Prediction of Breast Cancer Risk Level with Risk Factors in Perspective to Bangladeshi Women using Data Mining

Kawsar Ahmed

Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Bangladesh

Md. Ahsan Habib

Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Bangladesh

Tasnuba Jesmin

Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Bangladesh

Md. Zamilur Rahman

Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Bangladesh

Md. Badrul Alam Miah

Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Bangladesh

ABSTRACT

The upgraded and modern medical technologies are the most challenging task to detect cancer and provide accurate treatment. In Bangladesh about two million women are affected by 2nd most occurring deathful breast cancer due to them and their family member's unconsciousness and poverty. It requires about $400-500 for proper diagnosis and treatment. Most of the Bangladeshi women are uneducated and feel shy with society or husband to go doctor for checking breast cancer. So it also will be a good achievement of this work to find breast cancer with more efficiency. Breast cancer depends on some risk factors that may help to detect breast cancer using multi-layered approach. In this work, at first it is collected 100 peoples' information which consist of both cancer and non-cancer information having missing or duplicate information. So pre-processing and K-means clustering methods are performed to separate relevant and non-relevant data to Breast Cancer. Then risk factors are ranked using WEKA tools and are assigned a score according to rank. Finally, it is implemented an application software using Lotus Notes to predict Breast Cancer risk level which is easier, effective, efficient, secured, cheap and time saving with some suggestions. This technique will contribute equal opportunity to the underdeveloped and developing countries to detect, diagnosis, and treatment of breast cancer.

General Terms

Computer Science, Data Mining, Breast Cancer in Bangladesh

Keywords

Breast cancer in Bangladesh, Public health, Data mining, Risk factors of breast cancer, WEKA toolkit, Woman Health Conditions

1. INTRODUCTION

Cancer is a class of most deathful diseases caused by out-of-control cell growth. There are more than 100 different types of cancer, and each is classified by the type of cell that is initially affected. Cancer harms the body when damaged cells by dividing uncontrollably. Cancer is treated as most harmful disease in the world because most of the patients died when affected properly. All of cancers breast cancer is the most in Bangladesh and 2nd in the world deathful disease. Breast cancer is a cancer that begins in the tissues of the breast. There are two types of Breast cancer one is Ductal carcinoma and another is Lobular carcinoma. Ductal carcinoma begins in the tubes that move milk from the breast to the nipple. Most breast cancers are of this type. It may progress to invasive cancer if untreated. Lobular carcinoma starts in the parts of the breast, called lobules, which produce milk.

World Health Organization (WHO) says that the death rate of breast cancer is high in Bangladesh and ranked Bangladesh 2nd deathful country in the world for breast cancer of women. More than 89% of Bangladeshis people are Muslim, ranked 3rd in the world after Indonesia and Pakistan. Approximately 45% of the population is related in the agricultural sector. Bangladesh is a low-income country, defined by the World Bank as countries with a Gross National Income (GNI) less than US $1,005 per capita. About 40% of the population is underemployed; many participants in the labour force work only a few hours a week, at low wages. Approximately 60% of women are illiterate, and 27% of the population is undernourished. In Bangladesh only one doctor give service more than 3,300 people for urban areas and more than 15,000 people for rural areas [1]. So it is a burning question of Bangladesh about general population health.

Most of the women don't know about breast cancer before affected for a long period of time. Many of them suffered from husband as well as society. They believe that breast cancer is caused by an evil spirit, or as a punishment for bad deeds. Even some families give women very low priority. So it's need to increase the awareness among all population both rural and urban areas in Bangladesh as well as increase the number of doctors both areas.

