracy level of classification and the role of Information Gain(IG) in feature selection. Princy and Thomas [6] have used KNN and ID3 algorithms for detecting the risk rates of heart disease with a higher accuracy level. Sah and Sheetalani [7] have reviewed various classification techniques and found that the ensemble method of SVM and KNN gave a better accuracy level than conventional models. Jothi et al. [8] have reviewed various data mining techniques like SVM, decision tree, k nearest neighbour etc. and has suggested that single methods doesn't suffice the need of all types of detections. So a hybrid model with the available data set need to be developed for each type of diagnosis.Sheenaland Hardik [9], in their research have used step by step approach for data mining such as classification, clustering and association. Sanati-Mehrizy et al. [10] have studied and found that data mining is predominantly used in various medical applications such as hearing aid dataset, cancer cell mining, DNA speculation etc. Durairaj and Ranjani [11] have done a comparative study of data mining techniques in healthcare applications and the study shows that there was 97.77% of accuracy for cancer predictions and around 70% for estimating the success rate of In Vitro Fertilization (IVF) treatment. Ramageri [12] has done a survey and study on basic concepts of data mining and the different techniques that can be used for business and healthcare industry. Naidu and Rajendra [13] have proposed a new algorithm called Maximal Frequent Itemset Algorithm (MAFIA) in which K means clustering is also integrated and classification done using ID3 algorithm for heart disease detection which gave 85% accuracy. Farhad, Paymen and Parvin [14] where there was 86.25% cases were predicted correct by decision tree C4.5 algorithm which has been used for detecting the success rates of delivery in pregnant women and possibilities of type of delivery. Kaushar [15] has done a study of various tools that are used for implementing data mining techniques such as R, Python etc and suggested that the tool selection is dependent on the type of application. From these research works, we observe that many researches has been done in health sector using data mining and still there are chances of improving the efficiency of methods. In our work, we have incorporated the basic statistical techniques to predict contagious diseases.

III. METHODOLOGY

In this experiment, authors collected the data from government website called ncbi.gov.in; it is where all health data is stored and updated. The experiment is done on years of 2011-2015 that is total of five years. The following is the data set:

TABLE I
Dataset of Contagious Diseases

Disease/ Year	Diphtheria	Measles	Diarrhoea	Malaria
2011	0.1	1.3	91.1	8
2012	0.1	0.8	95.8	2.7
2013	0	3.3	76.5	20.2
2014	0	3.7	89.5	3.5
2015	0	3.4	94.4	2.3
:	:	:	:	:

The experiments are implemented using R Tool. R tool is used to do programming in R language. R is a programming language. It is a software used for statistical analysis, graphical representation and reporting. It is used here for linear regression to obtain results. This language has IDE (Integrated Development Environment) and is suitable for number of languages including Python. It is free of cost and maintained under GNU open source license. It can run on any modern operating system.

IV. ALGORITHMS AND TECHNIQUES

The techniques used in this paper are descriptive analysis using mean, standard deviation, covariance and also to use linear regression and Time series regression to find the correlation between the diseases and to infer the rate of diseases in order to take preventive measures.

A. Mean

An average is the central measure of a data set. The average is measured by mean, median and mode. Mean is the sum of all numbers divided by the no of numbers given in equation 1.

$$\text{Mean} = \frac{\sum_{i=1}^{n} a_i}{n}$$

Let there be n observations a1, a2 …an

$$\text{Mean} = \frac{a_1+a_2+\cdots a_n}{n} \quad \text{(1)}$$

E.g.: From the above table,

8 2.7 20.2 3.5 2.3

$$\text{Mean} = \frac{8+2.7+20.2+3.5+2.3}{5} = 7.34$$

B. Standard Deviation

After obtaining the mean using equation 1, now we measure the variance first. Variance is mean of squared deviations. Obtaining the variance, the square root of the variance is standard deviation. The standard deviation is a very good measure of dispersion and is the one to use when the mean is used as the measure of central tendency as given in equation 2 and equation 3.

