Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer

Shweta Kharya
Bhilai Institute of technology,
Durg, C.G. – India

Shika Agrawal
CSIT
Durg, C.G. – India

Sunita Soni
Bhilai Institute of technology,
Durg, C.G. – India

ABSTRACT
Naive Bayes is one of the most effective statistical and probabilistic classification algorithms. As health care environment is "information loaded" but "knowledge deprived". So to extract knowledge, effective analysis tools are constructed to discover hidden relationships in data. The aim of this work is to design a Graphical User Interface to enter the patient screening record and detect the probability of having Breast cancer disease in women in her future using Naive Bayes Classifiers, a Probabilistic Classifier. As breast cancer is considered to be second leading cause of cancer deaths in women today so early detection can improve the survival rate of women. The prediction is performed from mining the patient's historical data or data repository. Further from the experimental results it has been found that Naive Bayes Classifiers is providing improved accuracy with low computational effort and very high speed. The system has been implemented using java platform and trained using benchmark data from UCI machine learning repository. The system is expandable for the new dataset.

General Terms
Data Mining, Supervised Learning, Classification, Health Care.

Keywords
Breast Cancer, Naive Bayes Classifiers, UCI machine learning repository, Prediction, Posterior probability, Accuracy.

1. INTRODUCTION
Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns [4]. Data mining[5-8] is the set of techniques and tools applied to the non-trivial process of extracting and presenting/displaying implicit knowledge, previously un-known, potentially useful and humanly comprehensible, from large data sets, with object to predict automated form tendencies and behaviors; and to describe auto-mated form models previously unknown.[9-11] The term intelligent data mining[12] is the application of automatic learning methods[13,14] to discover and enumerate present patterns in the data. The model created in data mining can be Predictive and Descriptive in nature. A predictive model makes a prediction about values of data using known results found from different data. In this work classification technique of predictive model is used. Classification is supervised learning which maps data into predefined groups or classes .Classification applications includes image and pattern recognition, medical diagnosis ,detecting faults in industry and many more. Accuracy and the interpretability are the two parameters used to find the efficiency of classification model. Breast cancer is the most common cancer in women worldwide [15]. It is also the principle cause of death from cancer among women globally. The most effective way to reduce breast cancer deaths is its early detection. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy.

Breast cancer is the most frequently diagnosed cancer and is the leading cause of cancer death among women worldwide [16].

- Every 19 sec, somewhere around the world a case of breast cancer is diagnosed among women.
- Every 74 sec, somewhere in the world, someone dies from breast cancer.

In this paper evaluation of the performance of Naive Bayes Classifiers Model using the standard UCI datasets for prediction of presence of Breast Cancer is build. Further a GUI is designed to accept the patient's test results and predict the presence of Breast cancer diseases with more accuracy.

2. RELATED WORK
In [1] Naive Bayes Classifier has been applied to Wisconsin Prognostic Breast Cancer (WPBC) dataset (UCI Machine Learning repository:http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/), concerning a number of 198 patients and a binary decision class: non-recurrent-events of 151 instances and recurrent-events of 47 instances. The input features contain 12 relevant attributes describing the characteristics of cell nuclei. The testing diagnosing accuracy was about 74.24% in accordance of other well known Machine Learning techniques.

In paper [2] authors present the comparison of different classification techniques like Bayes Network, Radial Basis Function, and Pruned Tree and Nearest Neighbors algorithm using Waikato to Environment for Knowledge Analysis (WEKA) on large dataset. The data used in their investigation is the breast cancer data. It has a total of 6291 data and a dimension of 699 rows and 9 columns. In this 75% of overall data is used training and the rest is used for testing the accuracy of classification technique. According to the simulation result, highest accuracy is 89.71% which belongs to Bayes network with minimum time taken to build the model is 0.19 seconds and lowest average error is 0.2140 compared to others.