It is estimated that each year, 76,000 women die of breast cancer in South Asia (India, Bangladesh, Nepal, Myanmar, Pakistan, and Tibet) [2]. In Bangladesh, there is no national cancer registry but estimate an annual new breast cancer case burden of 30,000 women. It is projected that global breast cancer cases will grow from 1.4 million in 2008 to over 2.1

million cases in 2030 [3]. Breast Cancer Incidence and Mortality in Bangladesh will reach an unexpected level. So it's time to reduce number of breast cancer patient providing new technology day after day. At early stage breast cancer have no symptoms. But after developing or advanced breast cancer stage some symptoms. Like

- Bone pain
- Breast pain or discomfort
- Skin ulcers
- Swelling of one arm
- Weight loss
- Nipple discharge.
- Change in shape or appearance of the nipple

According to breast cancer stage there are different kinds of treatment. Treatment for breast cancer depends on the type and stage of the disease, the size the tumor, general health, medical history and age of patients. In most cases, the goal of treatment is to remove or destroy the cancer completely. Most breast cancers can be cured if found and treated early. Like other diseases treatment, breast cancer treatment has also some side effects. It may damages healthy cells and tissues, unwanted side effects sometimes occur. Side effects depend mainly on the type and extent of the treatment. Side effects may not be the same for each person. So it tells again that "Prevention is better than cure". And this paper will be helpful to maintain the moral sentence.

In general, cancer treatments may include:

☐ Chemotherapy medicines to kill cancer cells;
☐ Radiation therapy to destroy cancerous tissue; and
☐ Surgery to remove cancerous tissue

But it need more money that not bearable to low income people of developing country. So it will be a good work to introduce a system that can detect breast cancer with little amount and quickly. By thinking those, this paper introduces a system that can detect breast cancer quickly and need very low money.

2. BACKGROUND

Breast cancer, second deathful disease in Bangladesh can detect or predict by analysing electrical conical images, cancer region images or treatment information. Like Skin Cancer [4], Lung Cancer [5], Brain cancer [6] and Diabetics [7], this research paper also introduces to predict Breast cancer using treatment information. That treatment information has been collected according to Bangladesh and Bangladeshi patients. Treatment information is described briefly in below section.

Determining what causes cancer is complex. Many things are known to increase the risk of cancer, that are called risk factors of that cancer, including tobacco use, certain infections, radiation, lack of physical activity, obesity, and age. If any person is aware about these risk factors then cancer can prevent easily. Like of other cancers breast cancer also depends on some factors where some are most important and some are less important. Also some are editable and some are not editable. In this section discussed about the risk factors of breast cancer as well as their level of importance.

Age and gender is the most important factors of causing breast cancer. Compare to men, women are mainly affected by this cancer and developing of breast cancer risk is increased according to becoming older. In Bangladesh more than 50 years old people are mainly affected by breast cancer. Unmarried younger girl are also affecting rapidly by this

cancer because of their unconsciousness. This can be removed or prevented by maintaining some rules.

A person who was affected by any kind of cancer has an increased risk of developing another cancer of any type [6]. A person who has two or more close relatives (mother, father, sister, brother, or child) who are responsible for developing breast cancer has a risk factor of developing Breast cancer for his own. Rarely, members of a family will have an inherited disorder that makes the breast cell also affected and increases the risk of Breast cancer.

Sometimes some people develop breast cancer by genes (mostly found BRCA1 and BRCA2 genes). These genes normally produce proteins that protect us from cancer. If a parent passes you a defective gene, you have an increased risk for breast cancer. Women with one of these defects have up to an 80% chance of getting breast cancer sometime during their life.

Women who got their periods early (before age 12) or went through menopause late (after age 55) have an increased risk for breast cancer. These 2 risk factors are high risk factor of occurring breast cancer.

Women who have never had children or who had them only after age 30 have an increased risk for breast cancer. Being pregnant more than once or becoming pregnant at an early age reduces your risk of breast cancer.

If any person received radiation therapy as a child or young adult to treat cancer of the chest area, they have a much higher risk for developing breast cancer. A person who has taken any therapy is responsible for occurring different kinds of disease compare to other people who don't take any therapy [2], [6]. Women who took diethylstilbestrol (DES) to prevent miscarriage may have an increased risk of breast cancer after age 40.

Abortion is other most important risk factor of breast cancer. Taking alcohol regularly increase the risk level of breast cancer. Drinking alcohol increases the risk of breast cancer. Experts recommend no more than one alcoholic beverage a day for women.

Obesity is not only responsible for diabetics or breast cancer but also responsible for other diseases forcing directly or indirectly. Obesity is a condition in which a person has an abnormally high and unhealthy proportion of body fat. To measure obesity mainly used body mass index (BMI) that calculated by dividing a person's weight (in kilograms) by their height (in meters) squared. It provides a more accurate measure of obesity or being overweight than weight alone.

