IJCTA, 9(27), 2016, pp. 359-365 ISSN: 0974-5572
© International Science Press

Prediction of Breast Cancer through Classification Algorithms: A Survey

Samuel Giftson Durai*, S. Hari Ganesh and A. Joy Christy*****

ABSTRACT

One of the second leading causes of death for women is the breast cancer, which acts as serious threat to the woman society as the change of climate, urbanization and adoption of fast food culture that increases the occurrence of breast cancer in modern age. The techniques that are available in data mining provide a meaningful contribution to the field of medical diagnostics for the accurate prediction of the disease.Thepredictive analysis techniques of data mining build a knowledge prediction model by analyzing the present history of patients so as to analyze the future data. This paper reviews the results of various classification techniques that have been demonstrated in the research articles from the year 2012 -2016 and makes a comparison between the presented outputs of the earlier works. Moreover, experimentation is also conducted to verify the authenticity of the best clustering algorithm as a proof of concept.

Keywords: classification, breast cancer, decision tree, knowledge prediction model, predictive analysis

1. INTRODUCTION

Breast cancer, is a peculiar type of cancer that mostly affects women than any other human beings which is caused two major factors called modifiable or non-modifiable. Modifiable factors are those that can be controlled like habits and environmental issues. Non-modifiable factors are that cannot be controlled like gender and family history [1]. According to a survey, 1 out of 28 women in India is prone to breast cancer as the early detection techniques on the presence of breast cancer are still lacking in the exact prediction of disease. Moreover, the lack of awareness, proactive measures and treatment facilities increases the risks of survival. Early detection of the syndromes may direct to overcome the breast cancer through appropriate treatment.

Many scientific researches have been proposed for the automatic prediction of breast cancer. Data mining is yet another scientific approach that encompasses numerous techniques and algorithms for analyzing the hidden knowledge from large data sources [2]. Classification techniques of data mining acts a predictor of final results by processing the class labels of a dataset. Hence, the classification techniques can be widely accepted in medical domain for building knowledge prediction model for identifying the occurrence of disease.

The objective of this paper is to present a summary on the performance of existing classification techniquesso as to identify the research problems that need to be focused for future. Moreover, this paper also intended to conduct a comparative study of the classification techniques that highlight upon the best classification algorithms that accurately predicts the malignant of breast cancer dataset.

2. REVIEW OF LITERATURE

The review of literature presents a brief description of the existing researches on data mining techniques with respect to breast cancer. The narration of the algorithms used along with their scope and limitation is presented in the review.

* Lecturer, Dept of IT, Higher College of Technology, Muscat, Oman.

** AssistantProfessor, Dept. of Computer Science, H.H. The Rajah's College Pudukottai-622 001

*** Research Scholar, Dept. of Computer Science, Bishop Heber College, Trichy-620 017.

Chaurasia et al. [3] have investigated the performance of BFTree (Best First Tree), IBK (K-nearest neighbor classifier) and SMO (Sequential Minimal Optimization) classification techniques on breast cancer data. The authors have conducted the experiment in Weka data mining tool and have taken three evaluation criteria such as time, correctly classified instances and accuracy for assessing the superiority of each algorithm. The authors have stated that the performance of SMO algorithms has been better than the other two algorithms in terms of accuracy and low error rate. The authors have also identified the most important features for enhancing the prediction accuracy.

Shajahaan et al. [4] have explored the applicability of decision trees for breast cancer prediction and also analyzed performance of conventional supervised learning algorithms such as CART, ID3, C4.5 and Naïve Bayes. The experiment is conducted throughWekatool. The authors have highlighted five meaningful attributes that can be considered for the prediction. The authors have concluded that the random tree serves as the best classification algorithm for breast cancer with higher accuracy in prediction.

Shrivastava et al. [5] have used classification techniques for classifying the benign or malignant instances of breast cancer dataset. The authors have created a decision tree classifier model for the classification. The authors have used Weka for the experimentation for simplifying the prediction task. The authors have stated that most of the breast cancer analyses have done only through neural network and decision tree approaches. Hence the authors have also taken decision tree model using if-then rules for enhancing the performance of decision trees.

