Diagnosis of Breast Cancer using Decision Tree Data Mining Technique

Ronak Sumbaly
Department of Computer Science
BITS, Pilani – Dubai
United Arab Emirates

N. Vishnusri
Department of Electronics & Communication
BITS, Pilani - Dubai
United Arab Emirates

S. Jeyalatha
Department of Computer Science
BITS, Pilani – Dubai
United Arab Emirates

ABSTRACT
Cancer is a big issue all around the world. It is a disease, which is fatal in many cases and has affected the lives of many and will continue to affect the lives of many more. Breast cancer represents the second primary cause of cancer deaths in women today and has become the most common cancer among women both in the developed and the developing world in the last years. 40,000 women die in a year from this disease, which is one woman every 13 minute dying from this disease everyday.

Early detection of breast cancer is far easier to cure. This paper presents a decision tree based data mining technique for early detection of breast cancer. Breast cancer diagnosis differentiates benign (lacks ability to invade neighboring tissue) from malignant (ability to invade neighboring tissue) breast tumors. This paper also discusses various data mining approaches that have been utilized for breast cancer diagnosis, and also summarizes breast cancer in general (types, risk factors, symptoms and treatment).

General Terms
Breast Cancer, Data Mining, Decision Tree.

Keywords
Benign, BRCA, Breast Cancer, Carcinoma, Data Mining, Decision Tree, J48, Malignant, Survivability Rate, Tumor.

1. INTRODUCTION
Thousands of grandmothers, mothers, daughters fall victim to breast cancer every year. The human body comprises of millions of cells each with its own unique function. When there is unregulated growth of any of these cells it is termed as cancer. In this, cells divide and grow uncontrollably, forming an abnormal mass of tissue called as tumor. Tumor cells grow and invade digestive, nervous and circulatory systems disrupting the bodies' normal functioning. Though every single tumor is not cancerous.

Cancer is classified by the type of cell that is affected and more than 200 types of cancers are known. This paper is focused on Breast cancer. Breast cancer is the most common type of cancer among females across the world [2]. Recent years have seen an intense improvement in survival rates for women with breast cancer, which can be mainly attributed to an extensive screening and enhanced treatment.

The recent advances in data collection and storage techniques have made it possible for various medical companies and hospitals to keep vast amounts of data relating to their medical records pertaining to medication and symptoms of a disease. Formally, data mining is the process of running powerful algorithms on data to extract useful information. The uses and potentials of these methodologies have found its scope in medical data.

Predicting outcome of a disease is a challenging task. Data mining techniques tends to simplify the prediction segment. Automated tools have made it possible to collect large volumes of medical data, which are made available to the medical research groups. The results being an increasing popularity of data mining techniques to identify patterns and relationship among large number of variables, which make it possible to predict the outcome of the disease using pre-existential datasets. This paper presents the potential synergies between data mining techniques and breast cancer diagnosis.

Structure of the Paper
The paper is organized as follows: In Section 2, the objective of this paper is presented. In Section 3, the various acronyms being used in this paper have been given. An overview of breast cancer is presented in Section 4. Section 5 deals with the methodology involved (with analysis) in breast cancer diagnosis using 'Decision Trees'. The paper is concluded in Section 7, with future of breast cancer diagnosis in section 6.

2. OBJECTIVES
The present work is intended to meet the following objectives:

1. Summarize Breast cancer – Types, Risk Factors, Symptoms, Diagnosis and Treatment.
2. Demonstrate Decision Tree – Data mining technique for Breast cancer diagnosis using Wisconsin Breast Cancer datasets.
3. Survey a number of possible data mining techniques that can be applied on diagnosis of breast cancer.

3. ACRONYMS
1. BMI : Body Mass Index.
2. BRCA : Breast Cancer Susceptibility Gene.
3. BRCA 1 : Breast Cancer Gene 1.
4. BRCA 2 : Breast Cancer Gene 2.
5. CART : Classification and Regression Trees.
6. CNB : Core Needle Biopsy.
7. DCIS : Ductal Carcinoma In-situ.
8. FNAB : Fine-needle Aspiration Biopsy.
9. HRT : Hormone Replacement Therapy.
10. ID3 : Iterative Dichotomiser 3.