The diagnosis of Breast Cancer is a tedious and significant task. Advances in breast cancer treatment are facilitated by the knowledge gained from tumor biopsy and pathology. These techniques allow doctors to develop individual treatments for each patient based on their distinctive tumor biology. New techniques are constantly being researched to improve the detection, diagnosis, and treatment of breast cancer likely brain cancer, skin cancer [4], Lung cancers [5], Diabetics [7].

The treatment of breast cancer risk factors can be broken down into two categories: removable and irremovable risk factors. Removable risk factors are those risk factors that can be solved by consciousness or taking some rules. Irremovable risk factors are those risk factors that can't solve but keep those factors unworkable. Table 2 shows the categories of breast cancer risk factors.

A widely recognized formal definition of data mining can be defined as "Data mining is the non- trivial extraction of implicit previously unknown and potentially useful information about data" [8]. Data mining has some fields to analysis of data such as classification, clustering, correlations, association rule etc. Now-a-days data mining has been used intensively and extensively by many organizations. In-Healthcare, data mining is becoming increasingly popular [9]. Data mining provides the methodology and technology to analysis the useful information of data for decision making. Pre-processed data are broken down into two categories: relevant and non-relevant dataset to breast Cancer. Clustering technique mainly used to find out two types of data. Clustering is a process of separating dataset into subgroups according to the unique feature. WEKA toolkits are mainly used to find out highest affected risk factors as well as ranking of risk factors from dataset.

Day by day the number of Breast cancer person is increasing rapidly because of unconsciousness about breast cancer and risk factors of breast cancer. The main goal of this research is to develop a system that can be used by a person for testing his/her breast Cancer risk level and provides suggestions according to his/her breast cancer risk factor.

3. METHODOLOGY

3.1 Data Collection and Pre-processing

100 persons' data is collected from different diagnostic centre where only female are present of age 18 to 70. Obtained data also contains both patient and non-patient information. From the previous studies 20 risk factors were considered for Breast cancer assessment in Bangladeshi population, which includes-age, gender, hereditary, previous health examination, use of anti-hypersensitive drugs, late Menopause, early Menarche, food habit, obesity, genetic risk, environment, excessive alcohol, hormone therapy, abortion ,exercise, Depression, marital status, outdoor activities and affected any cancer before.

3.2 Pre-processed of Obtain Data

Due to missing and duplicate information the collected data are pre-processed by avoiding duplicate and supplying missing information. It mainly reduces the memory and normalizes the values used to represent information in database. And after analysis which factors are not link with breast cancer assign risk value to zero (0) and others are link with breast cancer assign to risk value one (1). Now classify the factors into 2 groups one contains relevant data to Breast Cancer who has risk value 1 and another contains remaining data that means who has risk value 0.

The process of partitioning and category of collected data into different subgroups where each groups have a unique feature is called clustering [10]. The goal of clustering is to classify objects or data into a number of categories or classes where each class contains identical feature. The main benefits of clustering are that the data object is assigned to an unknown class that have unique feature and reduces the memory. The K-means clustering [11] is a widely recognized clustering tool that is used for robotics, diseases and artificial intelligence application purposes [12]. Here k is a positive integer representing the number of clusters. The pre-processed data is clustered using the K-means clustering algorithm with the value of k equal to 2. This represents there is two clusters where one cluster contains relevant data to Breast Cancer and another contains remaining data.

3.3 Discover Most Responsible Factors

Discover most responsible of risk factors is known as the principle data mining problem that intends to find out the frequent items from the data warehouse. Here CFS Subset Evaluator Including locally predictive attributes in WEKA toolkit is used to discover the highest responsible factors of its efficiency and simplicity than other algorithms where relevant data in CSV format to Breast Cancer are used as input of this algorithm.

3.4 Ranking Risk Factors

Here Chi-squared Ranking Filter is used as Attribute Evaluator and Ranker is used as search method in WEKA to rank 13 included risk factors because of its efficiency and simplicity than others.

