DIAGNOSIS AND PROGNOSIS BREAST CANCER USING CLASSIFICATION RULES

Miss Jahanvi Joshi Mr. RinalDoshiDr. Jigar Patel
jahanvijoshib@gmail.comrinalhdoshi@gmail.com drjigarvpatel@gmail.com

Abstract— Breast Cancer is highly heterogeneous disease. Breast Cancer Diagnosis and Prognosis are two medical challenges to the researchers in the field of clinical research. Breast self-exam and mammography can help find early diagnosis of breast cancer. This is possible when in some situation or stage the treatment is possible. Treatment may consist of radiation, lumpectomy, and mastectomy and hormone therapy. The origin of this research for diagnosis a breast cancer depends upon a lump in the breast, a change in size or shape of the breast or a nipple. Men can have breast cancer, too, but the number of cases is small. The purpose of this research is to develop a novel prototype of clinical problem regarding to diagnose and manage patients with breast cancer. The primary dataset of breast cancer is carried out from UCI dataset repository for the purpose of experimental work. These experimental works justify the problem formulation of the clinical research using different classification technique.\

Keywords— Breast Cancer, Clinical Problem, Classification Rules, Data Mining, Health Care, Web Mining,Weka.

INTRODUCTION

Breast Cancer becomes dangerous disease in today's era. The most common type of this type of breast cancer is ductal carcinoma, which begins in the lining of the milk ducts. It is nothing but only thin tubes that carry milk from the lobules of the breast to the little nipple. Another type of breast cancer is lobular carcinoma, which begins in the lobules of the breast. Invasive breast cancer is breast cancer that has spread from where it began in the breast ducts or lobules to surrounding normal tissue. Breast cancer occurs in both men and women, although male breast cancer is rare.

According to the survey of United States in 2014, there are 232,670 females and 2,360 males having this type of new cases regarding the breast cancer. Among them 40,000 females and 430 males was death during the period this survey [1]. This survey is origin and motivation for our research work.

Early burning signs of breast cancer may absorb the detection of a new lump or a change in the breast skin. These are the signs and symptoms for the early detection of the breast cancer. By performing monthly breast self-exams, patient will be able to more easily identify any changes in her breast. If patient found abnormal changes in her breast she gives a path to contact healthcare experts. In some situations women are encourage for an excitement like breast sensitivity issue, breast examination by doctors and measurement with tailor.

Data Mining is a powerful tool and technique to handling this task. In data mining breast cancer research has been one of the important research topics in medical science during the recent years The classification of Breast Cancer data can be useful to predict the result of some diseases or discover the genetic behavior of tumors. There are many techniques to predict and classification breast cancer pattern. This paper empirically compares performance of different classification rules that are suitable for direct interpretability of their results.