Senturk et al. [6] have analyzed the performance of seven classification prediction models such as Discriminant Analysis (DA), Artificial Neural Networks (ANN), Decision Trees (DT), Logistic Regression (LR), Support Vector Machines (SVM), Naïve Bayes (NB) and K-nearest neighbor (KNN) for the early diagnosis of breast cancer through RapidMiner Tool. The authors have claimed that the SVM algorithm has outperforms than other six algorithms in the prediction of the existence and non-existence of the disease with 96%.

Theinet al. [7] have proposed an approach for distinguishing the classes of breast cancer through neural network. The authors have overcome the local optima issue of neural network differential evolution algorithm for determining the optimal value or near optimal value for ANN parameters. To overcome the issue longer training time and lower classification, the differential evolution algorithm is further collaborated with island based model. The island based model improves the accuracy and takes less training time by making an analysis between two different migration topologies.

Venkatesan et al. [8] have analyzed the breast cancer data using four classification algorithms namely j48, Classification and Regression Trees (CART), Alternating Decision Tree (AD Tree), and Best First Tree (BF Tree). The authors have conducted the experiment through Weka tool. The classifier has applied for two test beds –cross validation which uses 10 folds with 9 folds used for training each classifier and 1 fold is used for testing and percentage split uses 2/3 of the dataset for training and 1/3 of the dataset for testing. The authors have claimed that the decision trees have a standard construct and easy to understand from which the rules can be extracted. The authors have also stated that j48 classifier has the highest accuracy with 99%.

Williams et al. [1] have focused at two data mining techniques namely naïve bayes and j48 decision trees to predict breast cancer risks in Nigerian patients. The analysis is made to determine the most efficient and effective model. The authors have collected the dataset from cancer registry of LASUTH, Ikeja in Lagos, Nigeria which contains 69 instances with 17 attributes along with the class label. The dataset holds 11 non-modifiable factors and five modifiable factors. The experiment is conducted through Weka and the authors have claimed j48 decision tree is better for the prediction of breast cancer risks with the values of accuracy (94.2%), precision, recall and error rates.

Majali et al. [9] have presented a diagnostic system using classification and association approach in data mining. The authors have used Frequent Pattern (FP) in association rule mining for classifying the patterns that are frequently found with benign and malignant instances. The authors have also used decision tree algorithm for predicting the possibility of cancer with respect to age. The authors have implemented Fp-growth algorithm for generating the frequent itemset without candidate generation which improves the performance of algorithm. The authors have claimed that their algorithm is able to achieve 94% of prediction accuracy.

Sivakami [10] has presented a disease status prediction model by employing a hybrid methodology of Decision Trees (DT) and Support Vector Machines (SVM). To alarm the severity of the disease the strategy of the system consists of two main parts namely information treatment and option extraction, and decision tree- support vector machines. The author has compared the results of the proposed model with Instance-based Learning (IBL), Sequential Minimal Optimization (SMO) and Naïve Bayes (NB) and has proven that proposed algorithms works better than the comparative algorithms with 91% of accuracy.

Sumbaly et al. [11]have discussed j48 decision tree classification algorithm for breast cancer diagnosis along with the summarization on the types of breast cancer, risk factors, disease symptoms and treatment. The authors have proven that the j48 algorithm is able to produce 94.5% of accuracy with correctly classified instances and have also suggested that neural network and digital mammography would be the alternative approaches for breast cancer prediction.

The pictorial representation of the accuracy obtained in the surveyed algorithms has been presented in fig. 1. Five out of ten papers have suggested that the decision tree classification is best for the prediction of breast cancer. Moreover, the prediction accuracy of the torus and random topology is the highest prediction accuracy with 99.97% in the surveyed articles.

Table 1
Obtained Accuracy of Reviewed Articles

S. No	*authors*	*Algorithm*	*Accuracy*	*Tool*	*Data Source*
1	Chaurasia et al.[3]	BF Tree	95.46%	Weka	UCI Machine Learning Repository
		IBK	95.90%		
		SMO	96.19%		
2	Shajahaan et al.[4]	ID3	92.99%	Weka	University of Wisconsin hospitals
		C4.5	95.57%		
		CART	92.42%		
		Naïve Bayes	97.42%		
3	Shrivastava et al.[5]	J48	98.14%	Weka	UCI machine Learning Repository
4	Senturk et al.[6]	Discriminant Analysis	92.75%	RapidMiner	UCI machine Learning Repository
		Multilayer Perceptron	96.39%		
		Logistic Regression	95.55%		
		SVM	96.40%		
		K-NN	95.15%		
5	Thein et al.[7]	Torus Topology	99.97%	Java Programming	University of Wisconsin hospitals
		Random Topology	99.97%		
6	Venkatesan et al.[8]	AD Tree (Alternative Decision Tree)	97%	Weka	Swami Vivekananda Diagnostic Centre Hospital
7	Williamset al.[1]	J48 Naïve Bayes	94.2% 82.6%	Weka	Cancer Registry of LASUTH, Ikeja in Lagos
8	Majali et al.[9]	ARM + DT	94%	Java application	Wisconsin Data
9	Sivakami[10]	DT+SVM	91%	LIBSVM	UCI Machine Learning Repository

