Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining Techniques

Shomona Gracia Jacob[1], R. Geetha Ramani[2]

Abstract— **Data mining involves the process of recovering related, significant and credential information from a large collection of aggregated data. A major area of current research in data mining is the field of clinical investigations that involve disease diagnosis, prognosis and drug therapy. The objective of this paper is to identify an efficient classifier for prognostic breast cancer data. This research work involves designing a data mining framework that incorporates the task of learning patterns and rules that will facilitate the formulation of decisions in new cases. The machine learning techniques employed to train the proposed system are based on feature relevance analysis and classification algorithms. Wisconsin Prognostic Breast Cancer (WPBC) data from the UCI machine learning repository is utilized by means of data mining techniques to completely train the system on 198 individual cases, each comprising of 33 predictor values. This paper highlights the performance of feature reduction and classification algorithms on the training dataset. We evaluate the number of attributes for split in the Random tree algorithm and the confidence level and minimum size of the leaves in the C4.5 algorithm to produce 100 percent classification accuracy. Our results demonstrate that Random Tree and Quinlan's C4.5 classification algorithm produce 100 percent accuracy in the training and test phase of classification with proper evaluation of algorithmic parameters.**

Index Terms—**Breast Cancer Prognosis, Classification, Data mining, Feature Selection, Machine Learning**

I. Introduction

Data mining [1] is the process of hauling useful and related information from a database. Machine learning, [2-3] is concerned with the design and development of algorithms that allow computers to evolve behaviors learned from databases and automatically learn to recognize complex patterns and make intelligent decisions based on data. However the massive toll of available data poses a major obstruction in discovering patterns. Feature Selection attempts to select a subset of attributes based on the information gain .Classification [4-5] is performed to assign the given set of input data to one of many categories. Prognosis [6] is a prediction of outcome and the probability of progression-free survival (PFS) or disease-free survival (DFS) of a medical case.

Breast cancer ranks second as a cause of cancer death in women, following closely behind lung cancer. Statistics suggest [7-8] the possibility of diagnosing nearly 2.5 lakh new cases in India by the year 2015. Prognosis thus takes up a significant role in predicting the course of the disease even in women who have not succumbed to the disease but are at a greater risk to. Classification of the nature of the disease based on the predictor features will enable oncologists to predict the possibility of occurrence of breast cancer for a new case. The dismal state of affairs where more people are conceding to the sway of breast cancer, in spite of remarkable advancement in clinical science and therapy is certainly perturbing. This has been the motivation for research on classification, to accurately predict the nature of breast cancer.

Our research work mainly focuses on building an efficient classifier for the Wisconsin Prognostic Breast Cancer (WPBC) data set from the UCI machine learning repository [9-12]. We achieve this by executing twenty classification algorithms viz, Binary Logistic Regression (BLR), Quinlan's C4.5 decision tree algorithm (C4.5) ,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) classification algorithms. We also

Manuscript received May 07, 2012, revised June 5, 2012. This research work is a part of the All India Council for Technical Education(AICTE), India funded Research Promotion Scheme project titled "Efficient Classifier for clinical life data (Parkinson, Breast Cancer and P53 mutants) through feature relevance analysis and classification" with Reference No:8023/RID/RPS-56/2010-11, No:200-62/FIN/04/05/1624.

Shomona G.Jacob is a Full-time PhD research scholar in the Department of Computer Science and Engineering, Rajalakshmi Engineering College (affiliated to Anna University, Chennai), Thandalam, Chennai, India. Phone: 91-9841242291 (e-mail:graciarun@gmail.com)

Dr.R.Geetha Ramani is Associate Professor, Department of Information Science and Technology, College of Engineering, Anna University, Guindy, Chennai, India (e-mail: rgeetha@yahoo.com).

investigate the effect of feature selection using Fisher Filtering (FF), ReliefF, Runs Filtering, Forward Logistic Regression (FLR), Backward Logistic Regression (BaLR) and Stepwise Discriminant (Step Disc) Analysis algorithms to enhance the classification accuracy and reduce the feature subset size.

The following section reviews the past and current state of research in related areas of data mining.

II. Related Work

Previous research on application of data mining techniques in clinical research is briefly summarized in the following paragraphs.

