BreCaDS: Towards a Breast Cancer Decision Support System

April Roxelle Garcia
Ma. Beatrice Gerasmio
Mae Lizza Tuastumban
Milfe Ababa
Editha Dimalen

Abstract

Breast cancer is a common disease worldwide. In Southeast Asia, the Philippines has the highest mortality rate making it one of the top ten breast cancer hot spots in the world. Studies are conducted to identify if a person's genetic makeup and environmental factor have something to do with the disease. One genetic factor is Human Epidermal Growth Factor Receptor-2 (HER2). This study developed a computer-based decision support system (BreCaDS) that reveals a HER2 score which indicates if the cancer is aggressive or not through the given clinicopathological and cytological factors from pathology reports of breast invasive ductal carcinoma (BIDC) patients. The BreCaDS algorithm is based on the concept of pattern recognition and decision trees. Using 89 randomly picked data samples from breast cancer patients the system obtained a 75% reliability. This can aid clinicians in providing a better treatment course for breast cancer patients. BreCaDS is the first and pioneering study on breast cancer and HER2 on the Filipino population. It is recommended that with more data samples the reliability of the system can be improved.

Keywords: breast cancer, invasive ductal carcinoma, decision support system, HER2, metastasize, cytological data, clinicopathological data

---

GARCIA, GERASMIO, and TUASTUMBAN all graduated from MSU-IIT with the degree BS Information Technology major in Management Information System in March 2010. ABABA was a former Lecturer at the Department of Biological Sciences of the College of Science and Mathematics. DIMALEN, is an Associate Professor II of the Department of Information Technology, School of Computer Studies. She obtained her MS in Computer Science at the De La Salle University. She is currently pursuing her doctorate degree in Computer Science at Massey University in New Zealand.

Introduction

Cancer may be one of the biggest threats to human life. The Philippines is one of the Top Ten Breast Cancer Hot Spots in the world (Meneses, 2000). In breast cancer, a number of cells in the breast will develop abnormally and may metastasize in other parts of the body. One influencing factor is the Human Epidermal Growth Factor-2 (HER2). A HER2-positive breast cancer is caused by the overexpression or mutation of the HER2 gene. As a result, the affected cells grow and divide quickly and breast cancers become aggressive and is less responsive to hormone treatment (Pruthi, 2008). Research has shown that women with HER2-positive breast cancer have a more aggressive disease and greater likelihood of recurrence than women with HER2-negative breast cancer (Fyfe, 2006).

A literature review was performed to better understand how reliable this diagnostic software for medication. The review includes a study on related literature on breast cancer diagnostic software and how HER2 is related to breast cancer. This review presents existing medical computerized systems as sources of new techniques and methods that can be used in the development of this study.

To find out whether HER2 overexpression is a contributory factor to the existing breast cancer which is an indicator for aggressiveness, this study developed a computer-based decision support system that reveals the possibility of aggressiveness of a breast cancer and to recognize if a HER2 test might be needed. The idea of HER2 testing for breast cancer patients could be one of the best solutions. However, this test is expensive especially for the average Filipino breast cancer patient. In this study, a new technique is employed to aid clinicians in deciding if a HER2 test is indispensible to a certain breast cancer patient through a medical decision support system (BreCaDS) that shows the percentage likelihood towards HER2 overexpression given the clinicopathological factors (1) stage, (2) tumor size, and (3) nodal status and cytological factors (4) nuclear grade, (5) pleomorphism (6) characteristic of cell cytoplasm, (7) prominent nucleolus, and (8) observed necrosis from patient pathology reports of Filipino women.

BreCaDS applies the theory of supervised machine learning, pattern recognition and decision tree algorithm. Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses which then make predictions about future instances. The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features (Kotsiantis, 2007). Decision trees can handle high dimensional data. Their representations of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The classification and learning steps of a decision tree is fast and simple. In general, the actual decision tree algorithms are recursive. Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume (Kotsiantis, 2007). Decision trees present a clear, logical model that can be understood easily by people who are not mathematically inclined (Sherrod, 2001).

In order to predict the output and decide which attribute should be tested first, basically find the one with the highest information gain (Moore, 2006). The first step in BreCaDS algorithm is to create a decision tree out of the training data in which at this method an attribute is selected at each node of the tree. Through the mathematical concept of Entropy, Gini Index and Classification Error the system could come up with a tree with nodes that are at their proper place. This is the method in which attributes are selected at each node of the tree. Decision trees are built in order to capture underlying relationships in a dataset (Nashvili, 2004). In practice, there is no need to compute the impurity degree based on three methods. The use of either one of Entropy, Gini index or index of Classification error is already essential. Once information is gained for all attributes is computed then the optimum attribute that produces the maximum information gain is established (Teknomo, 2009).