11. IG : Information Gain.
12. MRI : Magnetic Resonance Imaging.
13. SEER : Surveillance, Epidemiology, and End Results.

4. OVERVIEW OF BREAST CANCER

4.1 Breast Cancer

Breast cancer [1] is a type of cancer originating from the breast tissue, commonly from the inner lining of the milk ducts or lobules supplying the ducts with milk. Breast cancer occurs in both men and women, although the former type is rare. It remains the number one form of cancer that woman are diagnosed with around the world. Even with enhanced treatment, the lack of early detection has put women at even higher risk of dying from this disease. Statistics reveal that there were 40,000 female deaths and 232,670 new cases recorded in the United States in 2014 [2].

4.2 Breast Anatomy

Figure 1 shows a cross sectional layout of the breast. The various different sections indicated are as follows:
Enlarged Breast (Top):
A : ducts
B : lobules
C : dilated section of duct to hold milk
D : nipple
E : fat
F : pectoralis major muscle
G : chest wall/rib cage

Breast Cross Section (Bottom):
A : normal duct cells
B : ductal cancer cells
C : basement membrane
D : lumen (center of duct)

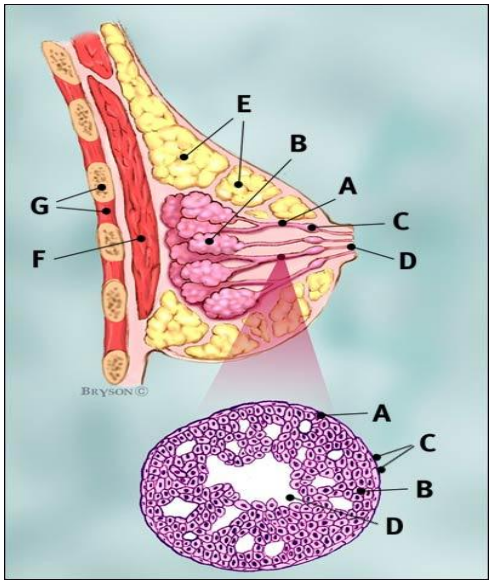

Figure 1. Breast Cross section (Bottom) Enlarged Breast (Top)

4.3 Types of Breast Cancer

Breast cancer can be of different types depending on the part of the breast it develops on. There have been two broad classification of breast cancer as shown in Figure 2.

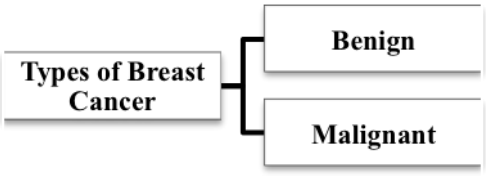

Figure 2. Types of Breast Cancer

1. **Benign Breast Cancer (Non-Invasive) [3]:** It is also known as carcinoma in situ. This type of cancer doesn't spread to neighboring tissue regions and hence is rarely a threat to life. These cells remain entirely in-situ (in their place of origin) because they have not yet developed the ability to spread outside of these ducts, either within the breast or elsewhere in the body. The cancer cells most commonly develop inside the milk ducts and hence it is also known as Ductal carcinoma in situ (DCIS) cancer. Both men and women can develop DCIS.

2. **Malignant Breast Cancer (Invasive) [3]:** Malignant or Invasive is the type in which the cancer has the potential to spread from the breast to other parts of the body and is a threat to life. Often they can removed but sometimes grow back. The most common type of invasive breast cancer is invasive ductal cancer. This accounts for 80 % of all cases of breast cancer.

3. **Other types of Breast Cancer [3]:** The less common type of breast cancer includes invasive lobular breast cancer, which develops in the cells of the milk-producing lobules, inflammatory breast cancer, tubular breast cancer, medullary breast cancer, and papillary breast cancer.

The distribution of various breast cancers in screening population is shown in Figure 3.