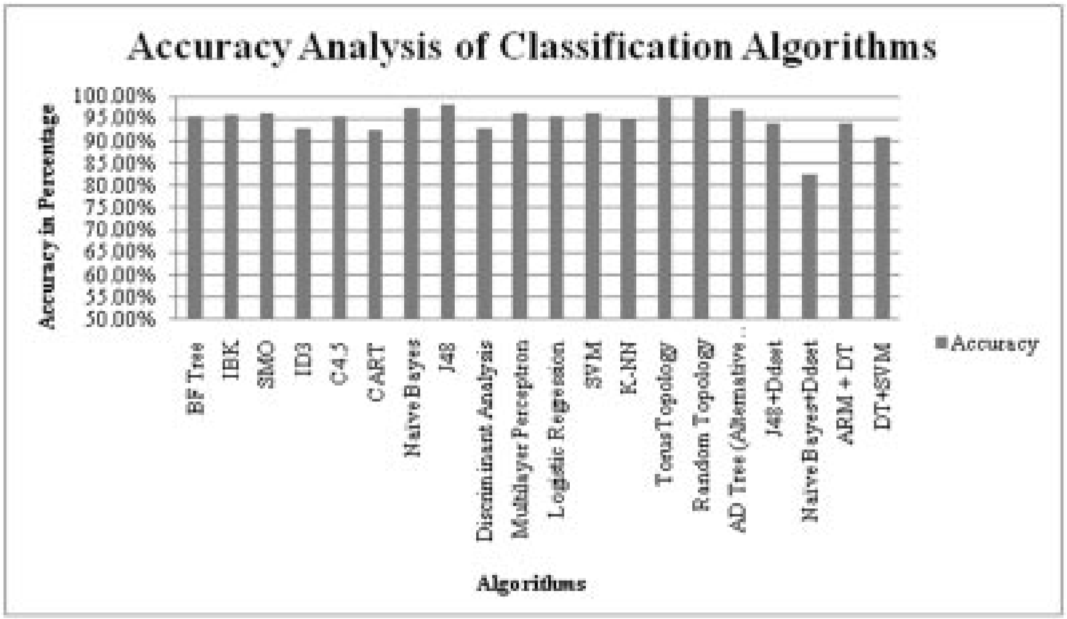


Figure 1: Comparative Analysis of Accuracy over Reviewed Articles

3. EXPERIMENTATION

The survey has clearly demonstrated that the decision tree algorithm is best suitable for the classification of breast cancer. To the proof of concept, in this paper an experiment is conducted to cross-verify the reviewed articles. Most of the experiments in the review is conducted by Weka tool, hence for a change of software, in this paper the experiment is conducted through MATLAB software. The dataset is taken from the UCI machine learning repository [12]. The dataset is created by Dr. William H. Wolberg, University of Wisconsin Hospitals, USA. Though the dataset contains 699 data points with 11 attributes, only 683 data points are taken into the consideration as the remaining data holds the missing values. The attribute that specifies the sample code number is also removed for the experiment. The description of the dataset is given in table.2.