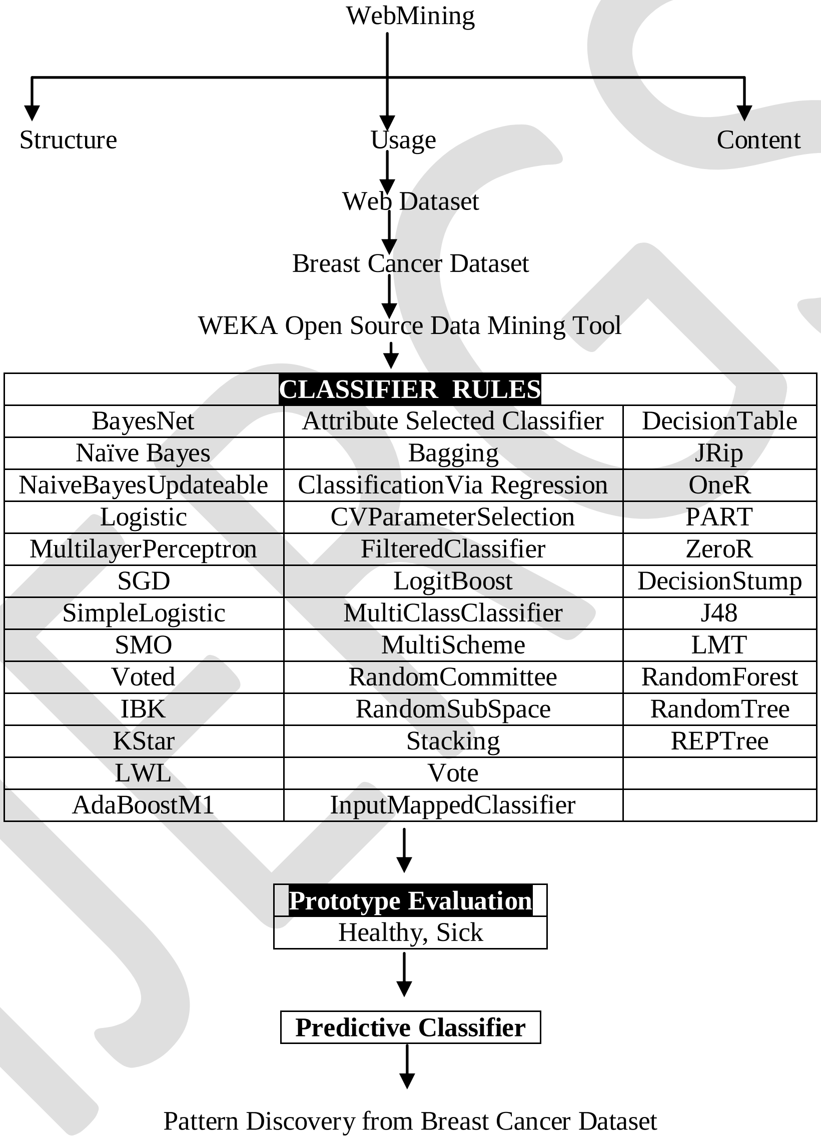
LITERATURE REVIEW

Author of this paper Tired to improve the website design, optimize website structure and built intelligent website. To achieve these targeted objectives, the author used machine learning approaches and for the user identification algorithm, session identification algorithm and Apriori algorithm applied on the preprocessed dataset. These processes worked on WEKA open source Data Mining s/w tool. The resulted outcome arrives on the bases of Apriori algorithm. The author gives the challenge that must apply on other. Data mining algorithm and the comparative analysis gives more optimum outcome [2]. Author of this paper Compare the performance of the supervised algorithm Naïve Bayesian, support vector machine, Radical basis neural network, decision tree, j48 and simple CART. The proposed work conducted to discover the batter –quality classifier for disease for detection. That is processed in WEKA with the dataset WBC, WDBC, pama diabetes and Brest tissue. The outcome of the result showed that the SVM RBF kernel responded well than the others. The author concludes with to achieve the potential outcome with accuracy well as the complexity will be calculated for feature expansion [3]. Author of this paper performed comparative study of classification method with different data set- PIMA Indian Diabetes, state Log Heart Disease, BUPA Liver –disorders, and Wisconsin Brest Cancer. This work target to study the performance to achieve higher accuracy level with lower error rate. There experimental findings used 10 fold cross validation method. The outcome by SVM method indicated promising level of accuracy level of 96.74% for PIMA Indian diabetes dataset and 99.25% with statLog heart Disease data set. They have used c4.5 decision tree technique with BAPA Liver-disorders dataset with an accuracy level of79.71% there ultimate finding with user techniques-Bayes Net, SVM, kNN and RBF-NN with the dataset Wisconsin Brest Cancer Data set have indicated as to combine multiple technique with different parameter [4]. Author of this paper explored the comparative performance of the classification and clustering algorithm using heart disease dataset. The work targeted to higher level of prediction accuracy by comparing both the techniques. The evaluation carried on the performance of classifiers of Bayes (Naïve Bayes, Naïve Bayes updateable), functions (SMO),Lazy (IB1,IBK),Meta Multi BoostAB, Multiclass