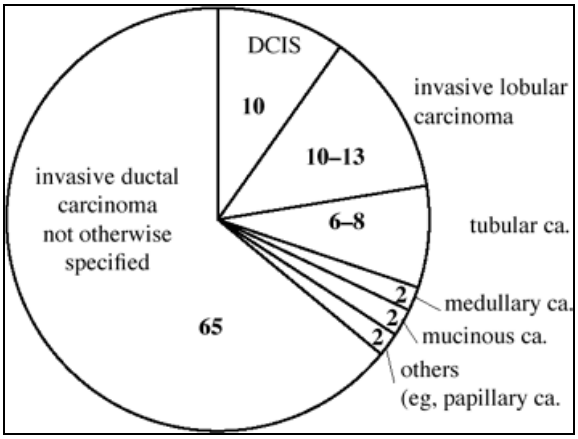

Figure 3. Distribution of Breast Cancers in Screening Population (Numbers – Percentage)

4.4 Risk Factors

Risk factor is anything that affects the chances of an individual of getting a disease, such as breast cancer. Presence of risk factors isn't a complete affirmative that a woman will develop breast cancer. There have been cases where many women with breast cancer have no apparent risk factors. Some factors need to be known by women so that they can

lower their risk of breast cancer. Since causes of breast cancer are not fully known. Researchers believe that these risk factors increase (or decrease) the changes of developing breast cancer.

Since breast cancer is a complex disease it is likely to be caused by a combination of risk factors. Some of the factors associated with breast cancer – can't be changed (Non-preventable) like age, genetic factor, heredity. While making choices can change other factors (Preventable) like overweight, lack of exercises, smoking [4]. The known risk factors for breast cancer are shown in Figure 4.

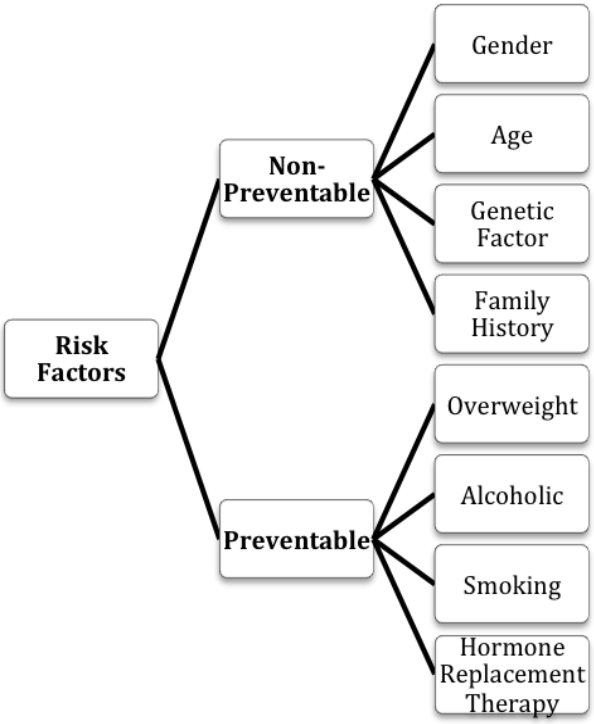

Figure 4. Risk Factors of Breast Cancer

Non-Preventable Risk Factors:

1. **Gender:** Main risk factor of breast cancer is being a woman. The disease is about 100 times more common amongst women than men.

2. **Age:** Risk of developing breast cancer increases with age. 2 of 3 invasive breast cancers are found in women with age 55 or older.

3. **Genetic risk factors:** 5% of breast cancer cases have strong inherited family risk. There are two autosomal dominant genes, BRCA1 and BRCA2 that account for most cases of familial breast cancer. Women with harmful BRCA mutation have 65% to 85% risk of developing breast cancer.

4. **Family History:** The risk of developing breast cancer increases (two-fold) if the woman's mother, sister, father or child has been diagnosed with breast or ovarian cancer. The risk increases if the relative was diagnosed before the age of 50.