4. RESULTS AND DISCUSSIONS

The dataset is saved as 'cancer.mat' and the class attribute is saved asaseparate file called 'class.mat' for classification. Initially the dataset is loaded for the execution.The source code that is shown in fig.2.is executed

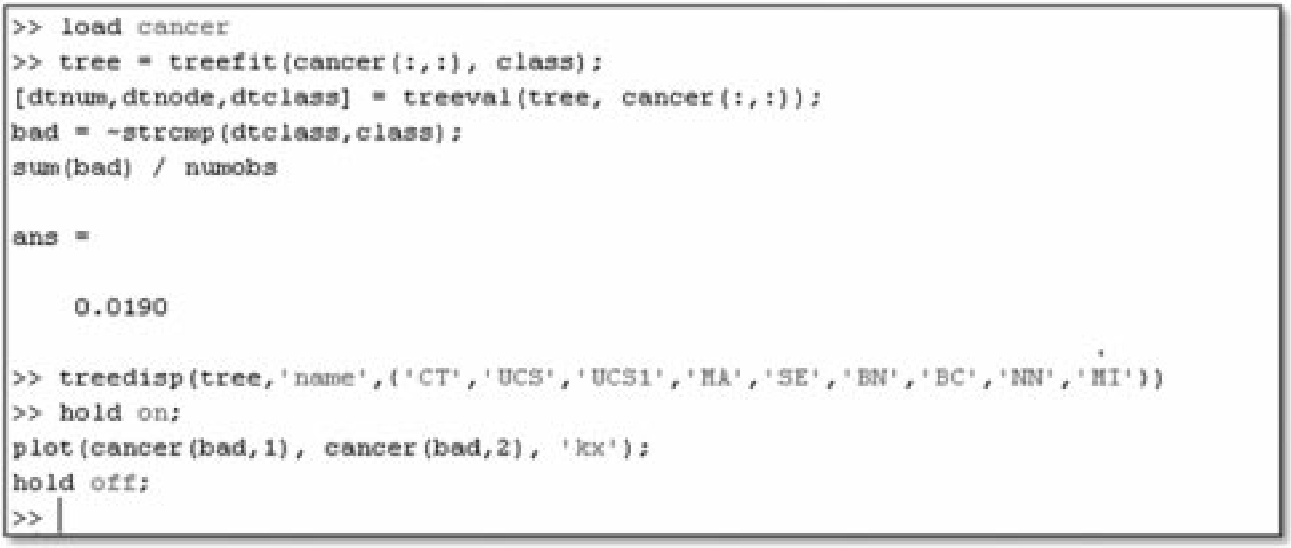
Table 2
Breast Cancer Dataset Description

S. No	*Attribute Name*	*Description*	*Range*
1	Clump Thickness	Assesses if cells are mono- or multi-layered.	1-10
2	Uniformity of Cell Size	Evaluates the consistency in size of the cells in the sample.	1-10
3	Uniformity of Cell Shape	Estimates the equality of cell shapes and identifies marginal variances.	1-10
4	Marginal Adhesion	Quantifies how much cells on the outdside of the epithelial tend to stick together.	1-10
5	Single Epithelial Cell Size	Relates to cell uniformity, determines if epithelial cells are significantly enlarged.	1-10
6	Bare Nuclei	Calculates the proportion of the number of cells not surrounded by cytoplasm to those that are.	1-10
7	Bland Chromatin	Rates the uniform "texture" of the nucleus in a range from fine to coarse.	1-10
8	Normal Nucleoli	Determines whether the nucleoli are small and barely visible or larger, more visible, and more plentiful.	1-10
9	Mitoses	Describes the level of mitotic (cell reproduction) activity.	1-10

for running the decision tree classifier. The value ‘ans’ denotes the error rate of the prediction which is 0.0190, which is 01%.

The knowledge prediction model that is derived by the decision tree classifier is shown in fig.3. which depicts the visualization of decision rule and group assignments for test data samples. The data points that are to be predicted are classified by the rules specified in the tree.

The pictorial representation of the misclassified data points are depicted in fig. 5. Out of 683 data points only 12 data points have been misclassified by the decision tree algorithm. Hence, the accuracy of the decision tree is 98%.



```
>> load cancer
>> tree = treefit(cancer(:,:), class);
[dtnum,dtnode,dtclass] = treeval(tree, cancer(:,:));
bad = ~strcmp(dtclass,class);
sum(bad) / numobs

ans =

    0.0190

>> treedisp(tree,'name',{'CT','UCS','UCS1','MA','SE','BN','BC','NN','MI'})
>> hold on;
plot(cancer(bad,1), cancer(bad,2), 'kx');
hold off;
>>
```