Overfitting occurs when a decision tree characterizes too much detail, or noise in the training data. Capturing all the exact nuances and extremities of the training data is no longer significant because these are the sources of error. To address the problem of overfitting, decision trees adopt a pruning strategy. Most decision tree algorithms use the technique of postpruning or backward pruning. This essentially involves growing

the tree from a dataset until all possible leaf nodes have been reached and then removing particular subtrees (Nashvili, 2004). Studies have shown that post-pruning will result in smaller and more accurate trees by up to 25% (Esposito et al. 1997).

Methodology

Breast tumor tissue samples were obtained from the Capitol University Medical City Foundation, Incorporated (CUMCFI), the collaborating hospital, in Cagayan de Oro City. The pathology report of the patients is obtained from the hospital's Laboratory Information System (LIS). Only patients who have HER2 cases was to be gathered and it was well noted that the tissue samples of patients were gathered from the previous cases of breast mastectomy and removal of breast mass. All the data gathered were categorized in a table (Table 1) in such a way that it could be easily read.

Table1. *Sample Table of the Training Dataset*

Nuclear Grade	Stage	Tumor Size	Proportion of Positive Nodes	Pleomorphism	Texture of Cell Cytoplasm	Prominent Nucleolus	Observed Necrosis	Her2 Score
3	2	C	Positive	Absent	Not Scanty	No	Absent	0
2	1	B	Negative	Absent	Not Scanty	No	Absent	0
3	2	C	Negative	Absent	Not Scanty	No	Absent	0
0	3	B	Positive	Present	Not Scanty	No	Absent	0
0	1	B	Negative	Absent	Not Scanty	No	Absent	2
3	2	C	Positive	Absent	Not Scanty	No	Absent	2
						No	Present	3

The extracted information from the pathology reports gathered from the Histopathology Laboratory was analyzed. Further analysis in the system's interface and its process were done to correctly output a reliable percentage of likelihood to HER2 between factors. The methods

and techniques used were properly analyzed to prevent mistakes of the system's output.

After the computations the final and pruned tree (Figure 1) was created out of the training dataset. For the training set, only 30 randomly picked sample data were used. Out of the 30 random samples, 14 data had a negative HER2 score and 16 data had a positive HER2 score. This set was the basis for creating a decision tree. With the computation of information gain of each attribute using the mathematical concept of entropy, gini index, and classification error a decision tree pattern was created. This will represent as the general tree pattern during testing to be soon compared during testing using the test data.

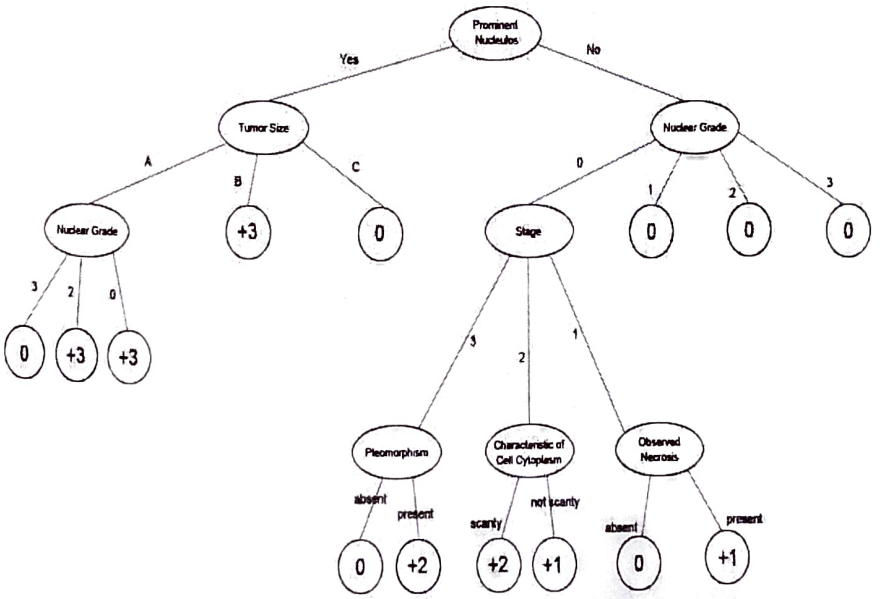

Figure 1. BreCaDS Final Tree (Pruned)

Results

There were 89 randomly picked data samples from surgical pathology reports gathered from the collaborating hospital. These data were divided into two sets. First set was the training data and the second

set was for testing data. During testing, BreCaDS obtained 75% reliability and 25% errors (Figure 2) compared to the actual pathology report results. For instance, out of the 15 random data samples, only 11 got the same result (Table 2) from the original pathology report.