Anagnostopoulos and Maglogiannis [13] employ a probabilistic approach to solve the Wisconsin Breast Cancer diagnosis problem, detecting malignancy among instances derived from the Fine Needle Aspirate test. For the diagnosis problem, the accuracy of the neural network in terms of sensitivity and specificity was measured at 98.6% and 97.5% respectively, using the leave-one-out test method. In the case of the prognosis problem, the accuracy of the neural network was measured through a stratified tenfold cross-validation approach. Sensitivity ranged between 80.5% and 91.8%, while specificity ranged between 91.9% and 97.9%, depending on the tested fold and the partition of the predicted period.

Mullins et.al, [14] applied a new data mining technique named 'Healthminer' to a large cohort of 667,000 inpatient and outpatient records from an academic digital system. HealthMiner approaches knowledge discovery using three unsupervised rule discovery methods: CliniMiner, Predictive Analysis, and Pattern Discovery. They tabulated the results for data trend characterization, discovery of medically known/unknown co-relations and identification of data anomalies using all the three unsupervised methods. Their results conclude that unsupervised data mining of large clinical repositories is feasible.

Mangasarian [15] performed classification on both diagnostic and prognostic breast cancer data. The classification procedure adopted by them for diagnostic data is called Multi Surface Method-Tree (MSM-T) that uses a linear programming model to iteratively place a series of separating planes in the feature space of the examples. If the two sets of points are linearly separable, the first plane will be placed between them. If the sets are not linearly separable, MSM-T will construct a plane which minimizes the average distance of misclassified points to the plane, thus nearly minimizing the number of misclassified points. The procedure is recursively repeated. Moreover they have approached the prognostic data using Recurrence Surface Approximation (RSA) that uses linear programming to determine a linear combination of the input features which accurately predicts the Time-To-Recur (TTR) for a recurrent breast cancer case. The training separation and the prediction accuracy with the MSM-T approach was 97.3% and 97 % respectively whereas the RSA approach was able to give accurate prediction only for each individual patient. Their drawback was the inherent linearity of the predictive models.

W.H. Wolberg [10-12] describes the accuracy of the system in diagnostically classifying 569 (212 malignant and 357 benign) Fine Needle Aspirates (FNA) and its prospective accuracy in testing on 75 (23 malignant, 51 benign, and 1 papilloma with atypia) newly obtained samples. The prospective accuracy was estimated at 97.2% with 96.7% sensitivity and 97.5% specificity using ten-fold cross validation. Using the standard error from the binomial distribution , they exhibited 95% confidence that the true prospective accuracy (the percentage of unseen cases that would be diagnosed correctly) lies between 95.8% and 98.6%.For prognostic data, the overall accuracy was estimated at 86%, with a 95% confidence region of ± 6%. Their results revolve around the clinical findings from mammogram images and reported a prospective accuracy of the projected system to be 86% by leave-one-out testing.

Falk et.al [16] has explored the results of Gaussian Mixture Regressors (GMR) on WPBC dataset and has concluded that the GMR performance is better than the performance of Classification And Regression Trees (CART) and Multivariate Adaptive Regression Splines (MARS) in predicting breast cancer recurrence time in patients who had a cancer excision.

Shekar Singh et.al, [17], presented work on breast cancer detection and classification based on H (Haematoxylin) & E (Eosin) stained histopathology and Feed Forward back propagation Neural Network (FFN). They concluded that FNN rendered fast and accurate classification and would be a promising tool for classification of breast cell nuclei. The overall accuracy of classification in the training, validation and testing mode were shown to be 96.34%, 95.54% and 95.80%.

Veerabhadrappa et.al [18] has compared the performance of three dimensionality reduction techniques on the Wisconsin Diagnostic Breast Cancer (WDBC), wine and zoo datasets. In the two approaches proposed, in level 1 of dimensionality reduction, features are selected based on mutual correlation and in level 2 selected features are used to extract features using Principal Component Analysis (PCA) or Locality Preserving Projections (LPP). Mutual correlation with PCA provided an average F-measure of 92.950, 85.146, and 87.073 for the Wine, Zoo and the Breast cancer datasets respectively whereas Mutual correlation with LPP provided an average F-measure of 95.148, 91.898, and 89.752 respectively.

A. Paper Organization

This paper is organized in the following manner. Section 3 portrays the proposed system design, clearly explaining each phase employed in the data mining process. In Section 4, we discuss the algorithms applied for feature relevance while Section 5 describes the classification algorithms. Section 6 reports the performance of the system with respect to the various algorithms employed while Section 7 concludes the paper.