Preventable Risk Factors:

1. **Overweight:** Obese women have a higher risk of being diagnosed with breast cancer than women with healthy weight. There is an increase risk of recurrence of breast cancer if a woman is overweight.

2. **Hormone Replacement Therapy:** Estrogen hormone therapy has been used to relieve symptoms of menopause and to help prevent osteoporosis but studies reveal that it also causes more risk of breast cancer.

3. **Alcoholic:** Alcohol consumption increases the risk of breast cancer, with the relationship being linear and dose dependent.

4. **Smoking:** It is linked with higher risk of breast cancer in younger and premenopausal women.

Other risk factors include:

- High BMI after menopause: Weight gain after menopause increases the risk of cancer breast.
- Lack of exercise
- Radiation Therapy to the chest (before age 30)
- Hormonal use – postmenopausal
- Late pregnancy at an older age,
- Race (African American-higher risk)
- High bone density

4.5 Symptoms of Breast Cancer

People who have breast cancer will initially notice only one or two symptoms and signs. Presence of these signs and symptoms do not actually mean that the person has breast cancer. The major symptoms of breast cancer are as shown in Figure 5 and few are described in this sub section [5].

1. **Breast Lump:** It is the most common symptom of breast cancer. Painless, hard mass with irregular edges verges to be cancerous but breast cancer can also be tender, soft or rounded.

2. **Change in Nipple:** Nipple retraction (turning inward) or nipple discharge (other than breast milk) also can be a major symptom of breast cancer.

3. **Skin Dimpling:** Puckering of skin on the breast is also considered one of the symptoms of breast cancer.

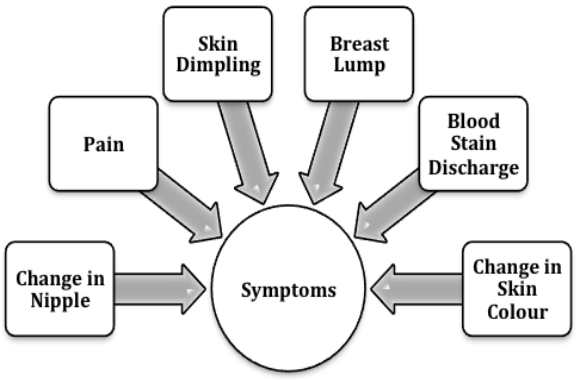

Figure 5. Symptoms of Breast Cancer

4.6 Diagnosis of Breast Cancer

Breast cancer can be found after symptoms appear, but for many women early breast cancer have no symptoms. Hence performing screening test before any symptoms develop is very essential. Early detection of breast cancer (before presence of symptoms), in the localized stage, increases the 5-year survival rate to 98 %.

The tests that can be performed are classified into a triple assessment routine as shown in Figure 6. Breast cancers tend to be larger and more likely to be spread beyond the breast if symptoms are predominant. In contrast, breast cancer that is found during screening tests is more prone to be smaller and confined to the breast.

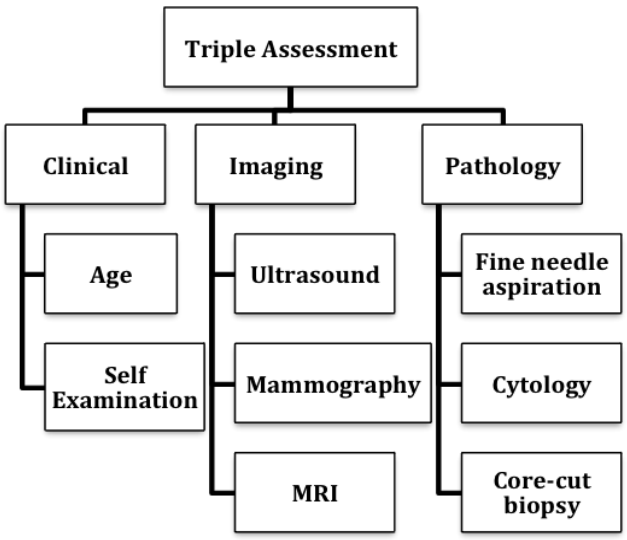

Figure 6. Diagnostic Tests for Breast Cancer

- **Clinical [18]:**
  - **Examination**: Women above a certain age and showing symptoms should have a physical exam to check for breast cancer. The breast is examined for lumps or suspicious areas (change in texture, size).

