Towards a Decision Support System by the study of Cell malfunctions for Breast Cancer

Sampurna Mandal, Supratim Bhattacharya, and Jayanta Poray
Department of Computer Science & Engineering
Techno India University, West Bengal
EM 4/1 - Salt Lake, Sector - V, Kolkata - 91, INDIA
Email: piu.sampurna@gmail.com, bhattacharya.supratim@gmail.com, jayanta.poray@gmail.com

***Abstract*—Breast cancer is one of the leading cause of death for women today; and it is the most common cancer in developed countries. The cause and degree of the breast cancer are very much associated with the malfunctions of its tissues and cells. It is very hard and rigorous task for the doctors to observe the clinical records for many affected patients and regulate the therapy manually. Therefore, it is very much necessary to properly process the bulk amount of clinical records (contain cell details) automatically and come with the best possible treatment for the affected patients. In this work we have proposed a decision support system with the help of two data mining techniques; namely, decision tree learning and association rules mining. Clinical data have been studied, pre-processed and analyzed with the help of a data mining tool (e.g., WEKA). Finally, as an outcome we come with the decision support tool for practical purpose.**

Keywords: Data Mining, Breast Cancer, Decision Tree, Association Rule, Decision Support System

I. Introduction

The event of breast cancer cases are increasing day by day. A new global study estimates that by 2030, the number of new cases of breast cancer in India will increase from the current 1,15,000 to around 2,00,000 per year[3]. In general cancer treatment and early successful diagnosis of the patients is a challenge since so many years. Since recent past doctors and researchers have been working very hard to find new ways to treat cancer. The variation and robustness of clinical diagonostic data for breast cancer patients are very huge. In general, the goal of Data Mining is to learn from robust data to generate many new and important information which is not addressed before. But it is not always easy to make any decision just by observing test cases. Two major shortfalls are: 1) information changes heavily time to time for cancerous patients; 2) it is very hard to find out the appropriate information (attributes) for optimal decision making purpose[2]. Here, we adapt the Data mining techniques for cancer treatment as a great support tool for doctors and physicians and facilitate them with decision making and estimation task. The need for biological data mining is that there is too much data but they are mostly unstructured. Data mining and machine learning depend on classification which is the most essential and important task. In other works of researchers,experiments have been performed on medical datasets for classification and feature selection using Data Mining techniques like Bayesian network,Rule based classifiers,Neural networks,Support Vector Machines. Many of them show good classification accuracy.

Data mining approaches in medical domain are increasing rapidly due to its effectiveness of classification and to generate the prediction system. Now-a-days it become a handy tool for medical practitionar. In addition to its importance in finding the ways to improve patient outcomes, it can also reduce the cost of the medicine, and help in enhancing clinical studies. Although there was a great deal of public education and scientific research,Breast cancer considered the most common invasive cancer in women, with more than one million cases and nearly 600,000 deaths occurring worldwide annually[3]. A good amount of research on breast cancer datasets is found in literature[4], [6], [7], [8]. Many of them used several data mining techniques and show a good or moderate classification accuracy.

In the same objective to achieve a certain classification accuracy, we have adopted the advantages of WEKA machine learning tool to analyze the Breast Cancer data for early diagnosis and proper preventive measures during therapy. We have used the Wisconsin Diagnostic Breast Cancer Dataset from UC Irvine[9] for our study. This dataset describes the detail of cell and tissue supported by the clinical results. The goal of our study is to classify the cancerous and noncancerous cases by investigating the diagnostic results precisely, study those cell attributes and generate the decisive model for the practitioner.

II. Problem Definition

In our work we have designed our model in such a way that we consider the dataset (UCI machine learning data for Breast Cancer) for decision making purpose. This dataset have been used in some other research work to satisfy some specific goal. Here we mainly study the effect of nine characteristic parameters on the

state of Breast cancer and the influence of the involved parameters on the performance of the decision tree learning and association rules mining models. We have used WEKA machine learning platform to implement our experimental model. In order to achieve the required result we thoroughly predict the various state, behavior and characteristics of breast cancer cells and tissue.