Let there be n observations a1, a2…an

From (1)

$$\bar{a} = \frac{a_1 + a_2 + \cdots a_n}{n}$$

Let $b_1 = (a_1 - \bar{a})^2$, $b_2 = (a_2 - \bar{a})^2$ and so on

$$\text{Variance} = \frac{b_1+b_2+\cdots b_n}{n} \quad \text{(2)}$$

$$\text{StandardDeviation} = \sqrt{variance} \quad \text{(3)}$$

For example- Using above data, mean = 7.34

We have deviations- 0.66 -4.64 12.86 -3.84 -5.04

Sq. of deviation- 0.4356 21.5296 165.3796 14.7456 25.4016

$$\text{Variance} = \frac{0.4356+21.5296+165.3796+14.7456+25.4016}{5} = 45.32416$$

$$\text{Standard Deviation} = \sqrt{45.32416} = 6.732$$

C. Time Series Regression

When data is in series of particular time intervals; it is called time series data. It is basically used in statistical analysis or trend analysis.

D. Covariance

Covariance is basically relation between two variables. It means that covariance measures the change between two variables. A positive covariance means both variables move in same direction while a negative covariance means both move in opposite direction.

The following steps show how covariance is calculated.The formula for covariance is as follows:

$$cov(a,b) = \frac{\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b})}{n-1} \quad \text{(4)}$$

In this formula, a represents the independent variable, b represents the dependent variable, N represents the number of data points in the sample, x- axis represents the mean of the a, and y- axis represents the mean of the dependent variable b.

E. Linear Regression

Linear regression is a statistical model to find the relationship between the variables in which one may a target and the others the predicting variables. The simple linear regression is of the form

$$Y = B0 + B1*x \quad \text{(5)}$$

where Y is the target to be predicted.

V. EXPERIMENTAL RESULTS

The analysis is based on statistical computation made using Mean, standard deviation and covariance.

The following table shows the mean and standard deviation calculated for each disease from 2011 to 2015. Both of them are calculated from equation 2 and equation 3. The mean and standard deviation are highest for diarrhoea and least for diphtheria. Diphtheria has mean 0.04 and SD IS 0.055. Measles is slightly more with mean of 2.5 and SD 1.34. Diarrhoea, the highest mean of 89.46 and SD of 7.668. Malaria has significant mean of 7.34 and SD of 7.54.

TABLE II

Mean and Standard Deviation of Disease

Year: 2011-2015		
Disease/ Year	**Mean**	**Standard Deviation**
Diphtheria	0.04	0.055
Measles	2.5	1.34
Diarrhoea	89.46	7.668
Malaria	7.34	7.54

Table III shows covariance of the years. Here difference, sum square and mean square is calculated. The results show the difference as 1 for all the years. The sum of square and mean square are same. For 2012

and 2014, 2012 has sum and mean square of 3577 and 2014 has value 195. Year 2013 has 5399 value of sum and mean square while 2015 has value 300. Similarly, year 2011 has value 6715 and 2014 has value 0, while year 2012 has 3577 and 2015 has 38 value of mean square and sum square.

TABLE III

Covariance – Difference, Sum Squares, Mean Squares

Year	DF	SUM SQ	MEAN SQ
2012	1	3577	3577
2014	1	0	0
2012:2014	1	195	195
2013	1	5399	5399
2015	1	300	300
2013:2015	1	0	0
2011	1	6715	6715
2015	1	0	0
2011:2015	1	1	1
2012	1	3577	3577
2015	1	38	38
2012:2015	1	157	157
2011	1	6715	6715
2014	1	0	0
2011:2014	1	1	1

Covariance- Covariance is used for comparison of two or more parameters. Here we use it to compare diseases and find their relationship. We take two diseases for e.g.Diarrhoea and Malaria and compare their data for all five years. As shown in figure, the parameters here are the years. The variation of the percentage of these two diseases in all these years is shown. To calculate covariance, the data should in form of dataframes.Data frames are tables or two dimensional array structure where columns are values and rows are set of values for each column. We use it to tabulate the data so it becomes convenient to study and analyze it. The columns are years and rows are diseases, the values are percentage of disease in each year. In R, to get covariance, the code requires the use of data frames because it is easier to get relationship between two variables when they are in tabulated form. Thus we create data frames.After creating data frames, it is used in code of covariance called ANCOVA. Combination of different years of the five are taken and comparisons are made. Thus we get relationship of Diarrhoea and Malaria of all five years.

The above results are graphically expressed using linear regression. The graphs are calculated using R Tool as shown in the following figure. The result shows the linear flow of the diphtheria disease. Diphtheria has lowest of percentage value compared to all other diseases. The highest was 0.1%.

