Patient Diagnosis of Breast Cancer Using Rule-Based Fuzzy Algorithm for Decision Support System

Mylene Domingo¹, Bobby Gerardo² and Bartolome Tanguilig III³

Abstract—Breast cancer is one of the most common cancers all over the world. Early diagnosis of breast cancer plays a very pertinent role in the survival of the patient. Data mining has a lot of benefits that can help the healthcare organization. The primary focus of the paper is on the algorithm used to identify the class of breast cancer that a patient may have. The type of breast cancer is classified by using a rule-based populated by features computed from a computerized image of a fine clear tissue of the breast mass. The algorithm for Diagnostic Decision Support System using Fuzzy rule based logic will be used to help support specialist to diagnose breast cancer in patients.

Keywords—Breast Cancer, Data mining, Fuzzy Logic, Rule Base.

I. INTRODUCTION

Breast Cancer is the most common invasive cancer among women in the world. The vast majority of breast cancer occurs in females than their male counterpart. It is one of the leading cause of deaths among women in the urbanized countries compared to developing nations. We can attribute some factors related to this notion like life expectancy ratio being one of the reasons. Breast cancer is prevalent among elderly women and women in developed countries live longer than those in poor countries. The different lifestyles of women in rich and poor nations are also believed to be a known contributing factor plus their eating habits. Some of the risk factors in developing breast cancer include: obesity, drinking alcohol, having children late or not having children at all, older age among others. Some of the notable symptoms of breast cancer are usually a thickened mass or tissue in the woman’s breast, change of breast shape, and liquid coming out from the nipple. Records show that the first noticeable symptom of breast cancer is when a woman feels a mass in the breast (lump). While the majority of lumps are not cancerous, it is still best to have it checked by doctors. Prevention is still the best way to address breast cancer. Women may significantly reduce the risk of breast cancer by maintaining ideal weight, being physically fit and active, and breastfeeding. However, there are cases that it is inevitable to prevent it and it leads to death. To reduce the death rate, early detection or diagnosis is needed. Early diagnosis requires a precise and dependable examination process to allow a physician to differentiate a benign from a malignant cancer [1]. An intelligent artificial technique is needed to improve the performance of specialist to diagnose breast cancer. All gathered data with regard to this disease will be taken into account to optimize the solution that they can devise to get more knowledge in battling breast cancer. Group of experts can utilize this data to develop different methods of screening for breast cancer and ultimately arrive at conclusive strategies.

Due to computerization of the records, a substantial amount of raw data is generated by the healthcare organizations [2]. Information Systems from various healthcare organizations gather data about the patient’s records, diagnosis of different illnesses, lab results, resources of the hospitals and others medical related information. Upper management and researchers can use the different extracted data from the system on their decision making. Physicians, nurses and other healthcare professionals are using this clinical information to diagnose diseases and recommend treatment and medication. The healthcare information makes easy for doctors to access relevant information needed for decision-making. Data from healthcare information system are very diverse and challenging to be analyzed and interpreted by traditional tools and techniques. Data mining techniques and methods can be used to help decision makers in discovering fascinating patterns in this complex data[3]. Healthcare data mining involves identification of risk factors associated with breast cancer, and physicians can use this tool to detect early diagnosis of breast cancer of patients. That is how data mining works smoothly in collating information and aids our doctors and medical specialists in formulating new strategies to combat this disease.

Physicians make diagnostic decisions and treatment recommendations based on its medical records. Health informatics allows doctors to have a faster access to more relevant information and thus makes more optimal decisions. Applying data mining in electronic medical records will give doctors analytical and predictive tools on the surface of the data. Traditional and modern approach can be combined using the collated records in arriving at near accurate

¹, ³ Technological Institute of the Philippines Graduate Programs, Quezon City, Philippines
² West Visayas State University, Institute of Information and Communication Technology, Iloilo Philippines.
decision.