In this dataset, there are 698 samples taken from different women and every sample is expressed by nine characteristic parameters namely, i) Clump thickness, ii) Uniformity of cell size, iii) Uniformity of cell shape, iv) Marginal adhesion, v) Single epithelial cell size, vi) Bare Nuclei, vii) Bland chromatin, viii) Normal Nucleoli, and ix) Mitoses cell division. According to the properties of these nine parameters, the breast cancer is classified into benign & malignant classes. Every single parameter is given a range between 1 to 10 and the resultant class is expressed by 2 for benign and 4 for malignant. Among total 698 number of records in the dataset there are 16 samples with missing or incomplete data. So we have used remaining 682 records in this machine learning platform for our experiment.

III. Analytical Framework

A. Cancer cell description

The breast cell attributes behavior can be helpful in determining whether the cell is normal or cancerous. Some cells can be visualized like in given figure 1. In the figure it is seen that the nucleoli in normal cell is approximately invisible but the cancerous cell has an enlarged one. In the mitoses attribute the cell division is so fast and uncontrolled that the cell size and shape varies a lot for the cancerous cell. Then according to the bare nuclei attribute, a cancerous cell is comparatively dry.

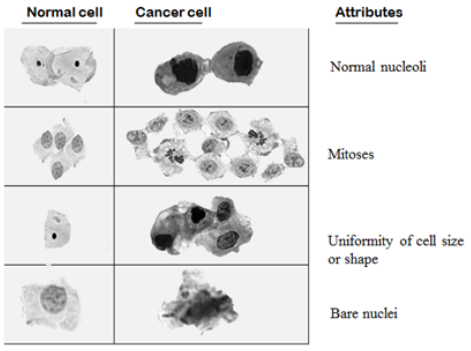

Fig. 1. Normal vs Cancerous cell attributes behavior

B. Decision tree

Decision tree is a popular classification method. Decision tree is used as a predictive model which maps observations about an item to conclusions about the item's target value. Rules produced by decision tree induction are easy to interpret and understand and hence can help greatly in appreciating the underlying mechanism that separate samples in different classes. One of the decision tree algorithms is C4.5. It builds decision trees from a set of training data using the concept of information entropy. It uses the information gain ratio criterion to determine the most discriminatory feature at each step of its decision tree induction process. Pruning helps to reduce the size of decision trees by removing sections of the tree that provide little power to classify instances. Pruning reduces the complexity of the final classifier and hence improves predictive accuracy by the reduction of over fitting.

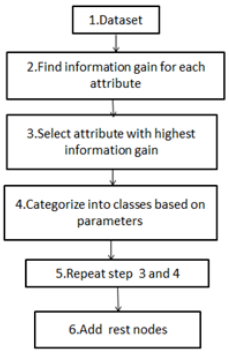

Fig. 2. Flowchart of C4.5

Algorithm:
Step 1: The leaf is labelled with the same class if the instances belong to the same class.

Step 2: For every attribute, the potential information will be calculated and the gain in information will be taken from the test on the attribute.

Step 3: Finally the best attribute will be selected based on the current selection parameter.

$$Entropy = -\sum_i p_i log_2 p_i$$

$$Info(S) = -\sum_{i=1}^{k} \frac{freq(C_i,S)}{|S|} log_2 \frac{freq(C_i,S)}{|S|}$$

where $Info(S)$ = entropy of sample training set S. C_i = class from 1 to n Taken $S = T$

$$Info_x(T) = \sum_{i=1}^{n}((T_i/|T|)).Info(T_i)$$

where $Info_x$ = entropy of each attribute

T_i = subsets of samples from 1 to n.

Information gain$(x) = Info(S) - Info_x(T)$

C. Association Rules Mining

Since its introduction in 1993 by Agarwal et. al.[1] the association rules mining has received a great amount of attention from several domains. Association Rules mining is the datamining process of finding the rules that may govern associations and causal objects between sets of items. It is used to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two subproblems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are frequent or large item sets.The second problem is to generate association rules from those large item sets with the constriants of minimal confidence. One of the popular algorithms is the Aprioiri algorithm. It iteravely reduces the minimum support until it finds the required number of rules with the given minimum confidence.

The algorithm is designed to mine association rules. In general if support $= freq(X,Y)/N$ and confidence $= freq(X,Y)/freq(X)$ correspondingly then the rule is $X \rightarrow Y$, where N is the no. of iterations. For example: There is some database containing a set of attributes. The support and confidence is calculated and the rules generated can be seen. For example in figure 3 there are 5 attributes going through 5 iterations. The rules generated are of the form $A \rightarrow D$ with support value of 0.4 and confidence value of 0.6. Similarly there can be more rules.