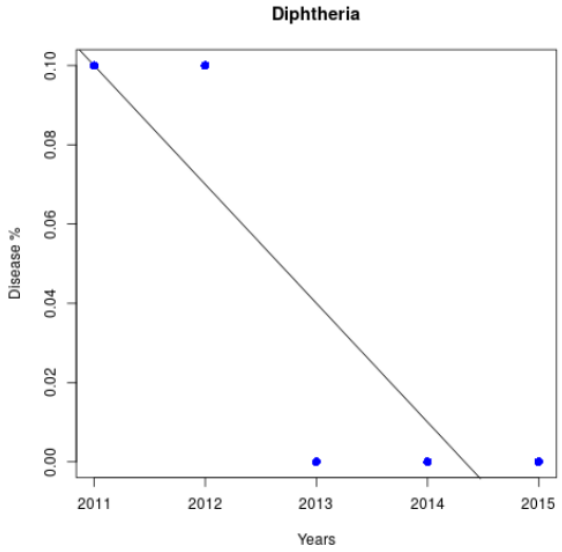

Figure 2. Linear Regression Graph of Diphtheria

Figure 3 shows the graphical analysis of Malaria diseases from the given data set the graphical results shows that there is an inclination from the year 2011 to 2015. The year 2013 shows maximum number of cases reported for malaria. Whereas year 2012 and 2015 are showing the lowest number of cases reported for malaria.Malaria shows little association to regression compared to other diseases. Except for 2013 it decreases from 2011 to 2015.

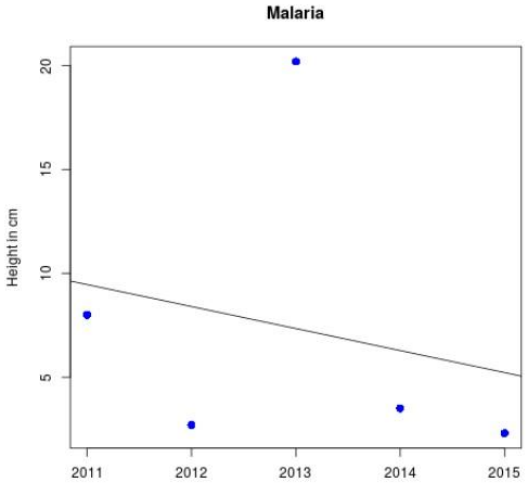

Figure 3. Linear Regression Graph of Malaria

There is a fluctuation of occurrence of measles where it got reduced in 2012 but has increased to a greater extent in 2014. Though it has started decreasing in 2015, the eradication is yet to be done which is clearly depicted in figure 4.

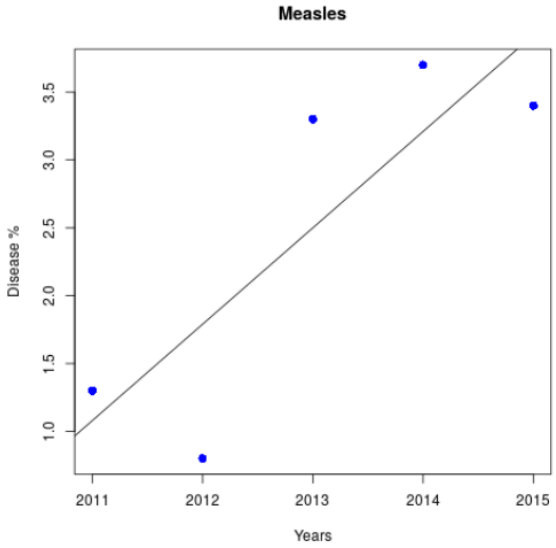

Figure 4. Linear Regression Graph of Measles

Diarrhoea is a disease of high concern when compared to all the other diseases among children. It has the highest percentage of occurrence and has increased after 2014. Figure 5 illustrates the percentage of Diarrhoea affected children against the year from 2011 to 2015. The data collected has not given any correlation between the diseases but to an extent there was an observation of occurrence of Diarrhoea in children whenever they were affected with one or more of the other three diseases. This observation has been given by the doctors with whom the consultation was taken when working on this research work. Malaria always had a correlation with Diarrhoea

