International Journal of Advanced Trends in Computer Applications
www.ijatca.com

A Quantitative Mining Model based Classification technique in active surveillance for DCIS

[1]Geeitha. S , [2]Dr. M. Thangamani
[1]Research Scholar
Kongu Engineering College
Department of Computer Science Engineering
[2]Assistant Professor
Kongu Engineering College
Department of Computer Science Engineering
[1]geethu.neelu@gmail.com, [2]manithangamani2@gemail.com

Abstract: *Breast Cancer is the second most leading cancer among women in the world. Detecting the disease at the earliest and parallel treatment may significantly increase the survival of the victim. A clinical trial has made an attempt to evaluate the low risk called ductal carcinoma in situ (DCIS), a monitoring approach rather surgery. This paper implements a quantitative model for mining the gene expression of DCIS for early detection of breast cancer under active surveillance that offers a close monitoring for the signs of progression of breast cancer. The proposed work determines whether the low-risk DCIS can undergo active surveillance without degrading the quality of life when compared to the conventional treatments. Our research proposes a classification technique using quantitative model based SVM, a hybrid technique for analyzing the gene expression of low risk patients.*

Keywords: *Breast Cancer, Ductal Carcinoma in situ (DCIS), Mining gene expression, Active surveillance, Classification.*

I. INTRODUCTION

Breast cancer is very severe disease in women all over the world. Ductal carcinoma in situ (DCIS) is a heterogeneous neoplasm with the potential probability of invasive breast cancer. Some of the risk factors of invasive cancer includes family background, genetic mutation, lifestyle of the patient, etc., It is a precancerous stage which can give rise to invasive breast cancer(IBC) if it progresses[1].Ductal carcinoma in situ is considered as a noninvasive carcinoma that ranges from low grade to high risk factor. These cells are confined to the basement membrane and do not invade normal breast parenchyma [2]. Many data mining techniques have been implemented in predicting the early stage of breast cancer among which decision tree(C5.0) and SVM has been proved best predictor with more than 80% accuracy than the fuzzy means and other classification algorithms and these mining techniques reveals the fact that classification provides better predictor algorithm than the clustering techniques[3]. Nowadays, molecular heterogeneity of IDC is vigorously emerging where the spectrum of molecular phenotypes of DCIS which is an immediate precursor to invasive breast cancer. But the identification of occurrence of molecular subtypes in IDC are also seen in DCIS [4] –[7]. Predicting the disease at the earliest is a challenging task where the data mining applications comes on the screen. The main objective of the work is to avoid or prevent the surgical treatment for the victim with low risk of the DCIS under active surveillance.

II. RELATED WORKS

Many data mining methods with new applications of classification techniques help doctors to detect the breast cancer at the earliest. Usage of AI (Artificial Intelligence) tools is also used to detect the occurrence of breast cancer and can also reduce the biopsies in the diagnostic process. A combined model of neural networks with AI methods can help in indicating and detecting the presence of breast cancers [8].There are some cases, where over diagnosis leads to overtreatment that exists particularly in low grade DCIS. This implies that the women with low grade DCIS can come under

active surveillance [9]-[11]. In [12] Emma J. Groen et.al proposed that DCIS is analyzed by applying the immunohistochemistry with genomic analysis on the affected specimens, but the size of the biopsies seems too small for these analyses. Therefore, laser microdissection or any alternative methods can be used to capture the cells and its tissue regions. An innovative approach may yield a construction for understanding the differentiation between the precancerous lesions and their environmental regions. In [13] Orlando Anunciacao et.al, explored about the importance of decision trees for detecting the high-risk DCIS from the dataset provided by Department of Genetics of faculty of Medical Sciences of Universidade with 94 samples in WEKA tool. A statistical validation is performed under permutation tests among which high-risk breast cancer group composed of 13 cases and only one holds the control, with a a p-value of 0.017. This reveals the fact that statistically significant associations is easy to detect with breast cancer with the help of the decision tree. In [14] Muhammad Umer Khan et al found a hybrid model based on fuzzy decision trees on SEER data. Some experiments have been performed using various combinations of decision tree rules, and different types of fuzzy rules associated with the inference techniques. They compared the performance of every gene for cancer prognosis and investigated a hybrid technique with fuzzy decision tree classification which is found to be more robust and balanced when compared with independently applied crisp classification.

