How to Cite:
Awasthi, A. K., & Sharma, M. (2022). Comparative analysis of forecasting models in healthcare (COVID-19). *International Journal of Health Sciences*, *6*(S3), 8649–8661. https://doi.org/10.53730/ijhs.v6nS3.8055

Comparative analysis of forecasting models in healthcare (COVID-19)

A. K. Awasthi
Department of Mathematics, School of Chemical Engineering & Physical Sciences, Lovely Professional University, Phagwara, Punjab, 144411, India
Email: dramitawasthi@gmail.com, amit.25155@lpu.co.in

Minakshi Sharma
Department of Mathematics, School of Chemical Engineering & Physical Sciences, Lovely Professional University, Phagwara, Punjab, 144411, India
*Corresponding author email: minakshisohareya@gmail.com

***Abstract*---**Knowledge discovery in databases (KDD) is another name of Data mining. It is an interdisciplinary area which focuses on extraction of useful knowledge from data in every sector like health, education, business etc. There are many fields to explore like business, health care, e-commerce etc but nowadays, as covid pandemic is affecting everyone and due to surge in coronavirus cases causing shortage of hospital beds, oxygen supplies, vaccine and turning away patients from hospitals, put creaky health infrastructure in spotlight. The plenty of data is available in the medical field of these conditions. To analyse the problems, there are many data mining approaches which can be used to extract useful patterns from these types of data to follow the upcoming trends. This study is to compare the various models like KNN, improved RF model and multilayer perceptron by using SPSS and python software. The data of COVID-19 has been taken from Kaggle's website which is based on the symptoms and the forecasted results has been shown. In results and conclusion, the performance of every model has shown along with this, it also shows the models and mathematical algorithms in various fields of healthcare accordingly which can be used and benefitted in medical industries.

***Keywords*---**data mining, knowledge discovery in databases, healthcare analysis, spss, python.

Introduction

Health care researchers are drowning in data but still starving for the knowledge. There is wealth of health data available but due to lack of effective analysis tool to develop the relationship between them or to find out the hidden patterns make it of no use. Traditional methods are time consuming and among all data Mining is one of the best techniques that have a good response from govt, Healthcare organisations, enterprise and many organisations. These techniques mainly used for interpreting big data, to ease the work in hospitals by helping all the employees of the hospitals to serve better treatments and facilities to the patients. With the help of these techniques' doctors can go through the stored data of patients and can draw out conclusion for the operations, OPDs, and many more works accordingly. To know the patient's future based on the patient's data stored by electronic means is one of the best features which can be used in healthcare for the wellbeing of patients. Data mining has found its role in almost all the fields of day-to-day life medical field also being no exception.

In the present prevailing situation of covid crisis internet can be effectively used for public health e.g., internet can be used to divide people into different age groups with various medical anomalies. The role of data mining is especially much more in our country where people still hesitate to get modern healthcare facilities e.g., vaccination etc. Uploading the information on a digital platform helps the government to identify the loopholes. The countries like USA, UK, European Nations have progressed a lot because all the government schemes, projects, development are all available online. Making the information and the government schemes online enables transparency, good governance and better public government relations. Data mining techniques can effectively use to address the medical problems in a country with less resources and high population density. The non-availability of specialist doctors in far flung areas can be best tackled by developing online medical consultation app in Play Store that enabling better health care facilities to general paper public.

Knowledge discovery in databases is a process to extract knowledge from big data. Data mining is a seven-step process to draw out the knowledge from big data. This consists of many approaches like clustering, classification, summarization, association, analysing variations and visualization. Through this paper, we focus on the data mining techniques which can be implemented in health sector. This study also presenting a brief literature survey and discussing various techniques for various disease. Healthcare facilities in our country are quite ambiguous and Data Mining Techniques helps in regulating these healthcare facilities by maintaining records with amazing array of patient's data. The traditional methods are not accurate and easy hence the modern technique like data mining is the need of hour for best treatments and day to day monitoring of the patients.