Classifier),Rule(Decision Table),tees (NB Tree)and the clustering algorithm of EM, Cobweb, Father First, Make Density Based Cluster ,Simple K-means' algorithm. The analyses lead to conclusion which state that the NB tree having higher prediction Accuracy compared to the clustering algorithm [5]. Author of this paper focus on the work statistical and data mining tool and technique for diagnosis disease. To achieve the more accurate outcome of data mining techniques they applied hybridization on the selected method used which illustrate the acceptable levels of accuracy then for enhance the accuracy of disease the hybridization data mining techniques. Hybrid data mining techniques produces more effective outcome in diagnosis of heart disease. Different hybrid technique like fuzzy artificial immune recognition system and K-nearest neighbor are applied together are applied tougher which produced accuracy of 87%.Neural network gives accuracy of 89.01% which is batter. One case of neural network and genetic algorithm is also discussed which produced better result in deter mining heart disease [6]. Author of this paper constructed the work to discover the effectiveness of preprocessing algorithm on dataset to investigate. The research conducted data mining algorithm on dataset on the Z –AlizadehSani dataset which is used to achieve more accurate results. The cost –sensitive algorithm are used along with base classifiers on naïve Bayes, sequential minimal optimization (SMO), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and C4.5 were a closed work with the SMO algorithm has to very high sensitivity (97.22%) and accuracy (92.09%)rates. AS a final point, they said that the proposed cost sensitive algorithm can be used on other diseases such as cancer [7]. Author of this paper focus on compared the performance of the classification technique like Sequential Minimal Optimization (SMO), IBK, BF Tree. The proposed work conducted to find out the accuracy classifier for breast cancer detection. that process in weka with the data set UCI machine learning in all over technique sequential minimal Optimization (SMO) achieve better outcome of the result with accuracy ,low error rate , and performance[8]. Author of this paper explore performance of classification technique using Wisconsin prognostic breast cancer (WPBC) data from UCI machine learning technique .the work targeted to high level of prediction accuracy by compare classification different technique.