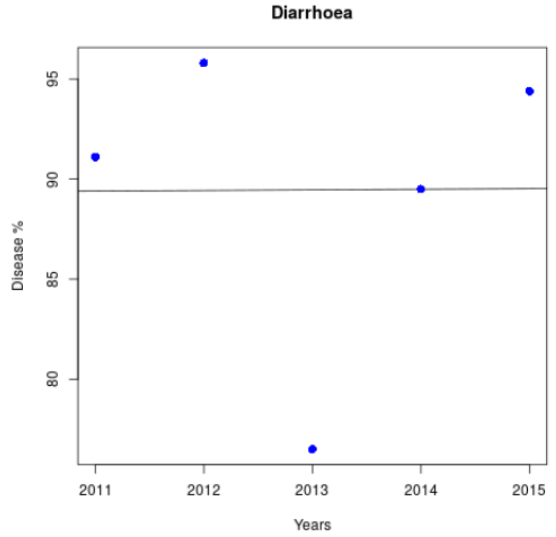

Figure 5. Linear Regression Graph of Diarrhea

The graphical representation of multi time series of the above mentioned diseases are given in figure 6 and figure 7 which describes the increase and/or decrease with respect to time which is year in this case for a period of 5 years.

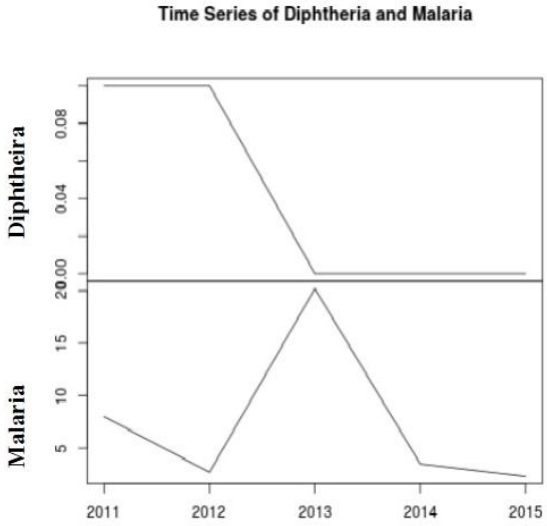

Figure 6. Time series of Diphtheria & Malaria

Diphtheria has lowest values compared to all other diseases. From 2011 to 2012 it is constant 0.1% and from 2013 to 2015 it is zero i.e. nil. Also from table 1 and 2 we can observe that Measles time series shows that first there is decrease in 2012. It is lowest in 2012 and increases from there, highest is 2014.Diarrhoea shows little randomness but its values are always high, even the lowest is 76.5. This shows that diarrhoea has affected a lot than other diseases.

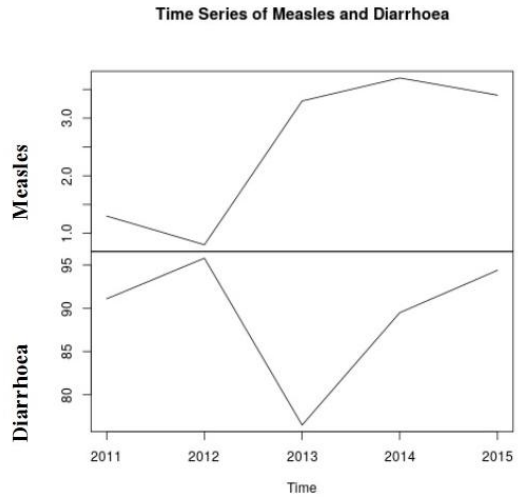

Figure 6. Time series of Measles and Diarrhea

IV. LIMITATIONS AND FUTURE SCOPE

The results obtained shows that the data is not linear, so linear regression does not work. The data is selected

only for five years from 2011 to 2015. It only emphasizes on a handful of contagious diseases pertaining to only Maharashtra. The source of the data i.e. NCBI has vast amount of data and different kinds of data like mortality rate, still born, vaccination, diseases etc. but here the data is only selected for some contagious diseases in Maharashtra. Since it is not linearly regressive, the percentage of children affected cannot be predicted by this method. So alternative means should be used for prediction of diseases.Here linear regression is used to study the disease. Other efficient methods can also be used here. The data is of only five years (2011-2015), it can be extended to 10 years also. Future prediction of 2016, 2017, 2018, and 2019 can be done. This is just a portion of data from government datasets. Similar to this, study can be done for mortality rate of new born babies, vaccination etc. This can also be extended to whole of India.