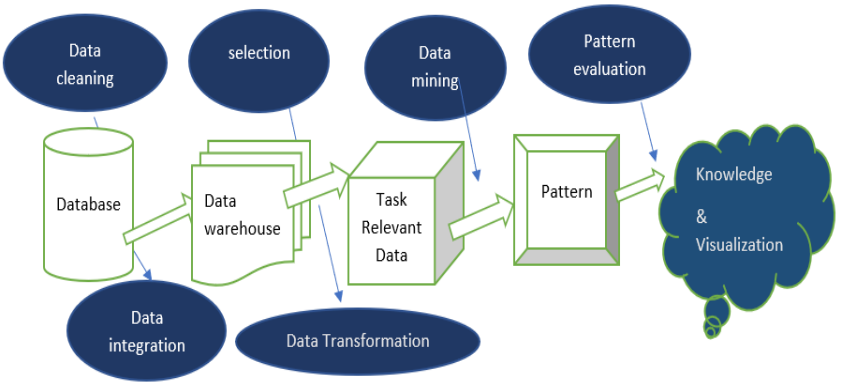
Challenges faced in healthcare

Most of the people don't have much awareness about it so the problem is how people can be benefitted from this. The main aim of this study is to provide better facilities at low costs. Insufficient man power in healthcare: The present situation shows us that how the doctors or trainers crisis occurs during this pandemic.

In2019 a survey was conducted where it is shown that India has roughly 20 health workers per 10,000 population so to meet the challenges which can be in the near future will be predicted and get ready to fight against that.

Knowledge Discovery and Data Mining

Data Mining



Healthcare with data mining

Health service is attention to the physical health of human being. According to WHO, the goals for a healthcare system is to provide good health facilities like first aid as it is the basic need of everyone, to provide health facilities in far flung areas also. Another main aim of healthcare system is to raise the ability or to meet the people's expectations regarding healthcare systems. It must ensure the fair funding facilities like the money paid by the people for providing health aids, medical facilities etc. or the revenue collected by state government or medical insurance funds etc. Nowadays there is huge amount of data in healthcare industries like patient records, infrastructure and all health resources like the equipment, various disease diagnosis, tests done in labs etc. That can make some problems such as lack of transparency, too much unnecessary care, ignorance towards patients, overburden of out patients and many more. All these can be reduced with the help of these techniques as nowadays people are in the same condition due to covid19. Early warnings may not lead to the deteriorating conditions of the hospitals, thus prediction in health care is an important task and the data mining techniques helps to discover hidden relationships in the data. The main aim is to improve the quality for better results, better treatments in low cost. With the aim of this healthcare facilities can reach in the far-flung areas also and the untimely deaths may reduce. A well-established company of medical equipment with the focused aim of hospitals and clinics increases their sales and got more return of investment. With the analysis of data fraud can reduces, patient records can be easily accessed and well treated. Several medical resources companies analysed their products for the usage in medical fields, clinics etc. and develop them according to the standards set for specific and modified treatment plans. With the advancement of the tools in healthcare

industry nurses can keep well check on individuals and responsible for day-to-day monitoring also.

Data mining models

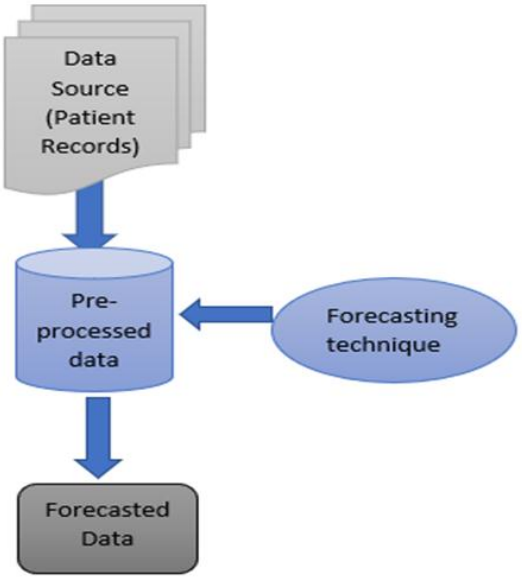
Data mining models are of two types: Predictive and Descriptive model.