the evaluation carried on the performance of classification like Binary Logistic Regression (BLR), C4.5 decision tree algorithm ,Partial Least Squares for Classification (C-PLS), Classification Tree(C-RT), Cost-Sensitive Classification Tree(CS-CRT), Cost-sensitive Decision Tree algorithm(CS-MC4), SVM for classification(C-SVC), Iterative Dichomotiser(ID3), K-Nearest Neighbor(K-NN), Linear Discriminant Analysis (LDA), Logistic Regression, Multilayer Perceptron(MP), Multinomial Logistic Regression(MLR), Naïve Bayes Continuous(NBC), Partial Least Squares -Discriminant/Linear Discriminant Analysis(PLS-DA/LDA), Prototype-Nearest Neighbor(P-NN), Radial Basis Function (RBF), Random Tree (Rnd Tree), Support Vector Machine(SVM) the analysis lad to conclusion random tree and Quinlan's C4.5 having 100% accuracy in classifying the Wisconsin prognostic Brest cancer data set [9].Author focus on performance of supervised learning algorithm viz decision tree, random tree ,ID3 ,CART,C4.5 ,and Naive Bayes. That processed in TANAGARA with Wisconsin breast cancer data set used for modeling breast cancer data. The result of random tree fives batter optimum result with highest accuracy rate [10].Author of this paper investigation on data mining three technique first one is Naïve Bayes, second one is back-propagated neural network, and last one C4.5 decision tree algorithms. For this processed use WEKA. Open source tool with data set SEER. The outcome of the result showed C4.5 better performance then other two techniques [11]. Author of this paper analyzed Brest cancer using data mining technique classification. They use machine learning technique like Decision Tree (C4.5), Artificial Neural Network and support vector machine for predicting a breast cancer. That process in WEKA tool kit with the Iranian center of breast cancer .this work targeted to analyzed the performance to achieve higher accuracy specificity and sensitivity result by SVM technique indicated promising level of accuracy level of 95.7%,97.1% and 94.5%[12].Author of this paper prediction on data mining technique on heart disease , Diabetes, Breast Cancer in heart disease data collected data from a hospital information system the heart disease prediction in that machine learning algorithms namely naïve bayes, K-NN, Decision List. In all technique classification accuracy of the naïve bayes algorithm is better when compared to other algorithm. Second one breast cancer as per survey of united state prediction data mining technique like C4.5, ANN and Fuzzy decision tree. ANN conducts better accuracy and good performance. Third one about diabetes as per base on the American diabetes association perdition by using homogeneity based algorithm genetic algorithm predicts batter accuracy. For feature work enhance they predates diff type of disease prediction using data mining technique [13]. Author of this paper focus on use three different type of machine learning technique for predicting of breast cancer. They use Iranian canter for breast cancer (ICBC) data set and implement machine learning technique like decision tree (C4.5) Support vector machine (SVM) and Artificial Neural network (ANN) compared the performance of technique and find sensitivity, specificity , accuracy. As per a conclusion SVM provide better performance with highest accuracy rate [14]. Author of this paper Prediction on breast cancer and heart disease with dataset Public Use data. They consisting of 909 Record for Heart disease and 699 for breast cancer. They use two type algorithms c4.5 and c5.0.as per a conclusion c4.5 get better result all over technique [15]. Author of this paper compared the performance of classification algorithms- Decision tree, Naïve Bayes, MLP, Logistic Regression SVM, KNN. That is process in WEKA with data set. The outcome of the result showed that the SVM responded well then the other. The author conduct with to achieve the potential outcomes with accuracy as well as the complexity will be calculated for future expansion [16]. Author of this paper focus on prediction of breast cancer using data mining technique.This process in WEKA data mining tool kit with data set SEER.They investigation on three data mining techniques like Naïve Bayes, the back-propagated neural network, and the C4.5Decision tree algorithms. The outcome of the result c4.5 algorithms shows much better performance then the other two techniques [17]. Author of this paper focus on data mining application on medical research for a predicting and discovering pattern base on detected symptom on health condition for process take a mammography, dermatology, orthopedic thyroids for data pre-processing execute classification for clinical test data lode test data for verification for classifier malady classification. They support decision tree generate by the quinlan's algorithm is smaller then the decision tree by the random tree classification technique [18]. Authors diagnosis breast cancer using clustering data mining techniques [19]. Authors of the paper develop pattern knowledge discovery framework using data mining technique. This framework is generic which is relate to different services for users [20].

PROPOSED WORK

The main area of the research work is web mining. Web mining has three categories – content, structure and usage. We use web usage mining for finding hidden pattern from the breast cancer dataset. The suggested model of the proposed work is followed.



CLASSIFIER RULES		
BayesNet	Attribute Selected Classifier	DecisionTable
Naïve Bayes	Bagging	JRip
NaiveBayesUpdateable	ClassificationVia Regression	OneR
Logistic	CVParameterSelection	PART
MultilayerPerceptron	FilteredClassifier	ZeroR
SGD	LogitBoost	DecisionStump
SimpleLogistic	MultiClassClassifier	J48
SMO	MultiScheme	LMT
Voted	RandomCommittee	RandomForest
IBK	RandomSubSpace	RandomTree
KStar	Stacking	REPTree
LWL	Vote	
AdaBoostM1	InputMappedClassifier	

Figure 1: Prototype for Breast Cancer Pattern Discovery (PBCPD)

In Fig. 1 we discuss novel prototype for breast cancer detection. Here Main work is start with Web Mining which has three dimensions- content, usage and structure. In this prototyping we select usage mining. In Web Usage Mining data is retrieved from web dataset. We select breast cancer dataset in Weka Open Source environment. In Weka we use 37 classification rules for diagnosis and prognosis breast cancer among patients. By Experimental Analysis we can get predictable pattern

IMPLEMENTATION WORK:

To classify breast cancer data set with high accuracy and efficiency different classifier rules are used for finding healthy patients. In this research WEKA open source mining tool is used for modeling breast cancer data. This tool proposes several classification rules from exploratory data analysis, statistical learning and machine learning.