ATTRIBUTE 1	ATTRIBUTE 2	ATTRIBUTE 3
A	B	C
A	C	D
B	C	D
A	D	E
B	C	E

Fig. 3. Example-Association Rules Mining

IV. Proposed Model

Fig 4 shows the functional block diagram of our proposed model. It consists of four steps: (a) Acquisition, (b) Preprocess, (c) Feature Selection and (d) Feature Extraction.

In acquisition step, we get the dataset. In our work here we have clinical records for cancerous cell. Then in preprocessing we prepare the data to fit with several data mining techniques. These Data Mining techniques guide us to select the suitable features and extract these for the purpose of classification of benign and malignant cases. In figure 4 we first discretize the dataset for generating nominal values, then pre-process the data to eliminate incomplete information. Thereafter we adopt

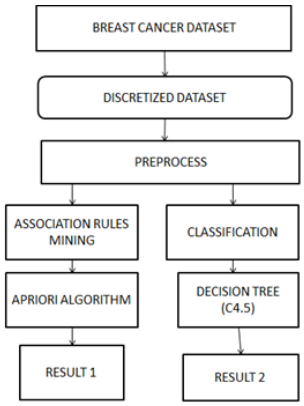

Fig. 4. Proposed Model

two Data Mining techniques namely, Association Rules Mining using Apriori Algorithms and classification using Decision Tree and come with respective results as shown in figure 4. This guides the doctors to consider appropriate therapy as an automatic support tool during their treatment.

V. Methodology

The attributes of the dataset are found as listed in Table 1.

Table1: WISCONSIN BREAST CANCER DATASET ATTRIBUTES

	Attribute	Domain
1	Sample Code No	id no
2	Clump Thickness	1-10
3	Uniformity(Cell Size)	1-10
4	Uniformity(Cell Shape)	1-10
5	Marginal Adhesion	1-10
6	sgl Epithelial(cell size)	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10
10	Mitoses	1-10
11	Class	2 or 4

In the clump thickness benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayers. While in the uniformity of cell size/shape the cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not. In the case of marginal adhesion the normal cells tend to stick together, where cancer cells tend to lose this ability. So loss of adhesion is a sign of malignancy. In the single epithelial cell size the size is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may

be a malignant cell. The bare nuclei is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors. The bland chromatin describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser. The normal nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible. In cancer cells the nucleoli become more prominent, and sometimes there are more of them. Finally, mitoses is nuclear division plus cytokines and produce two identical daughter cells during prophase. It is the process in which the cell divides and replicates. Pathologists can determine the grade of cancer by counting the number of mitoses cell division.

We next present our algorithm and further describe the dataset on which we have evaluated. Our first step is to discretize the dataset into three major groups.

1) Low
2) Mid
3) High

Algorithm 1 An algorithm for discretization of dataset

Input : Dataset in excel format with 9 parameters.
Output: : Dataset in csv file (space delimiter) format in discrete format with all 9 parameters.
Algorithmic Steps:

1) Obtain the ranges of high, middle and low.
2) Collect every cell value for computation for every parameter.
3) **For** parameter 1 to 9 do
4) **If** Cell value >= high value **Then Put** new Cell value= 'H' **Else If** Cell value >= middle value **Then Put** new Cell value= 'M' **Else Put** new Cell value= 'L'
5) **End if**
6) **Next**
7) **For** 10^{th} parameter
8) **If** Cell value = 2 **Then**
9) **Put** new Cell value = "Benign"
10) **Else If** Cell value = 4 then
11) **Put** new Cell value = "Malignant"
12) **End If**
13) Construct another excel file based on this discrete value.
14) Convert the excel file into csv(space delimiter) file.

As the parameter of the dataset ranges from 1 to 10 we made this discretization based on different ranges like:-

1) low:- 1 to 1, mid:- 2 to 6, high:- 7 to 10
2) low:- 1 to 1, mid:- 2 to 7, high:- 8 to 10
3) low:- 1 to 2, mid:- 3 to 7, high:- 8 to 10
4) low:- 1 to 3, mid:- 4 to 6, high:- 7 to 10
5) low:- 1 to 4, mid:- 5 to 6, high:- 7 to 10
6) low:- 1 to 4, mid:- 5 to 7, high:- 8 to 10

VI. Results

Based on these cell's behavior the dataset's attributes values can largely help in determination and diagnosis. On discretisation and preprocessing, each attribute shows its own graphical result divided into 3 categories based on high, medium and low values.