Experts can utilize as a support tool the extracted and analyzed data from the hospital information system. Data mining has a lot of mechanisms and techniques that can be applied to process data and eventually discover some hidden patterns. Based on the patients’ profile, history, physical examination, diagnosis and utilizing previous treatment patterns, new treatment plans can be potently suggested [4]. This paper aims to detect if breast cancer is very dangerous or benign by using Rule-Based Fuzzy Logic. The algorithm will be using three-stage methodology 1) generation of decision tree from the data 2) setting rules from the decision tree 3) transformation of regulations into fuzzy model.

II. REVIEW OF RELATED LITERATURES

Data mining is considered as an effective tool in diagnosis and treatment of many diseases like cancer, heart related diseases, diabetes, skin and other [2]. The health care sector collects sources of data huge amount of data to be mined and they used this data to discover patterns which will help them in doing decision making. This gives the idea of using data mining as a tool in diagnosing cancer in patient specifically on breast cancer.

A fuzzy rule based was used in the study of lung diseases diagnostic system by combing the positive and the negative knowledge. This was designed to determine different or multiple diseases of lung cancer in a single patient [12]. The use of Rule-based method is one of the objectives of the study. Based on the experience of the physicians, a set of rule was generated that will aid in the diagnoses of class of tuberculosis. This basically contributes and gives the researchers the idea on what method to use in the study.

In diagnosing Coronary Artery Disease, a rule-based DSS was used. Initially, a set of rules is constructed from a decision tree, induced by the training records. Then the resulted set of rules is transformed to a fuzzy model and finally the parameters of this fuzzy model are optimized [5]. This study gives the idea of using the fuzzy model with decision tree integration in the diagnosis of breast cancer. There is a similar study on fuzzy diagnostic system for coronary disease (CAD), where the researcher claimed that this disease affects millions of people every year which then leads to increase in mortality rate [11]. A use of information technology on this paper reduces the mortality rate and waiting time to see the oncologists. The researcher emulates the human intelligence. In the diagnosis of coronary artery disease, a fuzzy rule-based decision support system (DSS) is presented. From a novel data-driven four-stage methodology, the system is automatically generated. Initially, a decision tree was used to construct the set of rules. They extract a crisp rule from the decision tree. Then the model is transformed to a fuzzy, and they used this as parameters of fuzzy model optimization [10]. This study greatly influenced to use decision tree and fuzzy rule to be used in the early diagnosis of breast cancer.

A rule-based fuzzy logic Diagnostic Decision Support System was designed to help pulmonary physicians analyze the class of tuberculosis the patient has. In this system, the physician will input corresponding scores in each symptom which imply the intensity of every symptom the patient exhibits and its intensity varies depending on the condition. Then the system will summarize and allocate the class of tuberculosis the patient has [6]. This study helps in the development of a concept of using a fuzzy rule-based to determine whether the breast cancer is malignant or benign.

The use of fuzzy logic was used to identify the diseases from the symptoms that help to develop Fuzzy rules and this knowledge are stored and fired during further decision-making [7]. The processes are exhibited: The physician gathers knowledge about the patient from historical record, physical examination, laboratory results and other investigative procedures. The researcher uses the UCI dataset as a source of data for breast cancer diagnosis.

III. RESEARCH MODEL

Oncologists are the main recipients of this Rule-Based Fuzzy Decision Support System. The system will examine the type of breast cancer of the patient based on the tiny nuclei found in the image. They compute the features based on a computerized image of the aspirated breast mass. Fine Needle Aspiration (FNA) operation is a clinical procedure used to investigate the diseased spot of the breast [8]. They describe in the list the characteristics of the nuclei present in the image. In this regard, tissues obtain from the process has become an alternative way to breast cancer lesion. This breast cancer databases was obtained from the University of Winsconsin Hospitals, Madison from Dr. William H. Wolberg.

![Algorithm for Decision Support System for Breast Cancer](image)

Fig. 1: Research Model

A Decision Tree with a Fuzzy Logic implementation was used to develop the algorithm of the system. With fuzzy logic, the researchers were able to determine the class of breast cancer that the patient has.

To generate the fuzzy rule-based decision support system, a decision tree is used first to determine the number of rules to generate. Fuzzification is applied to a binary decision tree to convert the sharp transition of the nodes of the tree into a gradual one. Figure 1 show the model use on this research study.