In paper [3] authors analyze the performance of supervised learning algorithm such as Naive Bayes, SVM Gaussian RBF kernel, RBF neural networks, Decision tree J48 and simple CART. These algorithms are used for classifying the breast cancer datasets WBC, WDBC, Breast tissue from UCI Machine learning Repository (http://archive .ics.uci.edu/ml) .They conducted their experiments using WEKA tool. In which the accuracy percentage of Naive Bayes algorithm for WBC dataset yields to be 96.50%, for Breast tissue dataset comes to be 94.33% and for WDBC dataset it is 92.61%.

3. RESEARCH OBJECTIVES

The core objective of this research is to develop a Probabilistic Breast Cancer prediction system using Naive Bayes Classifiers that can be used in making expert decision with maximum accuracy. The system can be implemented in remote areas like rural regions or country sides, to imitate like human diagnostic expertise for treatment of cancer ailment. The system is user friendly and reliable as model is already developed.

4. PREDICTIVE DATAMINING REVIEW

4.1 Classifiers

Classification is a Supervised learning technique in which given a set of class labels as a training set, a model called classifier is built to predict future data objects for which the class label is unknown. Several classification models have been proposed, like Bayesian Classification, Neural Networks and decision trees.

5. MATERALS AND METHODS

Naive Bayesian Classifiers are statistical classifiers which can predict class membership probabilities such as the probability that a given sample will belong to a particular case. Naive Classifiers assumes that the effect of an attribute value on the given class is independent of the values of other attributes. Naive Bayesian Classifiers depends upon the BAYE'S THEOREM. The major steps involve in Naive Bayesian Classifiers is figure out in Figure 1.