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

III. Proposed Data Mining Framework

A. Overview

The proposed system design is diagrammatically presented in Fig 1. The data mining framework for the classifier is viewed from the perspective of both the training/learning phase and the test phase. The dataset is visualized and pre-processed before applying any of the data mining techniques The training phase then makes the learning process complete by generating all possible rules for classification after performing feature relevance followed by classification. The test phase determines the accuracy of the classifier when presented with a test data (unseen breast cancer case) and by viewing the returned class label.

B. Wisconsin Prognostic Breast Cancer (WPBC) Dataset

The description of the Wisconsin Prognostic Breast Cancer data is given in Table I. These are consecutive patients seen by Dr. Wolberg [9-11] since 1984. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The mean, standard error, and/or largest (worst case-mean of the three largest values) of these features were computed for each image, resulting in 30 features. The outcome is the target attribute (class label) and all other attributes (except ID) are predictor attributes whose values determine the result.

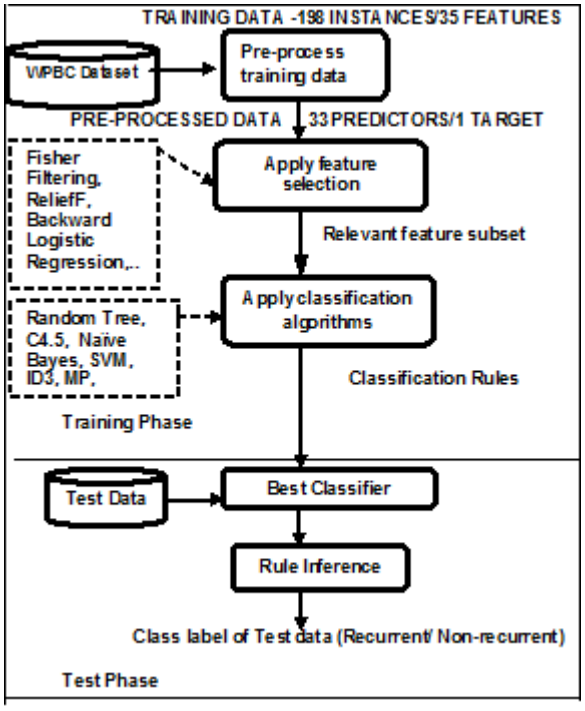

Fig 1. Proposed Data Mining Framework

C. Data Visualization and Pre-processing

The Wisconsin Prognostic Breast Cancer dataset is downloaded from the UCI Machine Learning Repository website [9] and saved as a text file. This file is then imported into Excel spreadsheet and the values are saved with the corresponding attributes as column headers. The missing values are replaced with appropriate values. The ID of the patient cases does not contribute to the classifier performance. Hence it is removed and the outcome attribute defines the target or dependant variable thus reducing the feature set size to 33 attributes. The algorithmic techniques applied for feature relevance analysis and classification are elaborately presented in the following sections.

TABLE I
WPBC Dataset Description

Attribute	*Significance*	*Attribute ID*
ID	Unique Identity of the patient	1
Outcome	Nature of the case (R-Recurrent/N-Non- recurrent)	2
Time	TTR(Time to recur)/DFS(Disease-free Survival)	3
Radius1,2,3	Mean of distances from centre to points on the perimeter	4,14,24
Texture1,2,3	Standard deviation of gray-scale values	5,15,25
Perimeter1,2,3	Perimeter of the cell nucleus	6,16,26
Area1,2,3	Area of the cell nucleus	7,17,27
Smoothness1,2,3	Local variation in radius lengths	8,18.28
Compactness1,2,3	Perimeter^2 / area - 1.0	9,19,29
Concavity1,2,3	Severity of concave portions of the contour	10,20,30
Concave points1,2,3	Number of concave portions of the contour	11,21,31
Symmetry1,2,3	Symmetry of the cell nuclei	12,22,32
Fractal Dimension1,2,3	Coastline approximation – 1	13,23,33
Tumour	Size of the tumour	34
Lymph node	Status of the lymph node	35

IV. Feature Selection Algorithms

The generic problem of supervised feature selection [19] can be outlined as follows. Given a data set $\{(x_i, y_i)\}$ $n_i=1$ where $x_i \in R_d$ and $y_i \in \{1, 2\ldots c\}$, we aim to find a feature subset of size m which contains the most informative features. The two well-performing feature selection algorithms on the WPBC dataset are briefly outlined below.