Figure 2: Execution of Decision Tree Algorithm in MATLAB

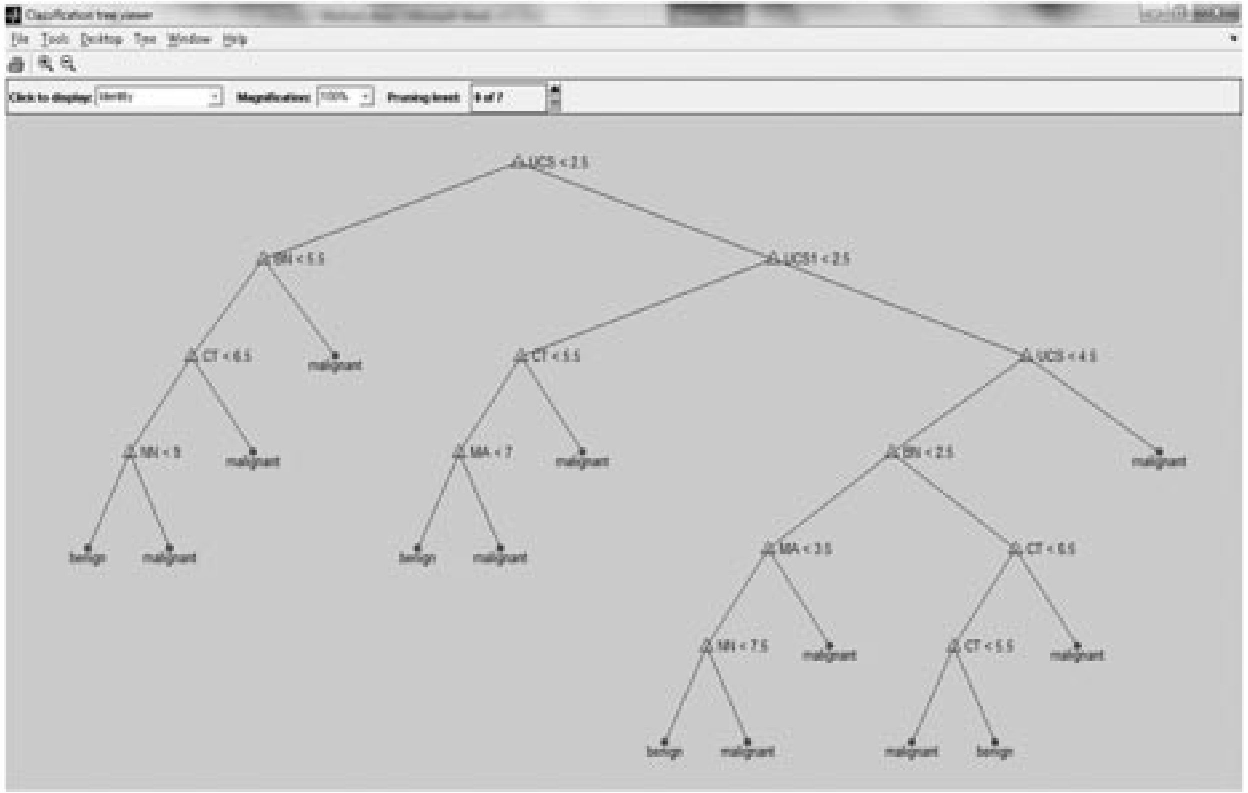


Figure 3: Knowledge Prediction Model for Decision Tree Classifier

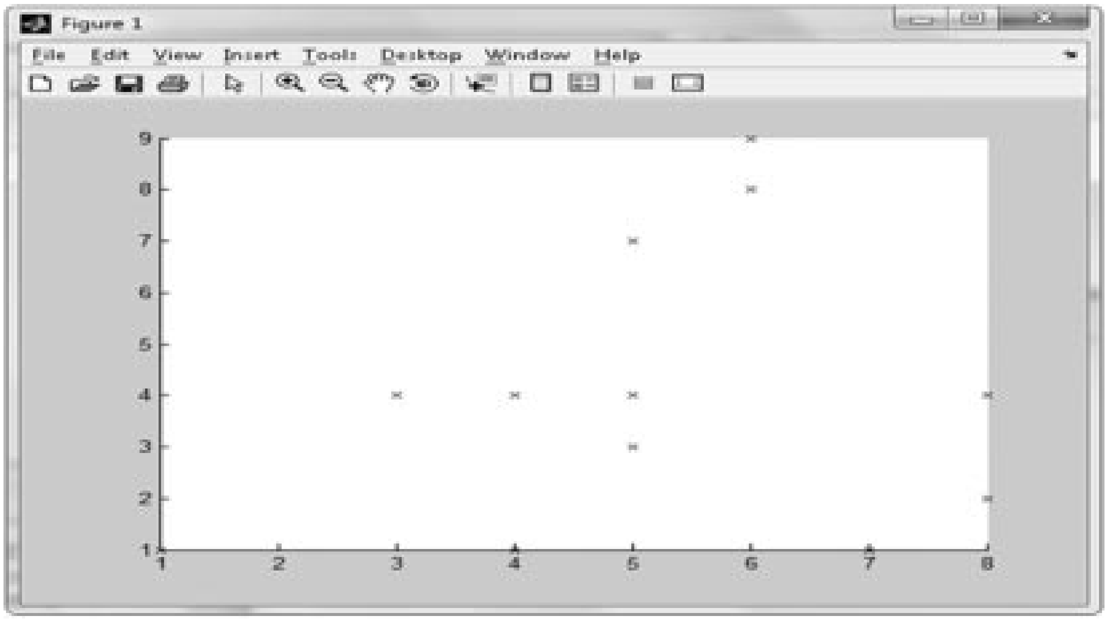


Figure 5: Misclassified Instances

5. CONCLUSION

This paper proves that the prediction of breast cancer is under the scope of classification algorithms in data mining. The algorithms build a knowledge prediction model using training dataset which is then used for the prediction of test data. This paper encompasses an extensive survey with the accuracy of 18 different types of classification algorithms. Among them, decision tree is the most predominant algorithm which has the highest prediction accuracy with 98%. To the proof of concept an experiment is also conducted with MATLAB software and the results are same with the surveyed papers. But the fact is, any automatic prediction must be 100% reliable and should not require any intervention of human. In future, the accuracy must be improved to 100%.

REFERENCES

[1] Williams, Kehinde, Peter Adebayo Idowu, Jeremiah AdemolaBalogun, and AdeniranIsholaOluwaranti. "Breast cancer risk prediction using data mining classification techniques." *Transactions on Networks and Communications*3**(2)**, 01-11, 2015.

[2] A. Joy Christy, S. Hari Ganesh, "Building Numerical Clusters Using Multidimensional Spherical Equation", *International Journal of Applied Engineering Research,* ISSN 0973-4562, **10(82)**, 629-634, 2015.

[3] Chaurasia, Vikas, and Saurabh Pal. "A novel approach for breast cancer detection using data mining techniques." *International Journal of Innovative Research in Computer and Communication Engineering*2**(1)**, 2456-2465, 2014.