4. RESULT

The experimental results are categorized into two sections. One is ranking factors discover and another is represents prediction tools to Breast Cancer.

4.1 Results for Ranking Risk Factors

The ranking of Breast cancer risk factors and highest responsible factors of causing breast cancer are shown in Fig. 1 and Fig. 2 respectively. Here we assign lowest factors score as 1 and highest score as 4. Each factor is subdivided into different sectors and assigns a different score that is presented in Table 1 that is used for predicting Breast cancer.

4.2 Results for Prediction to Breast Cancer

Finally using Table 1 the prediction tool to breast Cancer is implemented that is shown in Fig. 3. Fig. 4 represents the risk level (flowchart) of breast Cancer which is implemented using Table 1. In Table 1 weightage score high means high impact on causing breast cancer. Fig. 3 is principle graphical representation and Fig. 5-6, 8-9 is the graphical representation of Breast cancer different risk levels as example.

```
=== Attribute Selection on all input data ===

Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.078

Attribute Subset Evaluator (supervised, Class (nominal): 13 Patient):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 3,4,5,7,9 : 5
                Married?
                Menarche started early?
                Menopause stops at late ?
                Fatty foods Habits?
                Hormone Therapy Taken?
```

Fig 1: Most affected risk factors list using WEKA

5. COMPARE TO EXIT TECHNIQUES

Through Mammography [12] breast cancer can detect but it need more time, cost and increase radiation problem because

of X-ray. MRI has a role at high risk for breast cancer of women. Latest diagnostic techniques such as sestamibi scans, optical imaging and molecular diagnostic techniques [13] look promising, but need more investigation and time into their use. PET/computed tomography [14] a role in detecting local disease recurrence and distant metastasis in breast cancer patients but not time savable as well as cost. But this paper introduces a new technique that is easy, secure, time savable as well as cost savable.

6. CONCLUSION

Early detection of breast cancer is imperative for improving patient survival. Breast cancer ranks as one of the leading cancer types in the number of new cases diagnosed and is second only to lung cancer as the most prevalent cause of cancer death in women. For treatment it needs more money that is not bearable for developing countries' people like Bangladeshi people. Due to Bangladesh our technique is constantly being researched to improve the detection, diagnosis, and treatment of breast cancer.

Table 1. Factors corresponding Weightage & Score

Parameters	Weight age	Score
Age	Age≤24	1
	24<Age≤38	2
	38<Age≤50	3
	Age>50	4
Married?	Yes	1
	No	2
Menarche started early (before 12 years)?	No	1
	Yes	4
Affected Family members?	No	1
	Uncle, Aunty	2
	Brother, Parents etc.	5
BMI?	BMI≤24	1
	25≤BMI≤28	2
	BMI≥29	3
Menopause stops at late (after 53 years)?	Yes	4
	No	1
Fatty foods Habits?	No	1
	Yes	2
Alcohol Use?	Yes	3
	No	1
Hormone Therapy in Breast Area?	Yes	3
	No	1
Affected any cancer before?	Yes	4
	No	1
Abortion	Yes	3
	No	1
Exercise Regularly?	Yes	1
	No	2

Table 2. Removable and irremovable risk factors

Removable Risk Factors	Irremovable Risk Factors
1. Marital Status; 2. BMI; 3. Early Menarche; 4. Late Menopause; 5.Hormone therapy; 6. Alcohol Use; 7. Fatty foods; 8. Abortion and 9. Regular Exercise	1. Age; 2. Sex; 3. Affected family members; 4. Affected any cancer before;
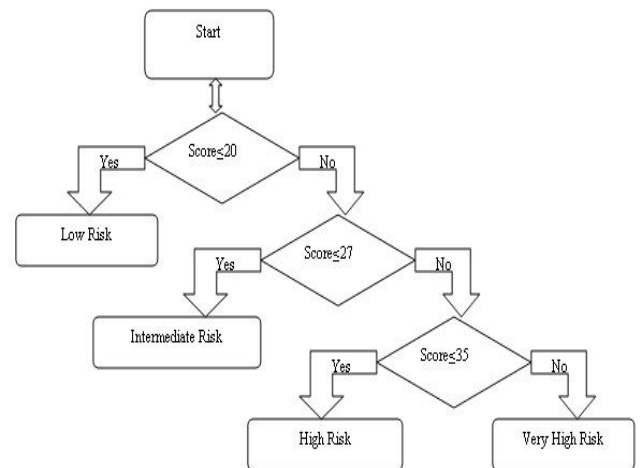
Fig 4: Flowchart Breast cancer risk prediction system


```
Attribute selection output

=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 13 Patient):
        Chi-squared Ranking Filter

Ranked attributes:
 9.0909    9 Hormone Therapy Taken?
 3.4749    7 Fatty foods Habits?
 3.3955    2 Affected family Members?
 2.6902   12 Alcohol?
 2.6864    5 Menopause stops at late ?
 1.9726    4 Menarche started early?
 1.5444    3 Married?
 0.5255   11 Abortion?
 0.3653    6 Exercise?
 0.0507   10 cancer before
 0         1 Age
 0         8 BMI?
```