Table 2. *Comparison of Results*

Data Number	Nuclear Grade	Stage	Tumor Size	Proportion of Positive Nodes	Pleomorphism	Texture of Cell Cytoplasm	Prominent Nucleolus	Observed Necrosis	Her2 Score On Pathology Report	BreCaDS Her2 Result
1	3	2	B	Positive	Absent	Not Scanty	No	Absent	0	0
7	3	2	C	Negative	Absent	Not Scanty	No	Absent	0	0
9	2	2	C	Positive	Absent	Not Scanty	No	Absent	0	0
10	2	1	B	Negative	Absent	Not Scanty	No	Absent	0	0
60	0	1	C	Negative	Absent	Scanty	No	Absent	1	
61	0	2	B	Negative	Absent	Not Scanty	No	Absent	1	1
67	0	2	B	Negative	Absent	Scanty	No	Absent	2	2
68	2	1	B	Negative	Absent	Scanty	Yes	Absent	2	
69	0	3	B	Positive	Present	Not Scanty	No	Absent	2	2
71	0	1	B	Negative	Absent	Not Scanty	No	Absent	2	
80	0	2	B	Negative	Absent	Scanty	Yes	Absent	3	3
81	0	2	C	Positive	Present	Not Scanty	Yes	Present	3	
82	0	2	B	Positive	Absent	Scanty	Yes	Absent	3	3
85	0	2	B	Positive	Absent	Scanty	Yes	Present	3	3
88	0	1	B	Negative	Present	Scanty	Yes	Absent	3	3

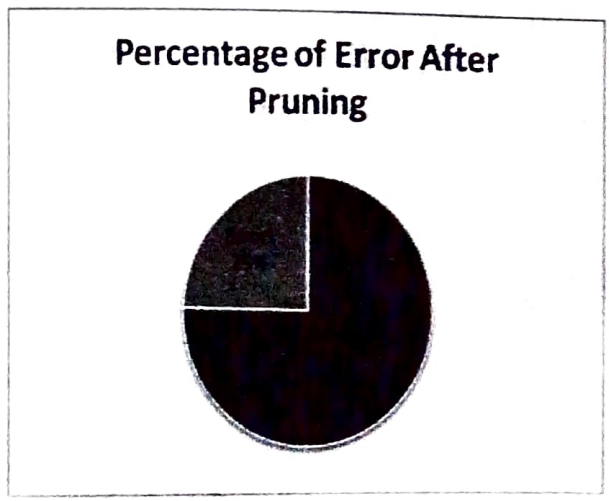

Figure 2. BreCaDS Percentage of Error after Pruning

In error handling, the system applies the theory of pruning, which was an error handling method applied by decision tree algorithm. In testing, BreCaDs relied on the number of data from the surgical pathology reports of patients. BreCaDS, did not need any respondents and the output (Figure 3) of the system highly depended on the number of cases gathered.