- **Imaging {18]:**
  - **Mammography**: It is an X-Ray of the breast. It is used to diagnose women who have breast symptoms. A mammogram shows abnormality in the breast, which include lesions.
  - **Ultrasound (Sonography):** The test uses sound waves to outline a part of the breast. It is usually helpful in women with dense breasts and is used to target a specific area found on the mammogram.

- **Pathology [18]:**
  - **Biopsy:** This test involves taking a sample of tissue cells from the breast and testing to see whether it is cancerous or not.
  - **Fine Needle Aspiration:** This test uses a thin hollow needle to withdraw a small amount of tissue (including fluids) from the breast and test it under the microscope to check whether cells are cancerous or not.

4.7 Treatment of Breast Cancer

The prognosis and treatment of breast cancer depend on the stage of the cancer and the type of breast cancer. Breast cancer diagnosed at a later stage requires a different treatment than when diagnosed in its early stages. A patient may have one treatment or a combination. Treatment of Breast Cancer can be summarized in Figure 7.

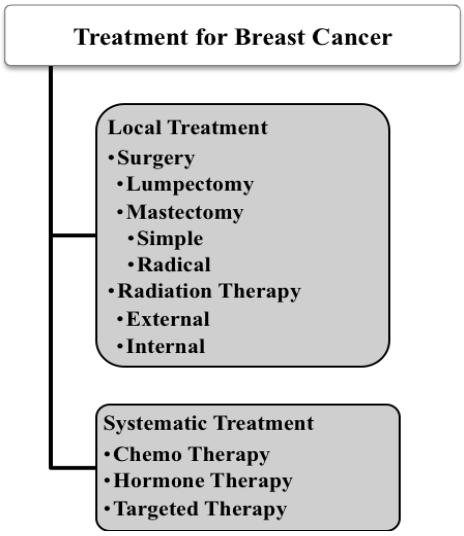

Figure 7. Treatment for Breast Cancer

- **Local Treatment [19]**
  - **Surgery:** It either involves removing the cancerous lump (tumor), which is known as breast-conserving surgery or mastectomy that is removal of the whole breast.
  - **Radiation Therapy:** It involves controlled dosages of radiation are used to kill cancer cells. Usually given to a patient after surgery and chemotherapy to kill any residual cancer cells.

- **Systematic Treatment [19]**
  - **Chemo Therapy:** It involves using anti-cancer drugs to kill the cancer cells.
  - **Hormone Therapy:** Breast cancer may be stimulated to grow by hormones oestrogen or progesterone, which is naturally developed by the body. This treatment involves lowering levels of hormones in the body and reversing the effect.

4.8 Investigation of Breast Cancer

After being diagnosed with breast cancer, a patient will have various treatment options depending on the stage of breast cancer and the doctor treating. Different factors are worked out for the best treatment, including the type, patients' age and general health. Figure 8 shows the flow diagram representing a few sets of options available for a patient after being diagnosed with breast cancer.