- *Predictive model*: This model focuses on forecast of the individual like credit risk etc. It is the most common used method as based on the future outcomes by its prediction people can strategies their business and plan accordingly. Based on the results a static model will be prepared.
- *Descriptive Model*: The descriptive model classifies customers or prospects into groups based on the analysing relationship between the data as in Clustering, Summarization, Association rule, Sequence discovery, etc.
- *Summarization:* Summarization is of great importance in data mining. It is like the summary or conclusion the data. In data mining summarization can be done by using various methods like mean, standard deviation etc. This gives the general information of the whole data.
- *Association*: Association is an unsupervised machine learning technique. It focuses on the association or discovering the relationships between two. Based on the symptoms and diseases a relationship gets formed to diagnose the disease. In this it focuses on the frequent items which can take together like the people suffering with covid will take the medicine of fever, cough and nasal drops. Apriori, Eclat and F-P growth algorithm are the types of association algorithms.
- *Classification*: It is the process to find a model which is well suitable for the problem according to the class objects whose class label is unknown. It accurately predicts the target for each class in the data. Basically, it divides the data into training and testing data. By analysing the pattern on training data prediction is applied on testing data by preparing a model based on training data It can be used to classify loan applicant as low, medium, high risk.
- *Trend analysis*: By analysing the data that what happened in the past it suggests that what will happen in the future. Nowadays as the number of cases are reducing and hence it follows the declining or down trend.
- *Regression Analysis:* It is a statistical process for estimating the relationships between the dependent variables or one or more independent variables. In this continuous values or range of numerical values can be predicted.
- *Decision Tree:* The name shows the tree shaped structures of decisions. In this all the branches gives their results and the best one to draw out as the decision for the classification of dataset. It includes ID3, C4.5 and CART (classification and regression) Trees.

Techniques in healthcare using data mining

There are mainly two types of techniques supervised and unsupervised learning techniques. These both techniques differ only by the labeled datasets or unlabeled datasets as the supervised learning technique uses labeled input and output data

while unsupervised learning models discover themselves the unlabeled datasets. Classification is a supervised learning technique while clustering is an unsupervised learning technique.



Healthcare techniques
Artificial Neural Network

It uses brain as a basis to develop algorithm for complex patterns and prediction problems as in our brain neuron processes the information. There are specialized projections called axon which allows neuron to transmit electric and chemical signal to another cells. Neuron receive these signals via root like extensions called as dendrites. In the same way ANN has billions of processing units (input and output units) which are interconnected by nodes. These input units receive various information and neural network tried to learn about the knowledge produced as a output. There is also a back propagation in which the network works backward from output to input units to adjust weight of its connection until the difference between the actual output and desired outcome shows less error as possible. It is also applied for improving the patient's disease management. (3) it helps to select the appropriate method which can be used in health care industry for best results (4) artificial neural network technique is most common technique used in major disease area like cancer etc according to a survey of artificial intelligence applications (5).

Bayesian Classifier or Naïve Bayes

Bayesian probability interpretation as partial beliefs and Bayesian estimation calculate the validity of a proposition based on i) prior estimate of probability ii) New relevant evidence. The posterior estimation is done. It is based on Bayes theorem and bayes theorem find the probability of the individual hypothesis in the given data which can be calculated as

$$P\left(\frac{h}{D}\right) = \frac{P\left(\frac{D}{h}\right).P(h)}{P(D)}$$

Where P(h) is the prior probability of the hypothesis h. $P\left(\frac{D}{h}\right)$ is the probability of D when h is true and P (D) is the likelihood of the data. Naïve Bayes is not a single algorithm but family of algorithms where every pair of features being classified as independent of each other and it is a family of simple "Probabilistic classifiers" and it is a supervised learning algorithm. There are many naïve bayes classifiers like Gaussian naïve bayes classifier, Multinomial naïve Bayes and Bernoulli naïve bayes.

Genetic Algorithms

Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

Decision Trees

Tree shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID), CART are decision tree techniques used for classification of a dataset. In decision tree, there are three types of nodes: chance nodes, decision nodes and end nodes represented by circle, square and triangle respectively. The decision tree starts with single node and then splits into its branches and then further splits into end nodes. To do overfitting there is pruning in which decision tree removes the parts of tree which are not of much use for decision making. In this the pruning can be done by replacing its branch with leaf. There are two types of pruning: Pre-pruning and post-pruning. This pruning is done till the accuracy cannot be further improved. There are various decision tree algorithms as listed below: I) Classification and regression tree (CART) II) Iterative Dichotomiser 3 III) C4.5 (successor of ID3) IV) Chi- squared automatic interaction detector (CHAID)

- *CART*: Classification and regression tree predicts the value and at the end shows predictive outcome. It takes the predictive value based on other outcomes.
- *Iterative Dichotomiser*: It is introduced by Ross Quinlan used to take decision from data set and it is the predecessor of C4.5 which takes its output as input for making decisions.
- *C4.5 algorithm*
- C4.5 is one of the decision tree algorithms. It is also known as J48 in WEKA tool and as its decision tree can be used for the classification and that is why it is known as statistical classifier. In this the normalized attribute selection measure is known as gain ratio and which can be measured as Gain(A)/Split info(A) = Gain ratio (A) where A shows the attribute in a data set D. It is quite time efficient and can be used for building more accurate decision trees.