VII. CONCLUSION

The data and the results which calculates the mean and standard deviation have shown the analysis and characteristics of the diseases over the five-year period. Some disease like Diphtheria are very low which is a good sign. Diseases like Diarrhoea are highest. This indicates that Diarrhoea is the disease which affects most of the children. It has highest percentage value for affected children. So priority should be given to Diarrhoea and preventive measures should be taken for it. The other disease which shows significant rise isMalaria, in 2013 it was 20.2. Though there is big difference compared to Diarrhoea, it is still relatively higher than other diseases. Measles highest point is 3.5. So it is still within limits and it shows it is not a big threat. This is also a good sign as Measles is very low, so children are immune to it. This analysis facilitates the healthcare department to take preventive measures and curb the diseases. The government data is thus studied and given in simplified form. The study of analysing the data of diseases in the fundamental descriptive form has led to further exploration of predictive analysis using data mining techniques.

REFERENCES

[1] Vijayashree, J., and N. Ch SrimanNarayanaIyengar. "Heart disease prediction system using data mining and hybrid intelligent techniques: A review.", International Journal of Bio-Science and Bio-Technology, Vol. 8, No. 4, 2016, pp. 139-148.

[2] Dr. B. Umadevi, M. Snehapriya, "A Survey on Prediction of Heart Disease Using Data Mining Techniques", International Journal of Science and Research (IJSR), Vol. 6, No.4, 2017, pp. 2228-2238.

[3] Sangeeta Oswal, Gokul Shah, "A Study on Data Mining Techniques on Healthcare Issues and its uses and Application on Health Sector", International Journal of Engineering Science and Computing,Vol.7, No.6, 2017, pp.13536-13538.

[4] Shinde S.B , Amrit Priyadarshi , "Diagnosis of Heart Disease Using Data Mining Technique", International Journal of Science and Research (IJSR),Vol.4, No.2, 2015, pp.2301-2303.

[5] Anchana Khemphila and Veera Boonjing, "Heart disease Classification using Neural Network and Feature Selection",2011 21st International Conference on Systems Engineering IEEE, 2011, pp. 406-409.

[6] Theresa Princy. R and J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT] IEEE, 2016.

[7] Rahul Deo Sah , Dr. Jitendra Sheetalani ,"Review of Medical Disease Symptoms Prediction using Data Mining Technique", IOSR Journal of Computer Engineering, Vol. 19, No. 3, 2017,pp. 59-70.

[8] Neesha Jothi et al, "Data Mining in Healthcare –A Review", Procedia Computer Science,Vol. 72, 2015, pp. 306-313.

[9] Sheenal Patel and Hardik Patel, "Survey of Data Mining Techniques used in Healthcare Domain", International Journal of Information Sciences and Techniques (IJIST), Vol.6, No.1/2, 2016, pp.53-60.

[10] Dr. Reza Sanati-Mehrizy et al, "A Study of Application of Data Mining Algorithms in Healthcare Industry', 120th ASEE Annual Conference &Exposition: American Society for Engineering Education, 2013, Paper ID #6905.

[11] M. Durairaj, V. Ranjani, "Data Mining Applications in Healthcare Sector: A Study", International Journal of Scientific & Technology Research, Vol. 2, No. 10, 2013, pp. 29-35.

[12] Bharati M Ramageri,"DataMining Techniques and Applications", Indian Journal of Computer Science and Engineering, Vol. 1, No. 4, pp. 301-305.

[13] Mounika Naidu.P, C. Rajendra, "Detection Of Health Care using Datamining Concepts through Web", International Journal of Advanced Research in Computer Engineering & Technology, Vol. 1, No. 4, 2012, pp.45-50.

[14] Farhad Soleimanian Gharehchopogh, Peyman Mohammadi, Parvin Hakimi,"Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study', International Journal of Computer Applications, Vol. 52, No.6, 2012, pp. 21-26.

[15] Kauser Ahmed P, "Analysis of Data Mining Tools for Disease Prediction", Journal of Pharmaceutical Sciences and Research, Vol. 9, No. 10, 2017, pp. 1886-1888.