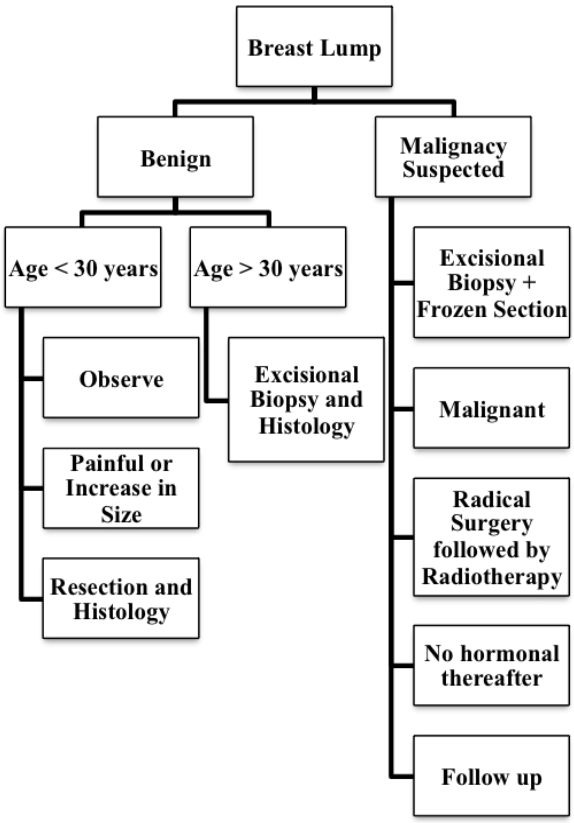

Figure 8. Flow Diagram for Breast Cancer Diagnosis

4.9 Prognosis of Breast Cancer

The chances of survival vary by the stages of breast cancer. Non-invasive and the early stages have a better chances of survival than that for the metastatic breast cancer (stage 4) which is the stage wherein the cancer has spread beyond the neighboring tissues. Table 1 shows the 5-year survivability rate of a cancer patient.

Table 1. Breast Cancer Survivability Rate [6]

Stage	Description	5 – year survival (%)	10 – year survival (%)
Stage – 0	No evidence of Primary Tumor	95	90
Stage – 1	Tumor <= 2cm	85	70
Stage – 2	Tumor > 2cm & <= 5cm	70	50
Stage – 3	Tumor > 5cm	55	30
Stage – 4 (Metastasis)	Any size with extending to - chest wall or skin	5	2

5. METHODOLOGY

5.1 Decision Trees

Decision tree [7] is a classifier that is expressed as a recursive partition of the instance space. It creates a predictive model, which maps observations about a node to conclusions about the nodes' target value. In a tree structure leaves represent the class labels and branches represent conjunctions of feature leading to the class labels. Figure 9 shows the illustrated example of binary decision tree.

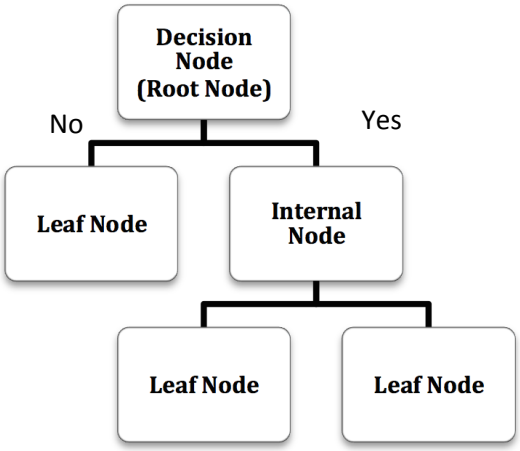

Figure 9. Illustrated example of binary decision tree

Sub section *B* discusses employing decision tree data mining technique to construct a predictive model to diagnose whether a tumor is benign or malignant depending upon various attributes associated with a particular medical record.

5.2 Proposed Data Model

Decision tree provides a powerful technique for classification and prediction in Breast Cancer diagnosis problem. Various decision tree algorithms are available to classify the data, including ID3, C4.5, C5, J48, CART and CHAID. In this paper we have chosen J48 decision tree algorithm [8] to establish the model.

10-fold cross validation [9] is used to prepare training and test data. After data pre-processing (CSV format), the J48 algorithm is employed on the dataset using WEKA (Java Toolkit for various data mining technique) [16] after which data are divided into "benign" or "malignant" depending on the final result of the decision tree that is constructed. Figure 10 shows the flow of the research conducted to construct the model. The algorithm for conducting the procedure is as follows:

ALGORITHM: BREASTDIAGNOSIS

INPUT: Wisconsin Breast Cancer data set pre-processed to satisfy prerequisites of the data mining technique.

OUTPUT: J48 Decision Tree Predictive Model with leaf node either benign or malignant.