[4] Shajahaan, S. Syed, S. Shanthi, and V. M. Chitra. "Application of Data Mining techniques to model breast cancer data." *International Journal of Emerging Technology and Advanced Engineering*3**(11)**, 362-369, 2013.

[5] Shrivastava, Shiv Shakti, Anjali Sant, and Ramesh Prasad Aharwal. "An Overview on Data Mining Approach on Breast Cancer data." *International Journal of Advanced Computer Research*3**(13)**, 256-262, 2013.

[6] Senturk, ZehraKarapinar, and Resul Kara. "Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven different algorithms." *Computer Science & Engineering*4**(1)**, 35-46,2014.

[7] Thein, HtetThazin Tike, and Khin Mo Mo Tun. "An Approach for Breast Cancer Diagnosis Classification Using Neural Network." *Advanced Computing*6**(1)**, 1-10, 2015.

[8] Venkatesan, E., and T. Velmurugan. "Performance analysis of decision tree algorithms for breast cancer classification." *Indian Journal of Science and Technology*8**(29)**, 1-8, 2015.

[9] Majali, Jaimini, et al. "Data Mining Techniques for Diagnosis and Prognosis Of Cancer."*International Journal of Advanced Research in Computer and Communication Engineering*4**(3)**, 613-616, 2015

[10] Sivakami, K. "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model."*International Journal of Scientific Engineering and Applied Science (IJSEAS)*–**1(5)**, 418-429, 2015

[11] Sumbaly, Ronak, N. Vishnusri, and S. Jeyalatha. “Diagnosis of Breast Cancer using Decision Tree Data Mining Technique.” *International Journal of Computer Applications***98(10)**, 16-24, 2014.

[12] https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/