III. Proposed Methodology

The capability of the SVM model shows the optimal results and also increases the diagnostic accuracy in classification and it has been proved under ROC curve [15]. Ensemble models are mainly applicable for the distributed data mining and online applications [16]. In this work, Naïve bayes, a data mining technique is implemented for classifying the DCIS data set. A Combined model of SVM with ensemble model is proposed to predict the DCIS in patients at the earliest. SVM is implied to segregate the two classes mainly hyper-plane and line, which easily maximizes the distances between the normal and infected gene expression. SVM model is effective in high dimensional space where it works with clear margin of separation. Though this model is best suited for DCIS analysis, still the model is not optimal as it is not very effective with noisy data and also does not provide any probability estimates. To address these issues an EM model is deployed to increase the accuracy of the classification. And 10 fold cross validation method is used to probability estimation. This combined model thus provides an optimal result for better prediction of DCIS to identify the low risk patients for active surveillance. The proposed methodology carries out three phases:

Phase1: Data acquisition and preprocessing

In our research work, 14 samples (**Table. 1**) of human gene expression of DCIS and invasive ductal carcinoma are taken from Geo dataset (NCBI source). These samples are tested in order to identify the earliest molecular marker genes that show under and over expression of ductal carcinoma. And it is preprocessed to remove noisy data. Cleaning and preprocessing is done using WEKA tool kit. Numerical values are converted to nominal type of data for easy classification.

Gene series of GSE21422			
Gene ID	**Category**	**ACCN**	**Bio.rep**
300535622	IDC	GSM535622	I5
300535621	IDC	GSM535621	I4
300535620	IDC	GSM535620	I3
300535619	IDC	GSM535619	I2
300535619	IDC	GSM535620	I1
300535612	DCIS	GSM535612	D9
300535611	DCIS	GSM535611	D8
300535610	DCIS	GSM535610	D7
300535609	DCIS	GSM535609	D6
300535608	DCIS	GSM535608	D5
300535607	DCIS	GSM535607	D4
300535606	DCIS	GSM535606	D3
300535605	DCIS	GSM535605	D2
300535604	DCIS	GSM535604	D1

Table: 1 Sample data set of IDC and DCIS gene data

.**Phase 2: Classification of DCIS gene expression**

After preprocessing gene expression is then uploaded in the tool for classifying. Naïve Bayes algorithm is used for classifying the training gene expression data. The below **Fig.1** depicts the

classification of DCIS data using NB model for predicting the low risk genes by partitioning the training data from the test data to segregate the enhanced expression genes from the under expression gene.

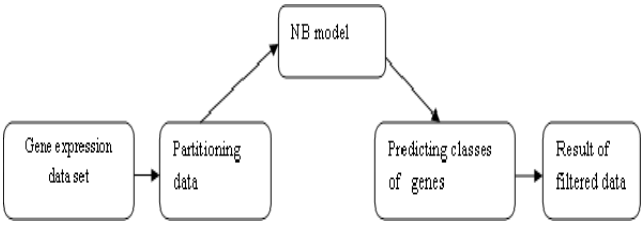

Fig: 1 NB model for Classification

Phase 3: Analyzing low risk factor for active surveillance

In this phase, second classification is carried out to analyze and differentiate between the DCIS and invasive carcinomas. This is to bring the insight in the development of different types of DCIS and its progression towards the invasive breast cancer on the molecular basis. Totally 14 samples are taken and analyzed using SVM based EM model which is an iterative method to find the best fit.

Procedure

Step1: Initialize the gene expression N with random positions and velocities on D dimension in the given space.

Step2: Setting the velocity vectors V_i= {1, 2, 3, N} to 0.

Step3: For each position Pi ε G of the particle P_i (i=1, 2…N) from the swarm, train the SVM classifier to compute the fitness function value

Step4: Detect the best global position Pg in the swarm that shows the minimal value of fitness function

Step5: For each candidate gene data P_j train the SVM classifier and compute the fitness function f (i).