Fig 2: Ranking of risk factors using WEKA

Fig 3: Graphical view of Breast cancer risk prediction

Fig 5: Graphical view of Breast Cancer Prediction with Low Risk Level

Fig 6: Graphical view of Breast Cancer Prediction with Intermediate Risk Level
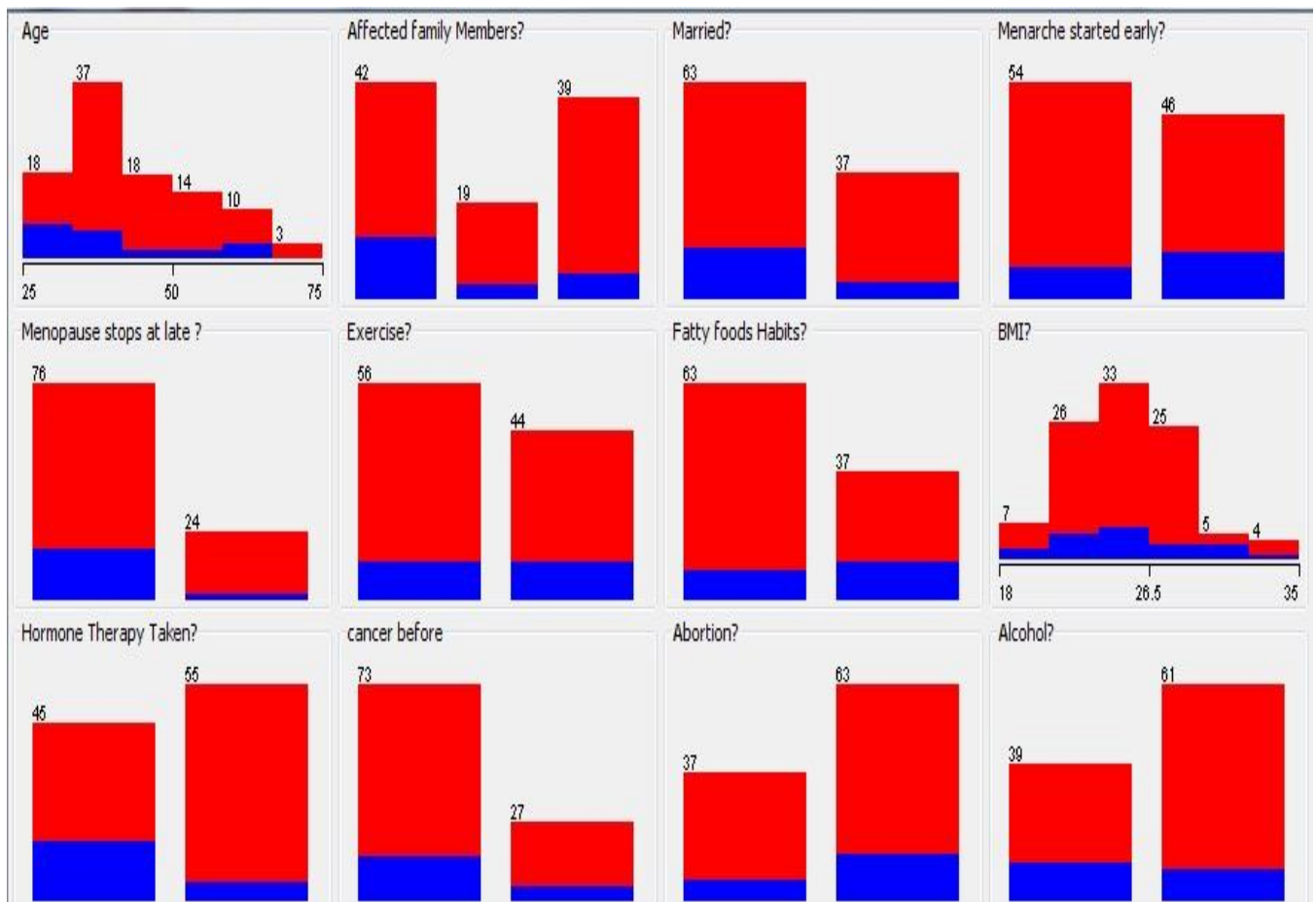
Fig 7: Visualization of Breast Cancer data where Red & Blue color represent patient & non-patient respectively