- *CHAID:* Chi-squared automatic interaction detector produces multiple branches of single or parent node and it is frequently used for descriptive analytics.

Random Tree Classifier

It is a machine learning algorithm and does ensemble classification. In this it takes the decision from all the branches and the highest no. of vote prediction will be the final decision tree. It is also time saving and efficient classifier.

Comparative study

In this, the case of coronavirus has been taken for the analysis by using various data mining techniques through SPSS. In this along with SPSS, PYTHON is also implemented .In this the classification technique has been applied and here I training sample, it shows the overall accuracy of 91.4% and in this only the corona result is dependent variable as it depends on symptoms and various other factors.

Classification

Sample	Observed	Predicted Negative	Positive	Percent Correct
Training	Negative	3749233	1432	100.0%
	Positive	352269	1475	0.4%
	Overall Percent	99.9%	0.1%	91.4%
Testing	Negative	1604487	655	100.0%
	Positive	151320	609	0.4%
	Overall Percent	99.9%	0.1%	91.4%

Dependent Variable: corona result

Multilayer Precptron

Number of Cases in each Cluster		
Cluster	1	5436599.000
	2	424881.000
Valid		5861480.000
Missing		.000

In the multilayer Precptron it shows the number of clusters formulated and it forms two clusters in which all the data is valid and there is no missing case.

Case Processing Summary

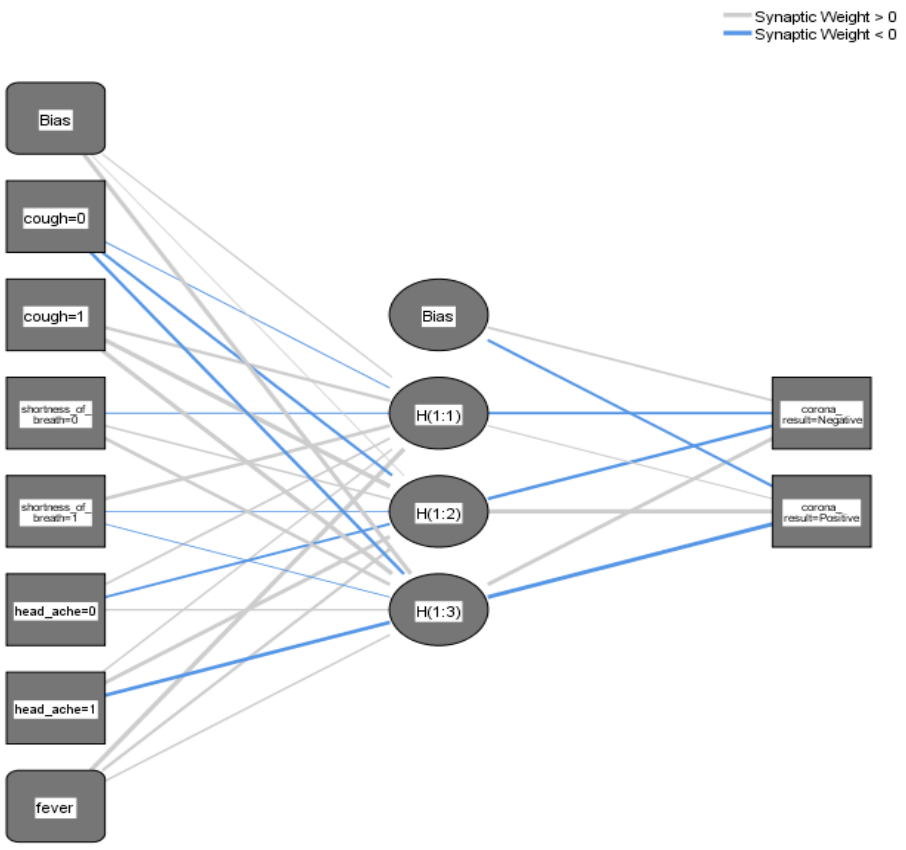
		N	Percent
Sample	Training	4104409	70.0%
	Testing	1757071	30.0%
Valid		5861480	100.0%
Excluded		0	
Total		5861480	

It divide the samples into training and testing as 70%, 30% respectively. In this the total number of cases is 5861480 and all the data is valid.On the basis of this valid data, it creats the network information as given below.There are 3 factors in input layer i.e., cough, shortness of breath and headache. Fever is covariate. In this, there is 1 hidden layer and 3 number of units in hidden layer.In output layer, it shows the corona result with cross entropy error.