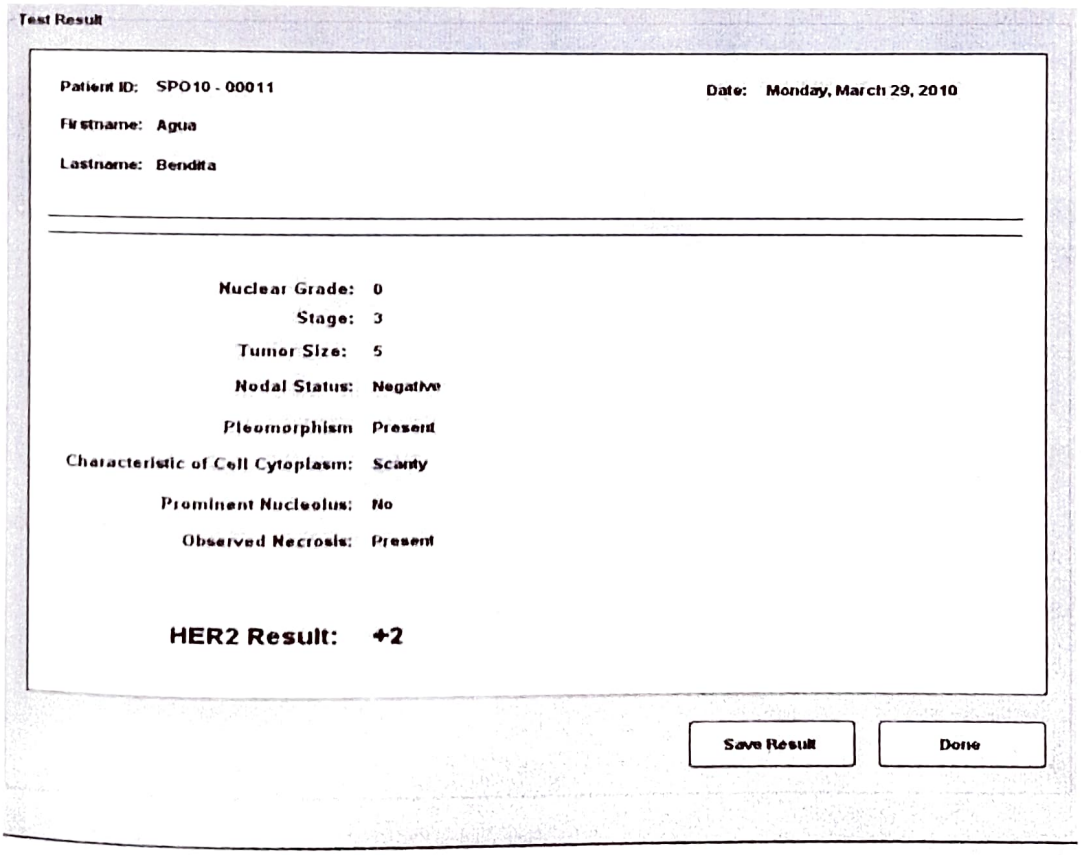

Figure 3. Her2 Score Output of BreCaDS

Conclusion

This paper sought to understand breast cancer behavior and HER2 overexpression development and its underlying probable combinations of underlying factors. The study showed that the BreCaDS could determine factors that were most likely significant to breast cancer aggressiveness and HER2 with 75% reliability. BreCaDS is an automated, fast, and effective system that can give HER2 overexpression result given the eight (8) clinicopathological and cytlogical factors from the patient's data. It has been shown to be reliable in helping clinicians and doctors on providing a better treatment course for breast cancer patients. BreCaDS is the first and pioneering study on breast cancer and HER2. It is recommended that with more data samples the reliability of the system can be improved.

References

Esposito et al., 1997 *Pruning / Decision Trees Algorithm*
[Online]: http://decisiontrees.net/?q=node/44
Accessed date: June 22, 2008

Fyfe, Gwen, M.D. 2006. *Her2-Positive Breast Cancer.* Genentech.
[Online]: http://www.her2genes.com/her2genes/her2.jsp
Accessed Date: July 6, 2008

Kotsiantis, S.B. 2007. *Supervised Machine Learning: A Review of Classification Techniques*
[Online]: http://www.informatica.si/PDF/31-3/11_Kotsiantis%20-%20Supervised%20Machine%20Learning%20-%20A%20Review%20of...pdf
Accessed date: August 27, 2009

Menesses, Danny. 2000. *Ang Hinaharap-Philippine Breast Cancer Network.*
[Online]: http://www.annieappleseedproject.org/anghinphilbr.html
Accessed Date: June 24, 2008

Moore, Andrew. 2006. *Decision Trees.*
Online]: http://www.autonlab.org/tutorials/.
Accessed Date: July 15, 2009

Nashvili, Michael. 2000. *Decision Trees and Data Mining*
[Online]: http://www.decisiontrees.net/
Accessed Date: January 14, 2010

Pruthi, Sandhya,M.D. 2008. *Breast Cancer and Her2.* Mayo Foundation for Medical Education and Research.
[Online]: http://www.mayoclinic.com/health/breast-cancer/AN00495
Accessed Date: July 6, 2008

Sherrod, Phillip. 2001._*Classification and Regression Trees.*
[Online]: http://www.dtreg.com/classregress.htm
Accessed Date: July 15, 20095

Teknomo, Kardi. 2009. *Tutorial on Decision Tree. How Decision Tree Algorithm Work?*
[Online]:http://people.revoledu.com/kardi/tutorial/DecisionTree/how-decision-tree-algorithm-work.htm
Accessed Date: December 12, 2009