Step6: Selecting the best global position Pg* of the gene expression and train the SVM with the detected feature subset

IV. RESULTS AND EXPERIMENTS

The main objective of this work is to analyze and identify the low risk molecules that lead to invasive ductal carcinoma for active surveillance. The work carries out by investigating the 14 instances of gene expression data. With these samples, training data set is classified after preprocessing to identify the DCIS gene expression to avoid any further surgery. The classification of different gene expression is evaluated using SVM classifier model. The below **Fig.2a and 2b** represents the classification of gene expression of DCIS and IDC. The classifier output reveals the fact that the patient with DCIS comes under the category of active surveillance.

Gene ID						
mean	300535620.7614	300535609.6765	300535606.3468	300535607.8856	300535619.7158	300535608.3894
std. dev.	6.2757	6.2757	6.2757	3.7096	6.2757	4.7529
Category						
IDC	2.5143	1.0048	1.0001	1.0043	4.4453	1.0311
DCIS	1.0154	4.4013	4.5583	1.577	1.0675	2.3795
[total]	3.5297	5.4061	5.5583	2.5814	5.5129	3.4106
ACCN						
GSM535622	1.5491	1.0001	1	1.0003	1.4466	1.0039
GSM535621	1.5243	1.0004	1	1.0005	1.4692	1.0056
GSM535620	1.2611	1.0021	1	1.0018	2.7233	1.0117
GSM535619	1.1799	1.0022	1	1.0017	1.8063	1.0099
GSM535612	1.0074	1.7322	1.0738	1.0462	1.0192	1.1213
GSM535611	1.0039	1.712	1.1073	1.0486	1.0131	1.1151
GSM535610	1.0021	1.6711	1.1521	1.052	1.0093	1.1134
GSM535609	1.0014	1.381	1.2652	1.0698	1.0085	1.274
GSM535608	1.0008	1.3183	1.3331	1.0718	1.0062	1.2698
GSM535607	1.0004	1.2771	1.4406	1.1215	1.0047	1.1557
GSM535606	1.0002	1.1418	1.6905	1.0564	1.0028	1.1084
GSM535605	1.0001	1.0995	1.7344	1.0553	1.0021	1.1086
GSM535604	1	1.0683	1.7613	1.0553	1.0017	1.1133
[total]	14.5297	16.4061	16.5583	13.5814	16.5129	14.4106
Bio.rep						
I5	1.5491	1.0001	1	1.0003	1.4466	1.0039
I4	1.5243	1.0004	1	1.0005	1.4692	1.0056
I3	1.1345	1.0006	1	1.0006	1.8595	1.0048
I2	1.1799	1.0022	1	1.0017	1.8063	1.0099
I1	1.1256	1.0015	1	1.0012	1.8638	1.0069
D9	1.0074	1.7322	1.0738	1.0462	1.0192	1.1213
D8	1.0039	1.712	1.1073	1.0486	1.0131	1.1151

Fig: 2a Classification of gene expression (DCIS and IDC)

Classifier output		
Incorrectly Classified Instances	14	100 %
Kappa statistic	-0.0769	
Mean absolute error	0.1429	
Root mean squared error	0.362	
Relative absolute error	103.934 %	
Root relative squared error	135.6119 %	
Coverage of cases (0.95 level)	0 %	
Mean rel. region size (0.95 level)	10.2041 %	
Total Number of Instances	14	

Fig: 2b Classifier output of 14 instances

A quantitative model based SVM is mainly implemented to identify the low risk molecules where active surveillance can be preferred at the earliest. The proposed work consists of 3 features such as mean area, standard deviation of area, average perimeter. SVM model is mainly used to classify the patients into two prognostic categories.

The first category includes patients with DCIS who is at the lower risk and prolong for survival time and the second is the victim with IDC stage that has to undergo some chemotherapy or radiation treatment depending upon the gene expression classification. The information gain ratio (**Fig. 3**) is implemented for uniquely identifying each sample to reduce a bias towards multivalued attributes taking the instances and size of the tumor cells.

```
Evaluator:     weka.attributeSelection.GainRatioAttributeEval
Search:        weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:      breast cancer data
Instances:     14
Attributes:    4
               Gene ID
               Category
               ACCN
               Bio.rep
Evaluation mode:    10-fold cross-validation


=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 1     +- 0        1   +- 0      1 Gene ID
 1     +- 0        2   +- 0      3 ACCN
 1     +- 0        3   +- 0      2 Category
```