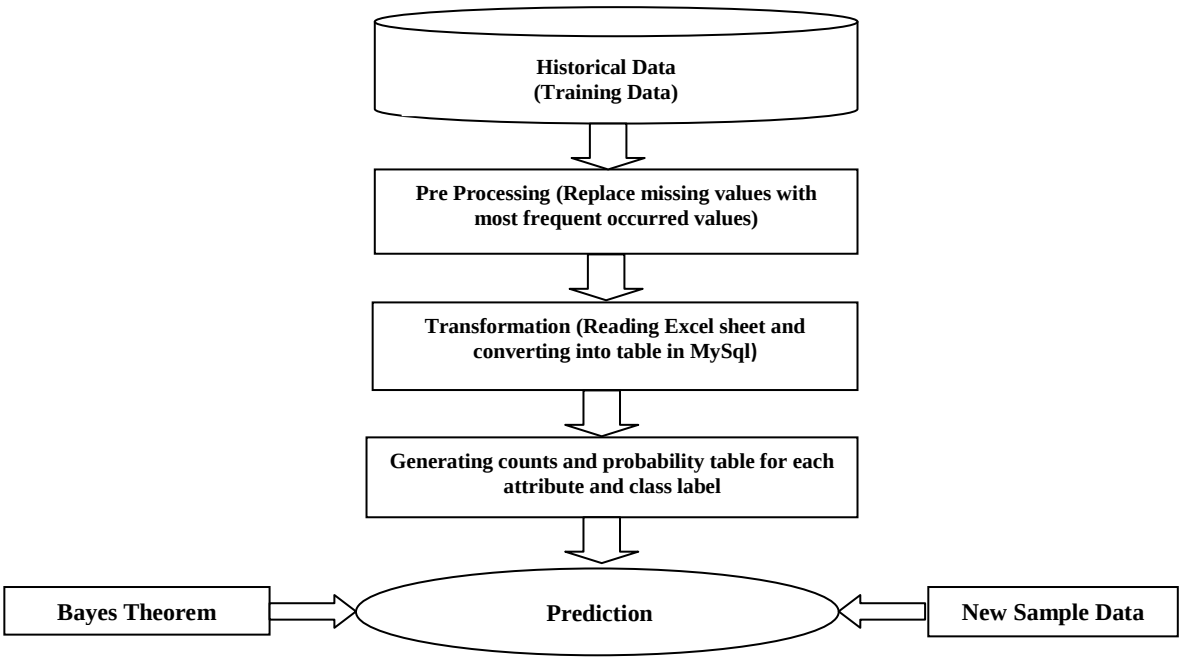

Fig 1.Main steps of Prediction using Naive Bayesian Classifiers

5.1. Breast Cancer Prediction System using Naive Bayesian Classifiers.

Naive Bayesian Classifier is based on Bayes Theorem. Naive Bayesian Model assumes that all the variables are mutually independent. Let D be the training set of tuples & their associated class labels. Each tuple is represented by N attributes such that a tuple will contain N values. Suppose there are m class labels from $C1, C2, \dots Cm$ for any new tuple X, then the classifier will predict that $X \in$ the class having highest probability condition on X. It shows that X belongs to the i^{th} class then i is having highest probability i.e

$$\text{If P } (\frac{Ci}{X}) > P\left(\frac{Cj}{x}\right) where\ 1 \le j \le m.$$

The class Ci for which $P(\frac{Ci}{X})$ is maximized is called maximum posterior hypothesis.

As $P(X)$ is constant for all the classes it is not considered & the formulas becomes

$$P(\frac{Ci}{x}) = P\left(\frac{X}{Ci}\right) * P(ci)$$

In order to predict the class label of X, calculate $P\left(\frac{x}{ci}\right) * P(ci)$ is evaluated for each class Ci and the predictor class label is class Ci for which $P\left(\frac{x}{ci}\right) * P(ci)$ is maximum.

5.2 Data Source

For training Wisconsin Datasets consisting of 699 records with 9 medical attributes have been used. Normalized dataset is available in

http://csc.liv.ac.uk/~frans/KDD/software/LUCS-KDD-DN/datasets/dataSet.html [16] in .num format. For our experiment the data in excel sheet is used directly. Table 1 shows different attributes with discretized values.

Table 1. Normalized Breast.D20.N699.C2 dataset with range of values

S.No	Attribute name	Range
1	Clump Thickness [1-10]	1
		2
2	Uniformity of cell size [1-10]	3
		4
3	Uniformity of cell shape [1-10]	5
		6
4	Marginal Adhesion [1-10]	7
		8
5	Single Epithelial cell size [1-10]	10
6	Bare Nuclei [1-10]	12
7	Bland Chromatin [1-10]	13
		14
8	Normal Nucleoli [1-10]	16
9	Mitoses [1-10]	18
10	Output (class label representing 2 type of breast cancer class)	19/20

5.3 Details of Attributes

Clump thickness

Benign cells tend to be grouped in monolayer's, whereas malignant cells are grouped in multilayer.

Uniformity of cell size/shape

The cancer cells tend to vary in size and shape. That is why these parameters are decisive whether the cells are cancerous or not.

Marginal adhesion

The normal cells tend to stick together, whereas cancer cells tend to lose this ability. So loss of adhesion is a sign of malignancy.

Single epithelial cell size

Epithelial cells that are notably enlarged may be a malignant cell.

Bare nuclei

Is a term used for nuclei that are not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors.

Bland Chromatin

It describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser.

Normal nucleoli

These 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.

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. Clinical experts can determine the score of cancer by counting the number of mitoses.[19]

From the breast.D20.N699.C2 datasets, Table of Count and probability is generated through our classifier model in MySql.

Table2. Count and Probability table of different discretized attribute with class Label

	values	Clump Thickness		values	Uniformity of Cell Size		Values	Uniformity of cell shape		Values	Marginal Adhesion		values	Single Epithelial Cell size	
		Yes	No		Yes	No		Yes	No		Yes	No		Yes	No
Count	**1**	283	20	**3**	281	3	**5**	352	1	**7**	373	34	**10**	458	241
	2	175	221	**4**	77	238	**6**	106	240	**8**	85	207			
Probability	**1**	0.62	0.1	**3**	0.83	0.01	**5**	0.77	0	**7**	0.81	0.01	**10**	1	1
	2	0.38	0.9	**4**	0.68	0.99	**6**	0.23	1	**8**	0.19	0.86			
	values	Bare Nuclei		values	Bland Chromatin		Values	Normal Nucleoli		Values	Mitosis		values	Class Label	
		Yes	No		Yes	No		Yes	No		Yes	No		Total	
Count	**12**	458	241	**13**	150	2	**16**	458	241	**18**	458	241	**19**	458	
				14	308	239							**20**	241	
Probability	**12**	1	1	**13**	0.33	0.01	**16**	1	1	**18**	1	1	**19**	0.66	
				14	0.63	0.99							**20**	0.345	