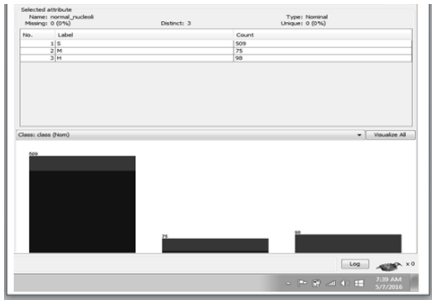

Fig. 5. Normal Nucleoli

Figure 5 shows normal nucleoli attribute and the no. of patient instances classified into benign and malignant group. The figure shows that out of 509 instances with low or small value has 30% malignant and 70% benign then out of 75 instances with medium value 90% are malignant and 10% benign then out of 98 instances with high value 99% are malignant and 1% benign cases.

The overall result is further shown in Figure 6 in the form of a Decision tree (Also known as J48 pruned tree). The detail of this observed result in WEKA platform classifies the benign (non-cancerous) and malignant (cancerous) classes based on selected cell attributes (e.g., size uniformity, bare nuclei, bland chromatin and shape uniformity) as shown in this figure. According to the observation the higher the value of the attributes the greater the tendency towards malignancy.On observing the J48 pruned tree and considering only the malignant category we can see that a combination of two medium values or one high and another medium value or only high values can help in determining the malignancy.

Also we have applied this modified dataset in WEKA Tool for further analysis; and we got the following results:

1) We observe that the value of clump thickness & bare nuclei tends to be on higher side. More than 20% of the values are on higher side, compared to other parameters where the range is 16% on higher side.

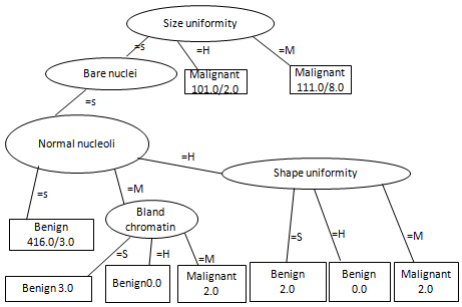

Fig. 6. Decision tree (J48 pruned tree)

INFO_GAIN	ATTRIBUTE NO.	ATTRIBUTE NAME
0.5788	2	size_uniformity
0.57	3	shape_uniformity
0.5411	6	bare_nucleoli
0.5034	7	bland_chromatin
0.4084	5	epithelial_size
0.3914	4	marginal_adhesion
0.3867	8	normal_nucleoli
0.3828	1	clump_thickness
0.0944	9	mitoses

Fig. 7. Ranking of attributes

2) More than 35% of the value for the parameter clump thickness, epithelial size & bland chromatin ranges in medium side.
3) Bare Nuclei's medium range value is in $< 11\%$ data whereas others had an average of 20%.
4) Malignancy is positive when clump thickness & bare Nuclei is on higher side but size uniformity & shape uniformity has a very sensitive effect, i.e when the value is $>= 4$ it shows a positive sign of malignancy in more than 80% cases.
5) More than 64% malignancy is positive only due to these two parameters.
6) For marginal adhesion & epithelial size, value ranges between 2 to 5. For more than 40% cases,malignancy is shown when these two parameters are low.
7) In case of bare nuclei & normal nucleoli, in 70% cases they tend to be low. They also show positive malignancy in 40% cases when they are low. They have 60% values in between 3 to 4. If the value is 5 or more then definitely it is malignant.

Selection of the most useful attribute is necessary. Information gain evaluator can help to find the attribute highest priority. The output presented in figure 7 shows information gain values of each and every attribute and their ranking. Size uniformity is selected as the best one. Then shape uniformity is taken into consideration. On selecting a set of attributes based on the ranking we get one or different kinds of rules. Among them the best and of utmost importance are selected.