For the purpose of the implantation work the data is taken from UCI for the purpose of to solve the research objective. To perform experimental work this research take WEKA as an open source data mining tool and then apply different classification algorithm for diagnosis and prognosis patients. **The dataset attributes descriptions are as under in Table 1:**

Table 1: BREST CANCER DATASET ATTRIBUTE

Attribute Name	Description
Age	Patient’s Age in years
Menopause	the period in a woman's life when menstruation ceases
Tumor-size	Patient’s tumor-size on her breast
inv-nodes	Node size in main portion of the breast.
Node-caps	Node is present or not in cap of the breast
Deg-malig	Stage of breast cancer
Brest	Left breast or Right breast or both breast
Breast-quad	Portion of the breast for example left-up, left-low, right-up, right-low, central.
Irradiate	Present or not (YES/NO)
Class	no-recurrence-events, recurrence-events (Reduce the risk of breast cancer)

Table 2: BREST CANCER DATASET CLASS

Class name	Description
Diagnosis	sick, healthy or (unpredictable) no class

For this research we use different 37 classification algorithm for the purpose of diagnosis of healthy and sick patients. The result of these classification rules are tabulated in table 3.

Table 3:RESULT ANALYSIS OF BREAST CANCER PATIENT

Clustering Technique	Healthy	Sick
BayesNet	76	24
NaiveBayes	75	25
NaiveBayesUpdateable	75	25
Logistic	76	24
MultilayerPerceptron	76	24
SGD	76	24
SimpleLogistic	76	24
SMO	76	24
Voted	74	26
IBK	98	2
KStar	98	2
LWL	79	21
AdaBoostM1	76	24
Attribute Selected Classifier	76	24
Bagging	81	19
ClassificationVia Regression	76	24
CVParameterSelection	70	30
FilteredClassifier	76	24
LogitBoost	78	22
MultiClassClassifier	76	24
MultiScheme	70	30
RandomCommittee	98	2
RandomSubSpace	77	23

Stacking	70	30
Vote	70	30
InputMappedClassifier	70	30
DecisionTable	71	29
JRip	77	23
OneR	73	27
PART	80	20
ZeroR	70	30
DecisionStump	72	28
J48	76	24
LMT	76	24
RandomForest	98	2
RandomTree	98	2
REPTree	74	26