To implement our work, Java has been used as front end and MySql as back end. Benchmark medical dataset (UCL machine learning data set) [17] i.e. breast.D20.N699.C2.num have been used. The Classification Model is built using Count and Probability Table as shown in Table 2. For training, entire records have been used and testing is also performed on some data sets nearly one fourth. The screen shots of Breast cancer Classifier Model is shown in Figure 2.

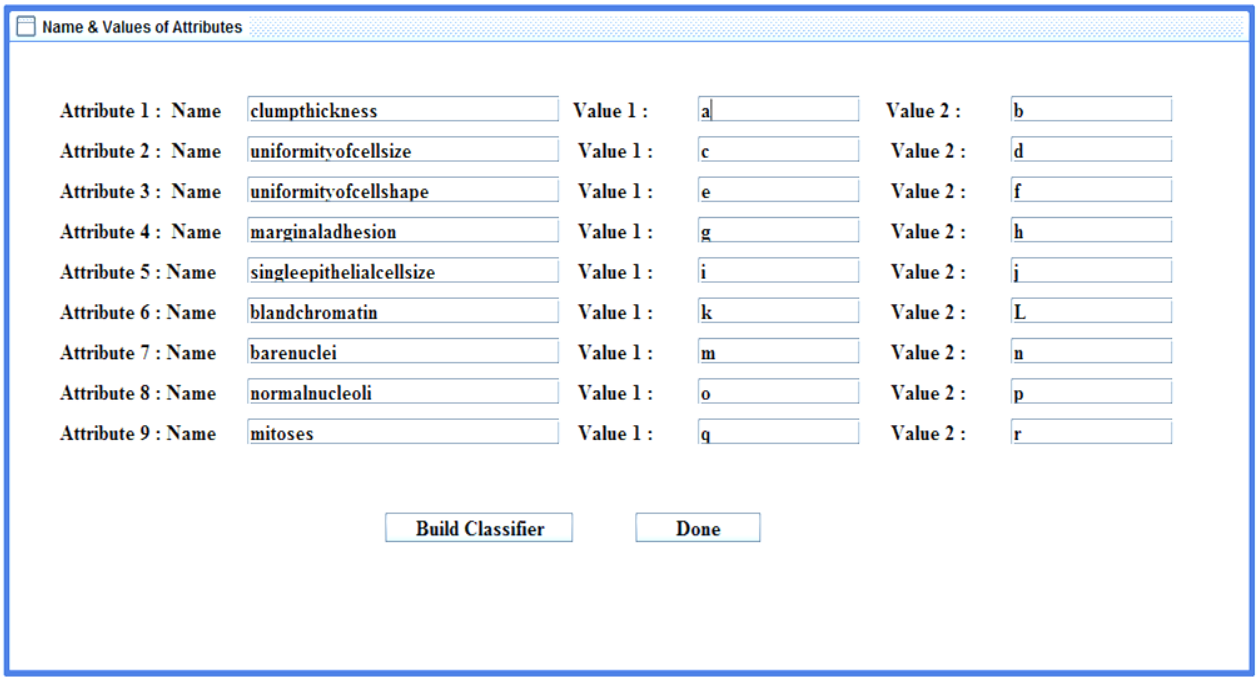

Fig 2.Interface designed for building the Naive Bayesian Classifiers Model

The screen shots of Breast Cancer Prediction System with 2 cases, breast cancer present (Malignant) and breast cancer absent (Benign) are shown in Fig 3 and Fig 4.