The fuzzy system will use the following type of fuzzy rules: [9]:

Rule Rj: If X1 is Aj1 and... Xn is Ajn then Class Cj with CF = CF
Where Rj is the label of the jth fuzzy if-then rule, Aj1, Ajn are antecedent fuzzy sets on the unit interval [0,1], Cj is the consequent class, and the CFj is the grade of certainty of the fuzzy if-then rule.

In this research work, FNA biopsy data has been obtained from UCI data set. UCI Machine Learning Repository is a collection of databases, domain theories and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

Specifically, the study used the Wisconsin Breast Cancer Dataset. They obtained this breast cancer database from the University of Wisconsin Hospitals, Madison from Dr. William [6]. WEKA was used to analyze the dataset. Weka is an open source software under the GNU General Public License.

In the first stage of the methodology, a decision tree was constructed on the set of data to minimize the expected value of the number of tests for classification. Each internal node of the tree corresponds to a principal component, and the rest of the outgoing branch corresponds to a possible range of that component. The leaf node corresponds to the class of breast cancer a patient has.

Data set composed of 699 numbers of instances and ten attributes plus the class attribute. The data includes 458 counts from benign cases and 241 instances belong to malignant cases. Ten features are measured per patient to determine the degree of the diseased spot. Below is the class distribution of the types of breast cancer.

**Class Distribution**

- **Benign:** 458 (65.5%)
- **Malignant:** 241 (34.5%)

**Attributes information**

<table>
<thead>
<tr>
<th># Domain</th>
<th>Sample code number id number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clump Thickness</td>
<td>1 - 10</td>
</tr>
<tr>
<td>2. Uniformity of Cell Size</td>
<td>1 - 10</td>
</tr>
<tr>
<td>3. Uniformity of Cell Shape</td>
<td>1 - 10</td>
</tr>
<tr>
<td>4. Marginal Adhesion</td>
<td>1 - 10</td>
</tr>
<tr>
<td>5. Single Epithelial Cell Size</td>
<td>1 - 10</td>
</tr>
<tr>
<td>6. Bare Nuclei</td>
<td>1 - 10</td>
</tr>
<tr>
<td>7. Bland Chromatin</td>
<td>1 - 10</td>
</tr>
<tr>
<td>8. Normal Nucleoli</td>
<td>1 - 10</td>
</tr>
<tr>
<td>9. Mitoses</td>
<td>1 - 10</td>
</tr>
<tr>
<td>10. Class</td>
<td>(2 for benign, 4 for malignant)</td>
</tr>
</tbody>
</table>

The second stage is the construction of rule-based classifier. It is a technique used to classify different records using a collection of “if…then” rules. There are 20-rules generated. Below are the rules generated based on the decision tree induction:

if uniformity cellsize = 3 AND bare nuclei = 5 then class = 4
if uniformity cellsize = 3 AND bare nuclei = 6 then class = 2
if uniformity cellsize = 3 AND bare nuclei = 7 then class = 4
if uniformity cellsize = 3 AND bare nuclei = 8 then class = 4
if uniformity cellsize = 3 AND bare nuclei = 9 then class = 4
if uniformity cellsize = 3 AND bare nuclei = 10 then class = 4
if uniformity cellsize = 4 AND bare nuclei = 4 then class = 4
if uniformity cellsize = 4 AND bare nuclei = 5 then class = 4
if uniformity cellsize = 4 AND bare nuclei = 6 then class = 2
if uniformity cellsize = 4 AND bare nuclei = 7 then class = 4

The last and the final stage is the transformation of a set of rules into a fuzzy model. There are 8-main rules generated out of 20 rules above. A set of rules is combined to provide a fewer number of rules to use with the same degree of accuracy and reliability.