Network Information

Input Layer	Factors	1	cough
		2	Shortness of breath
		3	headache
	Covariates	1	fever
	Number of Units[a]		7
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		3
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	corona result
	Number of Units		2
	Activation Function		SoftMax
	Error Function		Cross-entropy

a. Excluding the bias unit



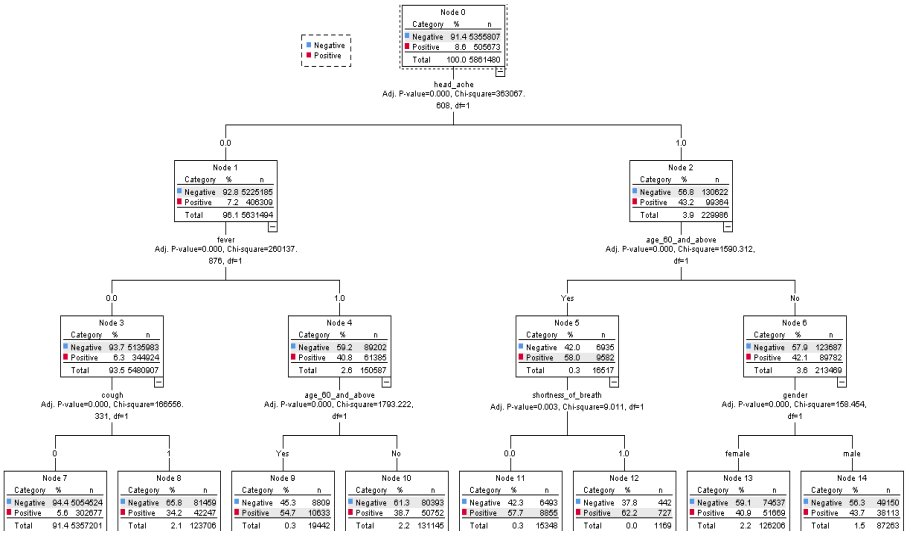
This is the representation of Network Function.

Model Summary

Training	Cross Entropy Error	1050810.172
	Percent Incorrect Predictions	8.6%
	Stopping Rule Used	Relative change in training error criterion (.0001) achieved
	Training Time	0:00:27.57
Testing	Cross Entropy Error	451140.951
	Percent Incorrect Predictions	8.6%

Dependent Variable: corona result

Random forest tree



The random forest classifier classifies the data on the basis of symptoms and after all the classification ,it selects the terminal nodes. It forms positive ans negative corona patients .From the classification it is shown that 99.7% of the accuracy is shown in negative cases and 4% accuracy for positive cases on the basis of this data.

Classification

		Predicted	
Observed	Negative	Positive	Percent Correct
Negative	5340063	15744	99.7%
Positive	485458	20215	4.0%
Overall Percentage	99.4%	0.6%	91.4%

Growing Method: CHAID

Dependent Variable: corona_result

Gain Summary for Nodes

Node	N	Percent	Mean
0	3361	100.0%	370.566

Growing Method: CHAID
Dependent Variable: recovered

The terminal nodes in the model has been shown in Gain summary for Nodes In this, it shows that it stops only after 100% classification and the mean is 370.566

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	corona result
	Independent Variables	test date, cough, sore throat, fever, shortness of breath, age_60_and_above, headache, gender
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	headache, fever, cough, age_60_and_above, shortness of breath, gender
	Number of Nodes	15
	Number of Terminal Nodes	8
	Depth	3

In this, the method developed and exceuted is CHAID. As the corona result is dependent on other variables like test date, cough, sore throat, fever, shortness of breath,age etc and hence it is said as dependent variable. The minimum cases has been taken for parent node is 100 and for child node is 50 .Its terminal nodes becomes 8 and the depth of these node is 3. By increasing more number of nodes, it increases the accuracy of Random Forest model.