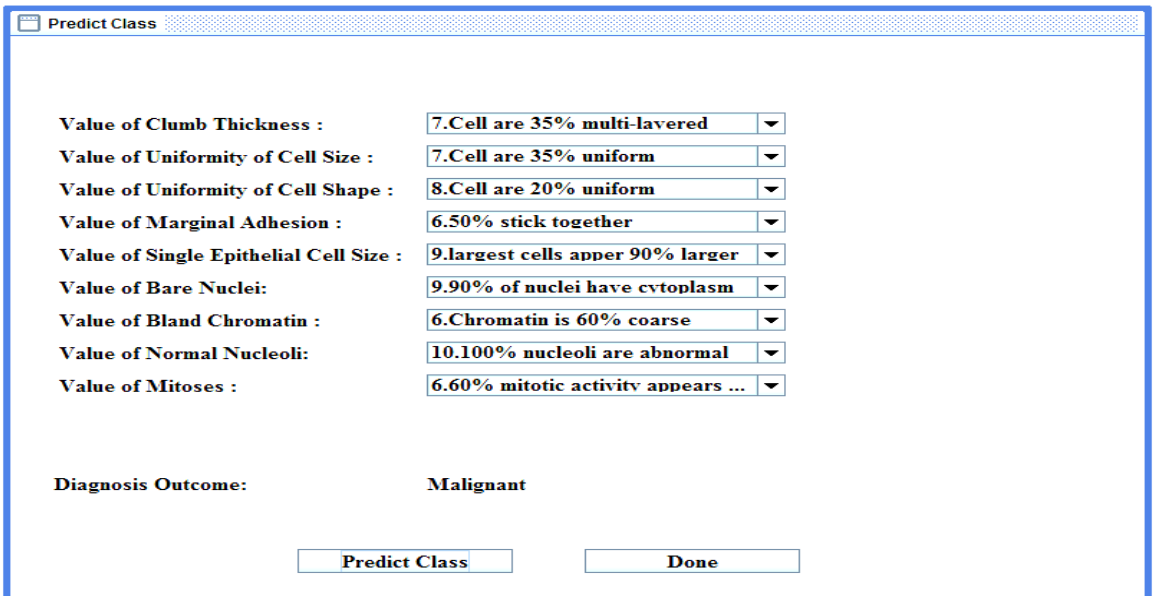

Fig 3. GUI of Predictive System with Malignant Case.