A. Fisher Filtering

It is termed Univariate Fisher's ANOVA ranking [20]. It is a supervised feature selection algorithm that processes the selection independently from the learning algorithm. It follows a filtering approach that ranks the input attributes according to their relevance. A cutting rule enables the selection of a subset of these attributes. It is required to define the target attribute which in this domain of research applies to the nature of the breast cancer (recurrent/non-recurrent) and the predictor attributes. After computing the Fisher score [21-22] for each feature, it selects the top-m ranked features with large scores. The next subsection directs focus on another technique of feature selection based on logistic regression.

B. Backward Logistic Regression

When the number of descriptors is very large for a given problem domain, a learning algorithm is faced with the problem of selecting a relevant subset of features [23].

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Backward regression includes regression models in which the choice of predictor variables is carried out by an automatic procedure. The iterations of the algorithm for logistic regression are given in Figure 2 as stated by Bewick [5].

Step 1: The feature set with all 'ALL' predictors.
Step 2: Eliminate predictors one by one.
Step 3: 'ALL' models are learnt containing 'ALL-1' descriptor each.

Fig. 2. Iteration 1 of Backward Logistic Regression

These iterations are further continued till either a pre-specified target size is reached or the desired performance statistics (classification accuracy) is obtained. After feature relevance, we classify the nature of the breast cancer cases in the Wisconsin Prognostic Breast Cancer dataset using twenty classification algorithms. The best performing algorithms are described in the following section.

V. Classification Algorithms

The classification algorithms that generated 100 percent accurate classification on the WPBC data are described below.

A. Random Tree Algorithm

Random trees [24-26] have been introduced by Leo Breiman and Adele Cutler. Random trees are a collection of tree predictors that is called forest. In most machine learning algorithms, the best approximation to the target function is assumed to be the "simplest" classifier that fits the given data, since more complex models tend to over fit the training data and generalize poorly [27]. The pseudo code of the Random Tree algorithm is given in Figure 3.

Input: The training set ⟶ TS, The set of attributes ⟶ X
Output: A random decision tree R
R = GenerateTree(X)
Procedure GenerateTree(X) //Provide number of attributes for split
If X is NULL then return leaf node
Else /*randomly select an attribute A as criterion for testing, create an internal node n with A as the attribute. Assume A has 'v' valid values*/
for i = 1 to v do
c_i = GenerateTree(X − {A})
Add c_i as a child of n
end for; end if; return n

Fig. 3. Pseudo code for Random Tree Classification

A sample rule generated by the Random Tree classification algorithm with Fisher Filtering feature selection algorithm is given in Figure. 4.

IF TIME < 48.5000 and IF Area3 < 1938.5000 and
IF TIME < 2.0000 and IF Radius3 < 20.7050
Then CLASS = N
ELSE IF Radius3 >= 20.7050 then CLASS = R

Fig. 4. Sample Rules from Random Tree Classification Algorithm

The methodology adopted by C4.5 algorithm is explained in the following sub-section.

B. Quinlan's C4.5 Decision Tree Algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [28]. Input to C4.5 consists of a collection of training cases, each having a tuple of values for a fixed set of attributes, A = $\{A_1, A_2...A_k\}$ and a class attribute. The goal is to learn from the training cases a function that maps from the attribute values to a predicted class. A sample rule generated by the C4.5 classification algorithm with the feature subset obtained by Backward Logistic Regression is briefly outlined in Figure 5.

IF TIME >= 50.0000
IF Radius1 < 12.7900
IF TIME < 88.0000
IF Compactness2 < 0.0553 then CLASS=R
IF Compactness2 >= 0.0553 then CLASS = N

Fig. 5. Sample Rule from C4.5 Algorithm for WPBC Dataset

VI. Performance Evaluation

The classification algorithms are ranked based on their accuracy in classifying the input datasets. Accuracy [1] of a classifier is measured in terms of how correctly the classifier places the input datasets under the correct category. This is denoted as the Misclassification rate which is computed as 1- Accuracy(C) where C denotes Classifier.