Fig 8: Graphical view of Breast Cancer Prediction with High Risk Level

Fig 9: Graphical view of Breast Cancer Prediction with Very High Risk Level

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] H. L. Story, R. R. Love, R. Salim, A. J. Roberto, J. L. Krieger, O. M., "Improving Outcomes from Breast Cancer in a Low-Income Country: Lessons from Bangladesh", International Journal of Breast Cancer, 2012, pp. 1-9. Article ID 423562.

[2] M. Brown, S. Goldie, G. Draisma, J. Harford, J. Health service interventions for cancer control in developing countries in Disease Control Priorities in Developing Countries. Oxford University Press, 2006, vol. 2, pp. 569-90.

[3] World Cancer Research Fund. Available from: http://www.wcrf.org/cancer_facts/women-breast-cancer.php/. Accessed March 10, 2013.

[4] K. Ahmed, T. Jesmin, Md. Z. Rahman, "Early Prevention and Detection of Skin Cancer Risk Using Data Mining" International Journal of Computer Applications, 2013, vol. 62, pp. 1-6.

[5] K. Ahmed, Abdullah-Al-Emran, T. Jesmin, R. F. Mukti, Md. Z. Rahman, F. Akter, "Early Detection of Lung Cancer Risk Using Data Mining", Asian Pacific Journal of Cancer Prevention, 2013, vol. 14 pp. 595-98.

[6] T. Jesmin, K. Ahmed, Md. Z. Rahman, Md. B. A. Miah, "Brain Cancer Risk Prediction Tool Using Data Mining", International Journal of Computer Applications, 2013, vol.61, pp. 22-27.

[7] K. Ahmed, T. Jesmin, U. Fatima, Md. M., Abdullah-al-E., Md. Z. Rahman, "Intelligent and Effective Diabetes Prediction System Using Data Mining Approach", ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY, 2012, vol. 5, pp. 215-21.

[8] Frawley, Piatetsky S., "Knowledge Discovery in Databases: An Overview", the AAAI/MIT Press Menlo Park C.A.

[9] H. C. Koh, G. Tan, "Data Mining Applications in Healthcare" Journal of Healthcare Information Management, vol. 19, pp. 64-72.

[10] Z. Nouir, B. Sayrac, B. Fourestié, W. Tabbara, F. B., "Generalization Capabilities Enhancement of a Learning System by Fuzzy Space Clustering" Journal of Communications, 2007, vol. 2, pp. 30-7.

[11] C. Ordonez, "Programming the K-Means Clustering Algorithm in SQL", in Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining, 2004, pp. 823-8.

[12] Daniel B. K., "Early Breast Cancer Detection Using Techniques other than Mammography", American Journal of Roentgenology, 1984, pp. 465-8.

[13] Medical Center University of Bonn, "Improving Methods for Breast Cancer Detection and Diagnosis", National Cancer Institute, 2003.

[14] S. V, S. C, W. L, Bourke A., "New diagnostic techniques for breast cancer detection" US National Library of Medicine National Institutes of Health, 2008, vol. 4, pp. 501-13.