Figure 8 shows the different association rules generated using uniformity of size,uniformity of shape and class after applying the Apriori algorithm. Among them rules no.2,3,4, 13 give significant results for determining malignancy. The first rule says that when there are 425 instances with uniformity of shape parameter's value to be small and the class they belong to is benign then 424 instances with small valued size uniformity attribute also follows showing confidence as 1.This rule does not seem to be so interesting to determine malignancy. Therefore on further analysis it is found that rule no.2,3,4 and 13 gives some thoughtful and result oriented knowledge. The second rule says that 92 instances with high value in shape-uniformity leads to malignant class with confidence of 0.99 The third rule says that 73 high valued instances in both size-uniformity and shape-uniformity leads to class malignant consisting 72 instances having confidence of 0.99. The fourth rule says that 101 instances with high valued size-uniformity is followed by 99 instances falling in malignant class showing confidence of 0.98. The thirteenth rule says that 111 instances with size-uniformity value as medium is followed by malignant class with 103 instances having confidence 0.93.

1. shape_uniformity=S class=Benign 425 ==> size_uniformity=S 424 conf:(1)
2. shape_uniformity=H 92 ==> class=Malignant 91 conf:(0.99)
3. size_uniformity=H shape_uniformity=H 73 ==> class=Malignant 72 conf:(0.99)
4. size_uniformity=H 101 ==> class=Malignant 99 conf:(0.98)
5. size_uniformity=S class=Benign 433 ==> shape_uniformity=S 424 conf:(0.98)
6. class=Benign 443 ==> size_uniformity=S 433 conf:(0.98)
7. shape_uniformity=S 457 ==> size_uniformity=S 442 conf:(0.97)
8. class=Benign 443 ==> shape_uniformity=S 425 conf:(0.96)
9. size_uniformity=S shape_uniformity=S 442 ==> class=Benign 424 conf:(0.96)
10. class=Benign 443 ==> size_uniformity=S shape_uniformity=S 424 conf:(0.96)
11. size_uniformity=S 470 ==> shape_uniformity=S 442 conf:(0.94)
12. shape_uniformity=S 457 ==> class=Benign 425 conf:(0.93)
13. size_uniformity=M 111 ==> class=Malignant 103 conf:(0.93)
14. shape_uniformity=S 457 ==> size_uniformity=S class=Benign 424 conf:(0.93)
15. size_uniformity=S 470 ==> class=Benign 433 conf:(0.92)

Fig. 8. Association Rules

The results as shown in Figure 9 shows the different statistical measures, like Kappa statistics, mean absolute error and all the other parameters which have their own significance and importance.

As an example by studying these measures, the argument between two or more observers that taken into account the fact that observers will sometimes agree or disagree simply by chance.The calculation is based on the difference between how much agreement is actually present (observed agreement) compared to how much agreement would be expected to be present by chance alone (expected agreement).The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or

Correctly Classified Instances	*654*	*95.8944%*
Incorrectly Classified Instances	*28*	*4.1056 %*

Kappa statistic	**0.9105**
Mean absolute error	**0.0636**
Root mean squared error	**0.1952**
Relative absolute error	**13.9707 %**
Root relative squared error	**40.9069 %**
Total Number of Instances	**682**

Fig. 9. Classification Results

equal to the MAE; The greater the difference between them the greater the variance in the individual errors in the sample. If RMSE=MAE then all the errors are of the same magnitude.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.959	0.042	0.977	0.959	0.968	0.961	Benign
	0.958	0.041	0.927	0.958	0.942	0.961	Malignant
WA→	0.959	0.041	0.96	0.959	0.959	0.961	

WA-Weighted Average

Fig. 10. Accuracy Measures

The figure 10 shows the detailed accuracy result with the help of some terms as True positive rate,False positive rate etc.The last row in the table gives the weighted average among the found result.

The figure 11 is the confusion matrix where rows show the actual instances and the column shows the predicted instances. a means benign and b means malignant.
The result shows that 425 instances are actually benign and predicted benign (True Positive).
18 instances are actually benign but predicted malignant (False Negative).
10 instances are actually malignant but predicted benign (False Positive).
229 instances are actually malignant and predicted malignant (True Negative).

Classified	a	b	
	425	18	a = benign
actual			
	10	229	b = malignant

Fig. 11. Resultant Confusion Matrix

VII. Conclusion

In this paper we have addressed the problem of breast cancer treatment. As the granularity of cases is robust, here we adapt the advantages of Data Mining techniques to build a Decision support system for practitioners. Here in particular we consider two Data Mining techniques, namely Decision tree and Association rules mining.We noticed that during the analysis of cell centric information of cancerous and non-cancerous cases, we get some results which encourage us to come with an automated Decision support system for therapy. As the analysis is very much data centric, the variation of clinical data/diagnostic information can vary the outcome. Some more Data Mining techniques could be adapted for future study and improvement of the result. Also we are considering the comparative analysis of the existing results with our achieved result.