A. Test Data

Nearly 20% of the training data is further applied to test and verify the accuracy of the designed classifier. The values are tested against the rules on which the classifier is trained to classify the new breast cancer case as recurrent/non-recurrent.

B. Experimental Results

The twenty classification algorithms are applied on the Wisconsin Prognostic Breast Cancer dataset after it is pre-processed. The feature subset size selected by the algorithms is given in Table II. The comparative classification accuracy is depicted in Table III.

TABLE II
Feature Subset Size Selected on the WPBC Dataset

S.No	*Feature Selection Algorithms*	*Attribute ID of selected features (Referring Table 1)*
1.	Forward Logistic Regression (FLR)	3
2.	Fisher Filtering (FF)	3,27,24
3.	Stepwise Discriminant Analysis (Step DiscAnalysis)	3,5,27,4,35
4.	Backward Logistic Regression (BaLR)	3,4,19,20,24,29,30
5.	ReliefF Filtering (RFF)	3,25,35,9,12,5,34,11,10,8,28
6.	Runs Filtering(RF)	0

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE III
COMPARISON OF CLASSIFIER PERFORMANCE ON WPBC DATASET WITH FEATURE SELECTION

S.No	*Classification Algorithms*	*Accuracy (%)*	*Feature Selection Algorithms*			
			Fisher Filtering	*Backward*	*Stepwise*	*ReliefF*
1	KNN	82.32	83.84	83.84	84.34	83.84
2	Naïve Bayes	70.71	75.25	75.76	77.78	77.74
3	Random Tree	100	100	100	100	100
4	C4.5	100	100	100	100	100

The classification algorithm K-Nearest Neighbor shows an improved accuracy of 1 to 2 % while Naïve Bayes Continuous Classification show an improved accuracy of 5 to 7 % with the selected features as graphically represented in Figure 7. The performance of the classification algorithms before feature selection is given in Table IV.

TABLE IV
CLASSIFICATION PERFORMANCE ON WPBC DATASET
BEFORE FEATURE SELECTION

S.No	*Classification*	*Accuracy (%)*
1	Binary Logistic Regression (BLR)	87.37
2	C4.5	100
3	C-PLS	68.18
4	C-RT	76.26
5	CS-CRT	76.26
6	CS-MC4	92.93
7	C-SVC	84.85
8	ID3	76.26
9	KNN	82.32
10	LDA	88.89
11	Log-Regression	81.31
12	MP	90.4
13	MLR	87.37
14	NBC	70.71
15	PLS-DA	83.84
16	PLS-LDA	84.34
17	PROTOTYPE-NN	76.77
18	RBF	76.26
19	RND TREE	100
20	SVM	79.29

The size of the feature set to be considered for classification is reduced to less than one –third of the original feature set and hence less storage space is required for the execution of the algorithms.

The graphical representation of the performance of the classification algorithms is portrayed in Figure 6. However, FLR feature selection algorithm does not produce 100 percent classification for Random Tree algorithm but improves the classifier accuracy of Naïve Bayes by 5%. BaLR, FF, Step Disc Analysis and ReliefF reduce the error rate of Naïve Bayes and K-Nearest Neighbor classifiers and also give 100 percent classification accuracy with Random Tree and C4.5 algorithm.

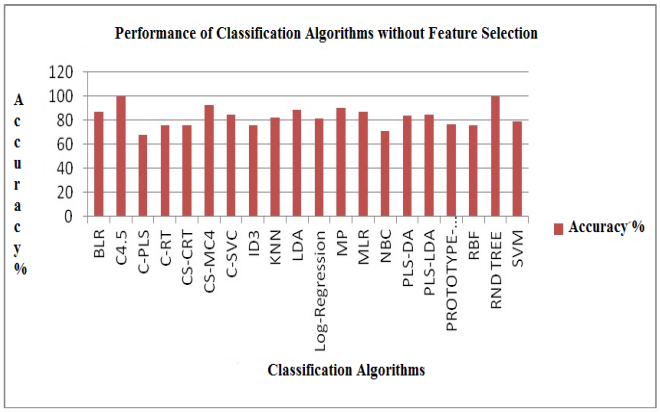

Fig.6. Classifier Performance before Feature Selection

It is to be noted that although the Random Tree and Quinlan's C4.5 algorithm produce 100 percent accurate classification, the size of the tree generated by C4.5 is much smaller than the tree obtained from the Random Tree algorithm.