Table 1
Comparative study of different prediction in healthcare

S. No	Techniques used	Algorithm	Field
1	Decision Tree	C4.5	Diabetes
2	Naïve Bayes	Classification	Heart Disease
3	Neural Network	Back Propagation	Eye disease
4	Decision tree (SPSS)	Chi square	Diabetics
5	SVM	-	Stroke mortality in brain
6	Neural networks	Trend analysis	Heart disease
7	Decision tree induction method	C5	Breast cancer
8	Artificial Neural Network	Not discussed	Breast cancer
9	Logistic Regression	Not discussed	Breast cancer
10	Artificial Neural network	Not discussed	Chest Disease
11	K- Nearest Neighbour	Classification	Diabetes, Cancer
12	SVM classifier	classification	Diabetes
13	Apriori algorithm	Prediction	Chronic Disease

Conclusion

In recent times there has been sharp hike in amount of medical data in the field of medical care which is gathered by various electronic means and this along with

rise in availability of economical and dependable computing equipment has encouraged many researchers to start exploring these collected data. However, despite all available data our healthcare system is still crumbling which shows that these data have not been properly used. This paper explores the application of data mining techniques using medical data effectively for disease prediction which can be diagnosis at proper time to enable good human health. It is observed that some of the techniques of data mining has already been employed for medical data and some can be in the coming time for better results. Huge amount of data available in the medical field has made it mandatory to use data mining techniques for arriving at a decision and prediction in the field of medical care such as to identify the kind of disease., availability of various medical facilities at different medical centres. More over the J48, decision trees in Weka shows the wonderful results in the healthcare sector. This paper checks various techniques in contrast with others it is seen that random forest classifier shows effective results. These techniques can be applied in various healthcare subfields also to predict and allow to prepare better for the upcoming situation.

References

1. Won J B. & Kyung Y (2020) Context Deep Neural Network Model for Predicting Depression risk Using Multiple Regression in *special section on Machine learning Designs implementations and Techniques IEEE access* Vol 8.
2. Khan A F, Mabrouk R, Daya A & Ahmad S B (2021) Detection and Prediction of Diabetes using Data Mining: A comprehensive Review in *IEEE Access* vol 9.
3. Dursun D, Gelen W, Mit K, (2005) "Predicting breast cancer survivability: a comparison of three data mining methods" *Artificial intelligence in Medicine* 34 pp 113-127.
4. Luo et al, 2017 hospital daily outpatient visits forecasting using combinatorial model based on Arima and Ses models, BMC health services Research,17:469
5. Deloitte,2018, Global health care outlook: The evolution of smart health care *Iranian journal of science and technology (IJST)* pp16-21
6. Jiang F, Jiang Y, Zhu H, Dong Y, Li H, Ma 2017Artificial intelligence in healthcare: past present and future Stroke Vasco Neurol. 2(4) 230–43.
7. Sheenal P, Hardik P,2016 survey of data mining techniques used in healthcare domain *international journal of information science and techniques (IJIST)*Vol.6, No. 1/ 2
8. Palaniappan S, and Awang R (2008, August) Intelligent Heart Disease Prediction System Using Data Mining Techniques *IJCSNS International Journal of Computer Science and Network Security* Vol. 8(8).
9. Wauls, Y. E., H. W. Ittmann, and L. Hanmer.2015 Decision support systems in health care *international research journal of engineering and technology (IRJET)*39.
10. Srinivasan B, & Pavya K. (2016) A study on data mining prediction techniques in healthcare sector *International Research Journal of Engineering and Technology (IRJET),* Vol 3 pp. 72-75.
11. Harshit k & Nishant S 2017 Review paper on big data in Health care informatics, care *international research journal of engineering and technology* (IRJET).

12. Yumusakc N & Temurtas F 2010 chest diseases diagnosis using artificial neural networks in expert Systems with Applications- *Elsevier*Volume: 37.
13. Tang H P, &Tseng H M. 2009 Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN Classification *IEEE access* pp 1-6
14. Kathyayini, R. &Jayaprakash, J. 2005 Association technique on prediction of chronic diseases using apriori algorithm *International Journal of Innovative Research in Science, Engineering and Technology,* vol 04, issue 06.
15. Balakrishnan S, & Narayanaswamy R. (2009) feature selection using FCBF in type ii diabetes databases *special issue of the International Journal of the Computer the Internet and Management* vol 17.