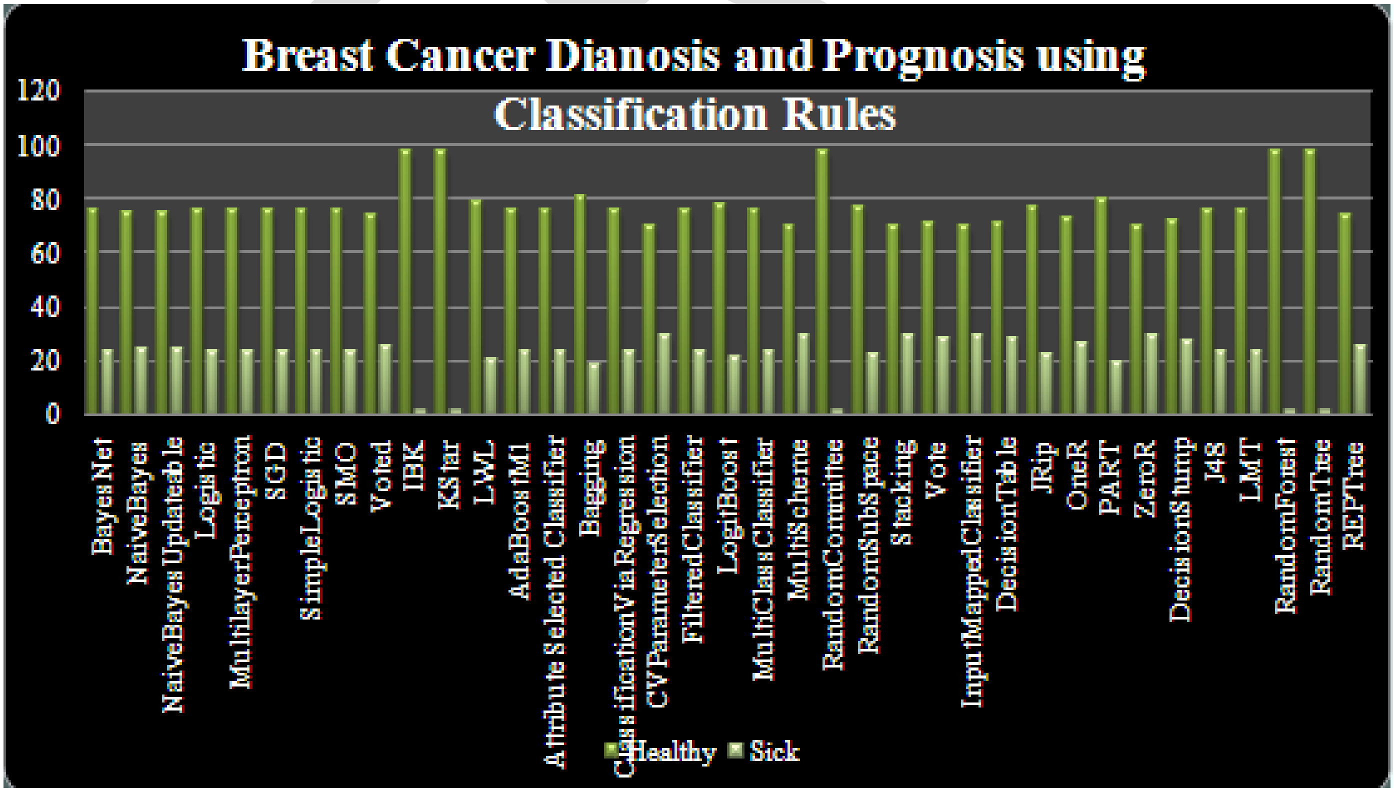


Figure 2: CLASSIFIER PERFORMANCE ANALYSIS

By this result analysis BayesNet , Logistic, MultilayerPerceptron ,SGD, SimpleLogistic, SMO, AdaBoostM1, Attribute Selected , ClassificationVia Regression, FilteredClassifier, MultiClassClassifier Classifier, J48, LMT classifier gives more accurate result. According to these classifier rules these research diagnosis 76% healthy and 24% sick patients.

CONCLUSION

Because In this paper various classification rules are compared to predict the best classifier. We develop new prototype for diagnosis predictable pattern discovery of breast cancer. Experimental results show the effectiveness of the proposed method. The base for this is knowledge discovery and data mining. The classifier is identified to determine the nature of the disease which is highly important for finding healthy breast cancer patients. By this work is useful to uncover patterns hidden in the data that can help the clinicians and doctors in decision making.

FUTURE WORK

For this research we use different classifier rules for diagnosis healthy patients. To study and experimental work we use weka open source data mining tool. This research is extending by using different clustering, statistical model and machine learning algorithm. In future we use tanagara and orange data mining tool. We can make generic prototyping for different domains like ecommerce, electricity or many areas.