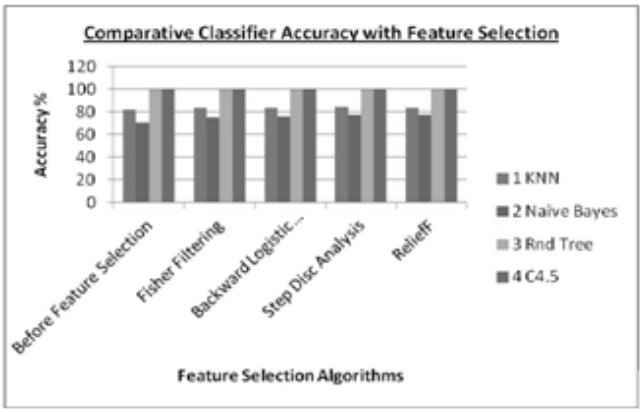

Fig.7. Classifier Performance Before and After Feature Selection

The Random tree algorithm produced a tree with 73 nodes and 37 leaf nodes which is much larger than the classification tree of the C4.5 algorithm that contains 55 nodes and 28 leaf nodes.

VI. CONCLUSION

In this paper we have considered the Wisconsin Prognostic Breast Cancer (WPBC) dataset for creating an efficient Classifier since it is highly essential in any clinical investigation to determine the nature of a disease, especially a life threatening ailment like cancer. The results of classification after feature selection are clearly outlined in this paper with necessary results. This will make it easier

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

for Oncologists to differentiate a good prognosis (non-recurrent) from a bad one (recurrent) and classify any new breast cancer dataset as being of a recurrent nature or non-recurrent one. Further accurate classification would enable clinicians to propose drugs for a new patient based on whether his/her features correspond to a good or bad prognosis. According to our findings, Fisher Filtering, Backward Logistic Regression, Stepwise Discriminant Analysis and ReliefF filtering algorithms have performed well in terms of improving classifier accuracy on this dataset. Random Tree and Quinlan's C4.5 classification algorithms have produced 100 percent accuracy in classifying the Wisconsin Prognostic Breast Cancer dataset. We also affirm that the Quinlan's C4.5 algorithm is the best performing classification algorithm on the WPBC dataset in terms of storage and classification accuracy since the decision tree generated is smaller and it also provides 100 percent classification accuracy.