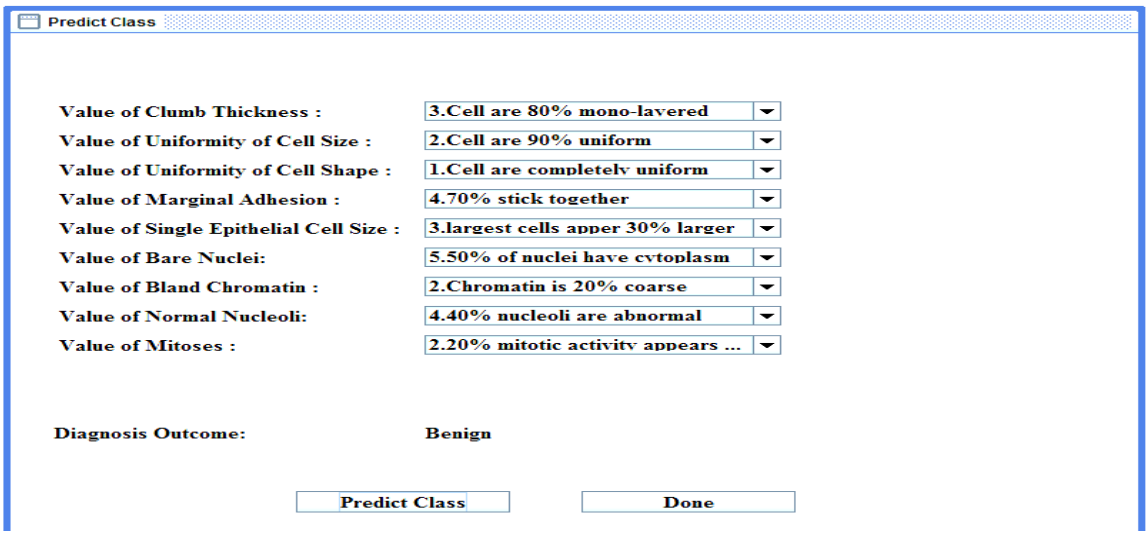

Fig 4.GUI of predictive system with benign Case

6. OBSERVATION AND EXPERIMENTAL RESULTS

We noticed that the Naive Bayesian Classifiers is sensitive as it predicts the disease on the ground of probability of disease being present in the findings. The breast.D20.N699.C2.num dataset is having almost 65.5% benign cases and remaining 34.5% malignant cases [16]. And the dataset incorporated in two class labels, one for Benign and other Malignant. The accuracy is found to be 93%.

- Training Data set-699 records
- Test Data set-200 records
- Accuracy of software
  - Based on correct prediction of TEST DATA with given class label.
  - Accuracy= Number of correct prediction/ Total number of records in a test data
  - Accuracy of software= Average accuracy of sample test data = 93%

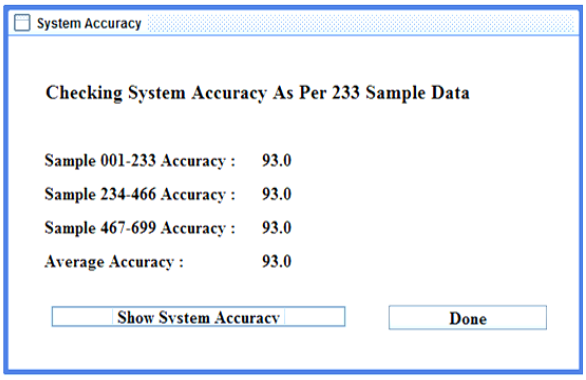

Fig 5. GUI for Software accuracy

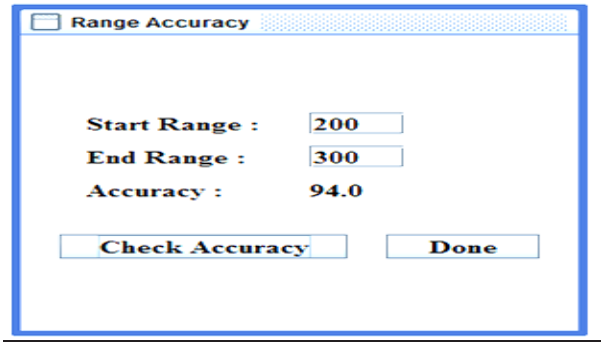

Fig 6.GUI for Range Accuracy

The GUI for the calculation of accuracy of in between range values of datasets is also designed.

7. DISCUSSION

The proposed system is web based ,user friendly, scalable and reliable that can be implemented in remote areas like rural regions in Primary Health center ,to imitate like human diagnostic expertise for finding the chances of having breast cancer in female in their coming future . The system can be expanded in the sense that more number of records or attributes can be incorporated [18] .In this work an intelligent and effective breast cancer prediction system using Naive Bayesian classifiers is designed. In this evaluation of the performance of NBC in terms of accuracy is achieved using benchmark data set (UCI machine learning repository) available in http://csc.liv.ac.uk/~frans/KDD/software/LUCS-KDD-DN/datasets/dataSet.html .Experimental results reveal that NBC is an efficient approach for extraction of significant patterns from breast cancer dataset. The maximum accuracy of 93% have been achieved .A GUI has been designed to enter the patient's records and presence of Breast cancer for a patient is predicted using the probability of disease being present in the probability of finding the symptoms.