REFERENCES:

[1] "Breast",from http://www.cancer.gov/cancertopics/types/breast access on [02-09-2014]
[2] "The research and application of web log mining based on the platform weka",Xiu-yuZhong,Scince Direct Elsevier 2011.
[3] "AN EMPIRICAL COMPARISON OF SUPERVISED LEARNING ALGORITHMS IN DISEASE DETECTION", S. Aruna, Dr S.P. Rajagopalan and L.V. Nandakishore IJITCS, August 2011.
[4] "PERFORMANCE ANALYSIS OF VARIOUS DATA MINING CLASSIFICATION TECHNIQUES ON HEALTHCARE DATA",Shelly Gupta, Dharminder Kumar and Anand Sharma, International Journal of Computer Science & Information Technology (IJCSIT) , August 2011
[5] "Improving the Performance of Data Mining Algorithms in Health Care Data",P. Santhi, V. MuraliBhaskaran, IJCST ,September 2011
[6] "Using data mining techniques in heart disease diagnosis and treatment", Shouman, M. Turner, T. ; Stocker, R.,IEEE 2012.
[7] "Diagnosis of Coronary Artery Disease Using Cost-Sensitive AlgorithmsHosseini,M.J. ,Sani,Z.A. Ghandeharioun,A. 2012IEEE.
[8] "A Novel Approach for Breast Cancer Detection using Data Mining Techniques" VikasChaurasia, SaurabhPal IJIRCCE Vol. 2, Issue 1, January 2014.
[9] "Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining Techniques" ShomonaGracia Jacob, R. GeethaRamani Proceedings of the World Congress on Engineering and Computer Science Vol I October 2012.
[10] "Application of Data Mining Techniques to Model Breast Cancer Data S. Syed Shajahaan ",S. Shanthi , V. ManoChitra IJETAE Volume 3, November 2013.
[11] "Predicting Breast Cancer Survivability Using Data Mining Techniques" AbdelghaniBellaachia,ErhanGuven http://www.siam.org/meetings/sdm06/workproceed/Scientific%20Datasets/bellaachia.pdf?q=data-mining- techniques
[12] "Using the Data Mining Techniques for Breast Cancer Early Prediction" Samar Al-Qarzaie,Sara Al-Odhaibi, BedoorAl-Saeed, and Dr.MohammedzAl-Hageryhttp://www.psu.edu.sa/megdam/sdma/ Downloads/Posters/Poster%2011.pdf
[13] "Disease Prediction in Data Mining Technique – A Survey",S.VijiyaraniS.Sudha,International Journal of Computer Applications & Information Technology January 2013
[14] "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence" AbbasToloieEshlaghy, Ali Poorebrahimi, MandanaEbrahimi, Amir R. Razavi and Leila GhasemAhmad,OMICS 2013
[15] "A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining", MohammadTaha Khan, Dr. ShamimulQamar and Laurent F. Massin, International Journal of AppliedEngineering Research 2012.

[16] "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques",G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani, International Journal of Innovations in Engineering and Technology (IJIET) August 2013.

[17] "Predicting Breast Cancer Survivability Using Data Mining Techniques", AbdelghaniBellaachia, ErhanGuven.

[18] ShomonaGracia Jacob, R. GeethaRamani "Mining of Classification patterns in clinical data through data mining algo" access from IEEE.

[19] Jahanvi Joshi, RinalDoshi and Jigar Patel. Article: Diagnosis of Breast Cancer using Clustering Data Mining Approach.International Journal of Computer Applications 101(10):13-17, September 2014.

[20] Mr. RINAL H. DOSHI, Dr. HARSHAD B. BHADKA and Ms. RICHA MEHTA, 2013. DEVELOPMENT OF PATTERN KNOWLEDGE DISCOVERY FRAMEWORK USING CLUSTERING DATA MINING ALGORITHM.International Journal of Computer Engineering & Technology (IJCET).Volume:4, Issue: 3, Pages: 101-112