REFERENCES

[1] J. Han and M. Kamber, *Data Mining; Concepts and Techniques*, Morgan Kaufmann Publishers, 2000.
[2] Mitchell, Tom M., *Machine Learning*. The Mc-Graw-Hill Companies, Inc., 1997
[3] S.B.Kotsiantis, *Supervised Machine Learning: A Review of Classification Techniques*, Informatica (31), 249-268, 2007.
[4] Tan, Steinbach, Kumar, *Introduction to Data Mining*, 2004.
[5] Vapnik, V. N., *The Nature of Statistical Learning Theory (2nd Ed.)*, Springer, Verlag, 2000.
[6] Sariego, J., "Breast cancer in the young patient". The American surgeon 76 (12): 1397–1401, 2010.
[7] American Cancer Society, Facts and Figures, 2010, http://www.cancer.org/acs/groups/content/@nho/documents/document/acspc-024113.pdf
[8] World Health Organization, Breast Cancer Statistics, http://www.who.int/cancer/detection/breastcancer/en/
[9] William H Wolberg, Olvi Mangasarian, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA
[10] W.H. Wolberg, W.N. Street, and O.L. Mangasarian, Image analysis and machine learning applied to Breast cancer diagnosis and prognosis, Analytical and Quantitative Cytology and Histology, Vol. 17, No. 2, pages 77-87, April 1995.
[11] W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian, Computer-derived nuclear ``grade" and breast cancer prognosis, Analytical and Quantitative Cytology and Histology, Vol. 17, Pages 257-264, 1995.
[12] W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian, Computerized breast cancer Diagnosis and prognosis from fine needle aspirates. Archives of Surgery 1995; 130:511-516
[13] Ioannis Anagnostopoulos and Ilias Maglogiannis, Neural network-based diagnostic and Prognostic estimations in breast cancer microscopic instances, Medical and Biological Engineering and Computing, Volume 44, Number 9, 773-784, 2006.
[14] Irene M. Mullins, Mir S. Siadaty, Jason Lyman, Ken Scully, Carleton T. Garrett, W. Greg Miller, Rudy Muller, Barry Robeson, Chid Apte, Sholom Weiss, Isidore Rigoutsos, Daniel Platt, Simona Cohen, William A. Knaus, Data mining and clinical data repositories: Insights from a 667,000 patient data set, Elsevier- Computers in Biology and Medicine, August 2005.
[15] D.S. O.L. Mangasarian, W.N. Street and W.H. Wolberg, *Breast cancer diagnosis and prognosis via Linear programming*, Operations Research, 43(4), pages 570-577, July-August 1995.
[16] Tiago H. Falk, Hagit Shatkay, Wai-Yip Chan, *Breast Cancer Prognosis via Gaussian Mixture Regression*, CiteSeerX- Scientific Literature Digital Library and Search Engine (United States), 2008.
[17] Shekar Singh, Dr.P.R.Gupta, Manish Kumar Sharma, Breast Cancer Detection and Classification of Histopathological Images, *International Journal of Engineering Science and Technology*, Vol. 3 No.5, May 2011, ISSN : 0975-5462.
[18] Veerabhadrappa, Lalitha Rangarajan, Bi-level dimensionality reduction methods using feature Selection and feature extraction, International Journal of Computer Applications (0975 – 8887) Volume 4 – No.2, July 2010
[19] Bing-Yu Sun, Zhi-Hua Zhu, Jiuyong Li, Bin Linghu, Combined Feature Selection and Cancer Prognosis Using Support Vector Machine Regression, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 6, pp. 1671-1677, Nov. /Dec. 2011.
[20] Tanagra Data Mining tutorials, http://data-mining-tutorials.blogspot.com/
[21] Tran Huy Dat, Cuntai Guan, Feature Selection Based on Fisher Ratio and Mutual Information Analyses for Robust Brain Computer Interface, *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.*
[22] Quanquan Gu, Zhenhui Li, Jiawei Han, *Generalized Fisher Score for Feature Selection, uai.sis.pitt.edu/papers*
[23] Viv Bewick, Liz Cheek, Jonathan Ball, *Statistics review 14: Logistic regression*, Critical Care. 2005; 9(1): 112–118. Published online 2005 January 13. doi: 10.1186/cc304
[24] L. Breiman, *Heuristics of instability and stabilization in model selection*, Annals of Statistics 24(6), 2350-2382, 1996
[25] Leo Breiman, Adele Cuttler, Random Trees, http://www.stat.berkeley.edu/users/breiman/RandomForests/
[26] Leo Breiman, Friedman, J. H., Olshen, R. A., & Stone, C. J., *Classification and regression trees*. 1984. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0412048418
[27] Gerard ´ Biau, Luc Devroye, Gabor ´ Lugosi, Consistency of Random Forests and Other Averaging Classifiers, *Journal of Machine Learning Research* (2008) 2015-2033
[28] Ron Kohavi and Ross Quinlan, *Decision Tree Discovery*, October 10, 1999.

AUTHOR'S PROFILE

Dr.R. Geetha Ramani is Associate Professor, Department of Information Science and Technology, College of Engineering, Anna University, Guindy, Chennai, India. She has more than 15 years of teaching and research experience. Her areas of specialization include Data mining, Bio-informatics, Evolutionary Algorithms and Network Security. She has over 50 publications in International Conferences and Journals to her credit. She has also published a couple of books in the field of Data Mining and Evolutionary Algorithms. She has completed an External Agency Project in the field of Robotic Soccer and is currently working on projects in the field of Data Mining. She has served as a Member in the Board of Studies of Pondicherry Central University. She is presently a member in the Editorial Board of various reputed International Journals.

Mrs.Shomona Gracia Jacob completed her M.E. in Computer Science and Engineering at Jerusalem College of Engineering, affiliated to Anna University, Chennai, India. She has more than 3 years of teaching experience. Presently she is pursuing her Ph.D in Computer Science and Engineering as a Full-time Research Scholar at Rajalakshmi Engineering College, affiliated to Anna University, Chennai. Her areas of interest include Data Mining, Bio-informatics, Artificial Intelligence and Machine Learning. She has presented and published papers in International Conferences and Journals.

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)