8. ACKNOWLEDGEMENT

Author acknowledge the guidance and support received from Mrs. Sunita Soni, Sr.Associate Professor, BIT ,Durg for motivating me for my research work. I am thankful to Mrs.Deepty Dubey and Shikha Agrawal, Assistant Professor ,CSIT , Durg for many helpful suggestions and Dr.Ani Thomas ,HOD ,Department of Computer Applications for the

use of resources provided by department. At last but not the least I am thankful to Bhilai Institute of Technology Management for timely support and encouragement in the field of research and development.

9. REFERENCES

[1] Diana Dumitru "Prediction of recurrent events in breast cancer using the Naive Bayesian Classification" Annals of University of Craiova, Math. Comp. Sci. Ser. Volume 36(2), 2009, Pages 92-96 ISSN: 1223-6934.

[2] Mohd Fauzi bin Othman,Thomas Moh Shan Yau,"Comparison of Different Classification Techniques Using WEKA for Breast Cancer ",IFMBE Proceedings 15, pp. 520-523, 2007.

[3] S. Aruna, Dr S.P. Rajagopalan , L.V. Nandakishore, "Knowledge based analysis of various Statistical tools in detecting breast Cancer",CCSEA 2011,CS & IT 02,pp.37-45,2011.

[4] K. Usha Rani,"Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.

[5] G. Perichinsky and R. Garc´ıa-Mart´ınez, Proc. Workshop Comput. Sc. Researchers (La Plata University Press, Buenos Aires, 2000), p. 107.

[6] G. Perichinsky, R. Garc´ıa-Mart´ınez and A. Proto, Knowledge Discovery Based on Computational Taxonomy And Intelligent Data Mining, CD of the VI Comput. Sc. Argentinean Congr. (Ushuaia, 2000).

[7] G. Perichinsky, R. Garc´ıa-Mart´ınez, A. Proto, A. Sevetto and D. Grossi, Data Mining: Supervised and Non-Supervised Intelligent Knowledge Discovery, Proc. II Workshop Computes Sc. Researchers (San Luis University Press, San Luis, 2001).

[8] G. Perichinsky, A. Servetto, R. Garc´ıa-Mart´ınez, R. Orellana and A. Plastino, Tax-omic Evidence Applying Algorithms of Intelligent Data Minning Asteroid Families,comput. Sci., Software Eng., Information Technology, e-Bussines & Applications (Rio de Janeiro, 2003), p. 308.

[9] M. Chen, J. Han and P. Yu, IEEE Trans. Knowledge and Data Eng. 8, 866 (1996).

[10] H. Mannila, Methods and problems in data mining, Proc. of Int. Conf. on Database Theory (Delphi, Greece, 1997).

[11] G. Piatetski-Shapiro, W. J. Frawley and C. J. Matheus, Knowledge Discovery in Databases: An Overview (AAAI-MIT Press, Menlo Park, California, 1991).

[12] S. Evangelos and J. Han, Proc. 2nd Int. Conf. Knowledge Discovery and Data Min.(Portland, United States, 1996).

[13] R. S. Michalski, J. G. Carbonell and T. M. Mitchell, Machine learning I: An AI Approach (Morgan Kaufmann, Los Altos, CA, 1983).

[14] M. Holsheimer and A. Siebes, Data mining: The search for knowledge in databases,Report CS-R9406 (University of Amsterdam, Amsterdam, 1991).

[15] S.kharya ," Using data mining techniques for diagnosis and prognosis of cancer disease" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[16] http://csc.liv.ac.uk/~frans/KDD/software/LUCS-KDD-DN/datasets/dataSet.html.

[17] UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml/]. Irvine, CA: University of California, Center for Machine Learning and Intelligent Systems.

[18] Jyoti Soni et.al ,"Intelligent and effective Heart Disease Prediction System using Weighted Associative Classifiers" International Journal on Computer Science and Engineering,Vol.3 No.6,ISSN:0975-3397,June 2011.

[19] Gouda I. Salama1 et. al," Breast Cancer Diagnosis on Three Different Datasets Using Multi- Classifiers", International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01, September 2012.