Below are the 8-main rules generated:

if uniformity cellsize = 1 AND uniformity cellsize <= 2 then class = 2
if uniformity cellsize = 3 AND bare nuclei = 1 AND bare nuclei <= 2 then class = 2
if uniformity cellsize = 3 AND bare nuclei = 3 AND sample code number <= 1170945 then class = 4
if uniformity cellsize = 3 AND bare nuclei = 3 AND sample code number > 1170945 then class = 2
if uniformity cellsize = 3 AND bare nuclei = 4 AND bare nuclei <= 5 then class = 4
if uniformity cellsize = 3 AND bare nuclei = 6 then class = 2
if uniformity cellsize = 3 AND bare nuclei = 7 AND bare nuclei <= 10 then class = 4

**IV. CONCLUSION AND RECOMMENDATION**

In this paper, it presented the use of rule-based fuzzy to diagnose the class of breast cancer that a patient has based on a computerized image of a fine clear tissue of a breast mass. There are two types of breast cancer known as benign or malignant. The proposed method is applied to Breast Cancer Wisconsin Diagnostic data set and was able to classify the entirely given sample on classes either malignant or benign.

Three different stages are used to create the algorithm. The first stage is the induction of decision tree from the given data set. Then they transformed into a fuzzy model the rule that they derive from the result of the decision ladder.

Rules are created based on the twenty guidelines generated from the decision tree. The set of rules created are then compressed and were used to developed the fuzzy model.

To develop a computerized decision support system for diagnosis of breast cancer class, this model can be used. It would greatly help in the decision making of the oncologist in giving the diagnosis.

Future researchers could use the features of the model and make the model more usable and probably use this to give corresponding treatments for patients.
ACKNOWLEDGMENT

Firstly, I would like to express my sincere gratitude to my adviser Dr. Bobby D. Gerardo for his continuous support of my Doctorate study and related research, for sharing his expertise and immense knowledge.

I am taking this opportunity as well to thank the Technological Institute of the Philippines most especially Dr. Bartolome Tanguilig III and Mr. Alexa Ray Fernando, for providing me with all the necessary facilities for the research. It would not be possible to conduct this research without their unwavering support.

Last but not the least, I would like to thank my family: my husband, my parents and my brother and sisters for supporting me throughout writing this thesis.

REFERENCES


Ms. Mylene J. Domingo finished her Bachelor of Science in Computer Science at the University of the East, Manila, Philippines in 2001 and her Master of Science in Information Technology at University of La Salette, Philippines in 2009. She is currently pursuing his Doctor in Information Technology at the Technological Institute of the Philippines, Quezon City, Philippines. She is currently the Department Head of the Department of Information Systems at the Institute of Information and Computing Sciences at University of Santo Tomas, Manila, Philippines. She is under the Business Analytics Track in Information Systems Department. Her research interests are in data mining, e-learning and algorithm development.

Dr. Bobby D. Gerardo is currently the Vice President for Administration and Finance of West Visayas State University, Iloilo City, Philippines. His dissertation: Discovering Driving Patterns using Rule-based intelligent Data Mining Agent (RiDAMA) in Distributed Insurance Telematic Systems. He has published more than 60 research papers in national and international journals and conferences. He is a referee to international conferences and journal publications such as in IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Knowledge and Data Engineering. He is interested in the following research fields: distributed systems, telematics systems, CORBA, data mining, web services, ubicots computing and mobile communications.

Bartolome T. Tanguilig III took his Bachelor of Science in Computer Engineering at Pamantasan ng Lungsod ng Maynila in 1991. He finished his Master’s Degree in Computer Science from De La Salle University, Manila, Philippines, in 1999, and his Doctor of Philosophy in Technology Management from Technological University of the Philippines, Manila, in 2003. He is currently the Assistant Vice-President for Academic Affairs and concurrent Dean of the Graduate Programs of Technological Institute of the Philippines, Quezon City. Dr. Tanguilig III is a member of the Commission on Higher Education (CHED) Technical Panel for IT Education (TPITE), the chair of the CHED Technical Committee for IT (TCT), the founder of Junior Philippine ITE Researchers (JUPITER), Vice President – Luzon of the Philippine Society of IT Educators (PSITE), board member of the PCS Information and Computing Accreditation Board (PICAB), member of the Computing Society of the Philippines (CSP), and a program evaluator/accreditor of the Philippine Association of Colleges and Universities Commission on Accreditation (PACUCUA).