References

[1] Agarwal, R., Imielinski, T., and Swami, A. N. ,1993. *Mining association rules between sets of items in large databases.* In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[2] Delen D, Patil N, *Knowledge extraction from prostate cancer data*. . The 39thAnnual Hawaii International Conference on System Sciences; 2006; 1-10.

[3] National Cancer Institute, *Surveillance, Epidemiology, and End Results (SEER) Program Public-Use Data (1973-2008).* Cancer Statistics Branch; 2011.

[4] A.M. Elsayad and H.A. Elsalamony. , *Diagnosis of Breast Cancer using Decision Tree Models and SVM*, International Journal of Computer Application (0975-8887),Volume 83-No 5,December 2013

[5] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR , *Using three machine learning techniques for predicting Breast cancer recurrence* , Health Med Inform 2013,4:2.

[6] S.S. Shajahaan, S. Shanthi, V. Mano Chitra, *Application of Data Mining Techniques to Model Breast Cancer Data.,*. . International Journal of Emerging Technology And Advanced Engineering, ISSN 2250-2459,ISO 9001:2008 Certified Journal, Volume 3,Issue 11,November 2013.

[7] Samar Al-Qarzaie, Sara Al Odhaibi, Bedoor Al-Saeed and Dr. Mohammed Al-Hagery. , *Using the Data Mining Techniques for Breast Cancer Early Prediction* , 2013

[8] V. Chaurasia and S. Pal., *Data Mining Techniques: To predict and resolve breast cancer survivabilty.*, International Journal of Computer Science and Mobile Computing,vol.3 Issue 1, January-2014

[9] UCI machine Learning Repository, *http://archive.ics.uci.edu/ml/* ,

[10] A. Bellaachia and Erhan Guven., *Predicting Breast Cancer Survivability using Data Mining Techniques.*, Dept.of Computer Science, The George Washington University. Washington DC 20052

[11] Kawsar Ahmed, Abdullah Al-Emran, Tasnuma Jesmin, Roshney Fatima Mukti, Md Zamilur Rahman and Farzana Ahmed*Early detection of lung cancer risk using Data Mining.* DOI:http://dx.doi.org/10.7314/APJCP .2013.14.1.595

[12] Jyoti Soni, Ujma Ansari , Dipesh Sharma and Sunita Soni*Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction.* International Journal of Computer Applications (0975 8887) Volume 17 No.8, March 2011

[13] Htet Thazin, Tike Thein1 and Khin Mo Mo Tun2, *Department of Computational Mathematics, University of Computer Studies, Yangon,Myanmar. An approach for breast cancer diagnosis classification using neural network.*, Advanced Computing: An International Journal (ACIJ), Vol.6, No.1, January 2015

[14] George Tzanis, Christos Berberidis, and Ioannis Vlahavas,*Biological Data Mining* , Department of Informatics, Aristotle University of Thessaoniki, Greece

[15] Thangaraju P1, Barkavi G2, Karthikeyan .T, *Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques* , International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014

[16] D. Lavanya and Dr.K. Usha Rani. *Ensemble decision tree classifier for breast cancer data.* International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012

[17] Neha Patel and Divakar Singh. *An Algorithm to Construct Decision Tree for Machine Learning based on Similarity Factor.*, International Journal of Computer Ap-plications (0975 8887) Volume 111 No 10, February 2015

[18] Salvatore Ruggieri,*Efficient C4.5.Dipartimento di Informatica*, Universita di Pisa Corso Italia 40, 56125 Pisa Italy

[19] Jonathan Tyrer, Stephen W. Duffy and Jack Cuzick. *A breast cancer prediction model incorporating familial and personal risk factors*, Department of Epidemiology; Mathematics and Statistics; Cancer Research U.K.; Wolfson Institute of Preventive Medicine; Charterhouse Square; London EC1M 6BQ; U.K. STATISTICS IN MEDI-CINE Statist. Med. 2004; 23:11111130 (DOI: 10.1002/sim.1668)