Fig.3. Evaluation report using Gain Ratio Attribute

Using EM model with 10 fold cross validation , the gene expression data set is again split into training and testing sets where the model uses the probability distribution to each instance shown in **Fig.4a and 4b**. statistical methods to produce a probabilistic output based on the latent variables.

```
Scheme:        weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation:      breast cancer data
Instances:     14
Attributes:    4
               Gene ID
               Category
               ACCN
               Bio.rep
Test mode:     evaluate on training data

=== Model and evaluation on training set ===


EM
==

Number of clusters selected by cross validation: 6
```

Fig. 4a. Evaluation using EM model

```
Clustered Instances

1       4 ( 29%)
2       5 ( 36%)
4       5 ( 36%)


Log likelihood: -8.49726
```

Fig.4b. EM model Analysis

The prediction probability of IDC survival rate is very difficult task. Two classification models NB model and SVM model are used to predict and analyze the gene expression. The below **Fig. 5** represents an outline of DCIS and IDC patient of breast cancer.

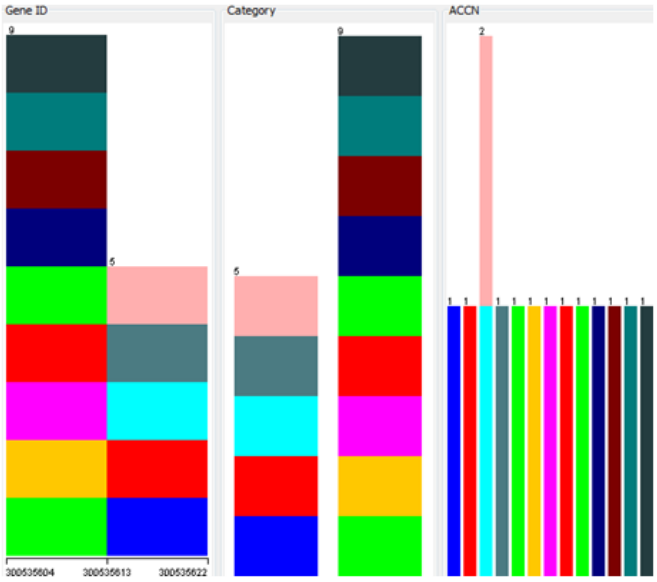

Fig.5. Identification chart of IDC and DCIS

V. CONCLUSION AND FUTURE WORK

To afford clinical treatments in the diagnosis of breast cancer becomes economically a challenging task in the medical field. Various data mining tools have been implemented for classification and analysis of breast cancer. In this paper, a quantitative model based SVM is used to differentiate the cancer stage into two phases. NB model plays a vital role in predicting the particular gene expression for breast cancer prognosis. Second classification is performed using SVM model. In order to increase the accuracy in classification EM model is designed to bring the optimal solution. Perhaps, there occurs some side effects due to this active surveillance when the patients are progressing towards the IDC. In future, analysis can be made to identify whether the patient needs active surveillance or surgery by designing a tool for analysing DCIS with high risk molecules and estimate the survival rate of the cancerous patients.

REFERENCES

[1] Erik S. Knudsen, Adam Ertel, Elai Davicioni, Jessica Kline,Gordon F. Schwartz, Agnieszka K. Witkiewicz, Progression of ductal carcinoma in situ to invasive breast cancer is associated with gene expression programs of EMT and myoepithelia, Springer, Volume 133, Issue 3, pp 1009–1024,

[2] Mohammed Badruddoja, Ductal Carcinoma In Situ of the Breast: A Surgical Perspective, International Journal of Surgical Oncology, PMC, 2012

[3] Uma Ojha, Savita Goel, A study on prediction of breast cancer recurrence using data mining techniques, Cloud Computing, Data science & Engineering –Confluence, 7th International Conference, IEEE, 2017

[4] Bryan BB, Schnitt SJ, Collins LC: Ductal carcinoma *in situ* with basal-like phenotype: a possible precursor to invasive basal-like breast cancer. Mod Pathol. Vol. 19, pp. 617-621, 2006.

[5] Livasy CA, Perou CM, Karaca G, Cowan DW, Maia D, Jackson S, Tse CK, Nyante S, Millikan RC: Identification of a basal-like subtype of breast ductal carcinoma *in situ*. Hum Pathol. Vol. 38, pp. 197-204, 2007.

[6] Steinman S, Wang J, Bourne P, Yang Q, Tang P: Expression of cytokeratin markers, ER-alpha, PR, HER-2/neu, and EGFR in pure ductal carcinoma *in situ* (DCIS) and DCIS with co-existing invasive ductal carcinoma (IDC) of the breast. Ann Clin Lab Sci. Vol. 37, pp. 127-134, 2007.

[7] Paredes J, Lopes N, Milanezi F, and Schmitt FC: P-cadherin and cytokeratin 5: useful adjunct markers to distinguish basal-like ductal carcinomas *in situ*. Virchows Arch. Vol. 450, pp. 73-80, 2007.

[8] D. Voth, Using AI to detect breast cancer, IEEE Intelligent Systems, Vol.20 No.1, pp.5-7, 2005.

[9] L.J. Esserman, I.M. Thompson, B. Reid, Overdiagnosis and overtreatment in cancer: an opportunity for improvement, JAMA J Am Med Assoc, Vol. 310, pp. 797-798, 2013.

[10] L.E. Elshof, K. Tryfonidis, L. Slaets, A.E. van Leeuwen-Stok, V.P. Skinner, N. Dif, e al., Feasibility of a prospective, randomised, open-label, international multicentre, phase III, non-inferiority trial to assess the safety of active surveillance for low risk ductal carcinoma in situ – the LORD study Eur J Cancer, Vol. 51, pp. 1497-1510, 2015.

[11] J.S. Wong, Prospective study of wide excision alone for ductal carcinoma in situ of the breast , J Clin Oncol, 24,Vol. 24, pp. 1031-1036, 2006.

[12] Emma J. Groen, Lotte E. Elshof, Lindy L. Visser, Emiel J, Th. Rutgers, Hillegonda A.O. Winter-Warnars, Ester H. Lips, Jelle Wesseling, Finding the balance between over- and under- treatment of ductal carcinoma in situ(DCIS), Elsevier, ScienceDirect, Vol.31, pp. 274-283,2017.

[13] Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo and Rueff Jose, "A Data Mining approach for detection of high-risk Breast Cancer groups," Advances in Soft Computing, vol. 74, pp. 43-51, 2010.

[14] Khan M.U., Choi J.P., Shin H. and Kim M, "Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare", Conf Proc IEEE Eng Med Biol Soc., pp. 48-51, 2008.

[15] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed, "Using data mining for assessing diagnosis of breast cancer," in Proc. International multiconfrence on computer science and information Technology, pp. 11-17, 2010.

[16] Nikunj C. Oza, Ph.D., NASA Ames Research Center, USA, Ensemble Data Mining Methods, https://ntrs.nasa.gov/search.jsp?R=20060015642, 2017.

Author Profile

Ms. S. Geeitha has completed Master of Engineering in Computer Science and Engineering in Anna University Chennai. She has 10 years experience in teaching. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing and ontology. She has published 10 international journals and has presented 14 papers in national and international conferences in the above fields. She has got senior educator and researcher award from National foundation for entrepreneurship development award. She is currently working as Head of the Information Technology Department and Assistant Professor in Mahendra Engineering College for women, Thiruchengode, Namakkal Dt..

M. Thangamani completed her B.E., from Government College of Technology, Coimbatore, India. She completed her M.E in Computer Science and Engineering from Anna University and PhD in Information and Communication Engineering from the renowned Anna University, Chennai, India in the year 2013. **Dr. M. Thangamani** possesses nearly 23 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published nearly 80 articles in refereed, indexed, SCI Journals, books and book chapters and presented over 67 papers in national and international conferences in above field. She has delivered more than 60 Guest Lectures in reputed engineering colleges and reputed industries on various topics. She has got best paper awards from various education related social activities in India and Abroad. She has organized many self-supporting and government sponsored national conference and Workshop in the field of data mining, big data and cloud computing. She has received the many National and International Awards. She continues to actively serve the academic and research communities and presently guiding nine Ph.D Scholars under Anna University. She is on the editorial board and reviewing committee of leading research SCI journals. She has Editor-In-Chief in International Scientific Global Journal in Engineering Science and Applied Research (ISGJESAR)