PROCEDURE:

1. Acquire dataset from Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository.
2. Pre-process data for applying J48 decision tree data mining technique.
    a. Remove Sample Code Number from attribute list
    b. Numeric to nominal type of data conversion of Class attribute. (2 – Benign, 4- Malignant)
3. Pre-processed dataset uploaded in WEKA toolkit for analysis.
4. Information Gain algorithm applied in WEKA and IG of respective attributes record.
5. Decision Tree J48 algorithm implemented, generating a decision tree with leaf nodes as the class label (benign and malignant).
6. Diagnosis of new patients is achieved by cross referencing new attribute values in the decision tree and following path till the leaf node reached which would either specify benign or malignant tumor.

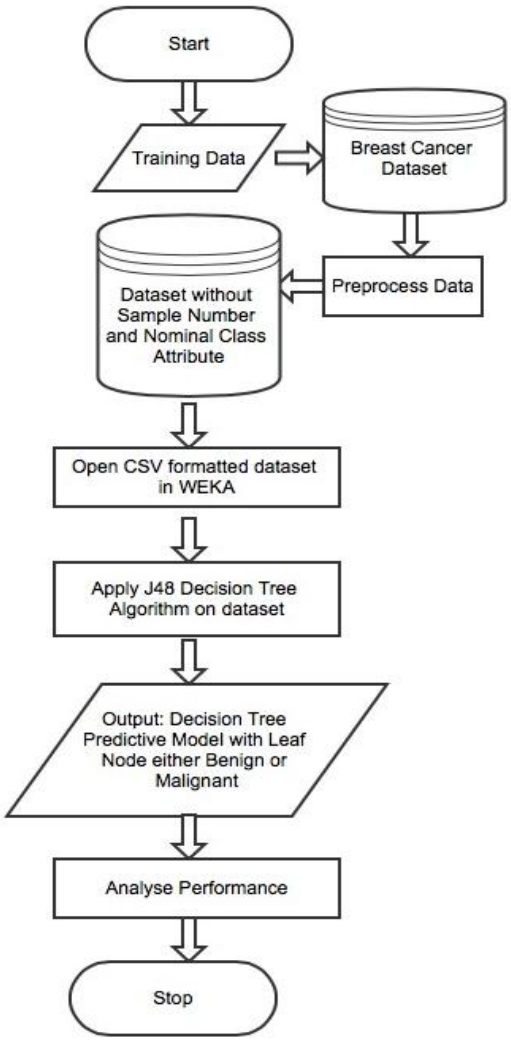

Figure 10. Flow Diagram for Breast Lump Detection

5.3 Data Description and Pre-Processing

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository [10] is used to differentiate benign (non-cancerous) from malignant (cancerous) samples. Table 2 shows a brief description of the dataset that is being considered.

Table 2. Description of Breast Cancer Dataset

Dataset	No. Of Attributes	No. Of Instances	No. Of Classes
Wisconsin Breast Cancer (Original)	11	699	2

Details of the attributes present in the dataset are shown in Table 3.

Table 3. Wisconsin Breast Cancer Dataset Attribute

S.No	Attribute	Domain
1	Sample Code Number	Id number
2	Clump Thickness	1 – 10
3	Uniformity of Cell Size	1 – 10
4	Uniformity of Cell Shape	1 – 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 – 10
11	Class	2(Benign) or 4(Malignant)

- **Clump Thickness:** Monolayer grouping in benign and multi layer grouping for cancerous cells.
- **Marginal Adhesion:** Normal cells stick together while cancer cells lose their ability. This is also relating factor to a single epithelial cell size, which is enlarged for a malignant cell.
- **Bare Nuclei:** Benign tumors have nuclei, which are not surrounded by cytoplasm.
- **Bland Chromatin:** Cancer cells have coarse chromatin.
- **Mitoses:** Uncontrollable levels of mitoses (cell-division) are seen in cancer cells.

The dataset comprises of 699 instances of breast cancer patients with each, either having malignant or benign type of tumor. Figure 11 shows the distribution of the patient based on the class label (malignant or benign).

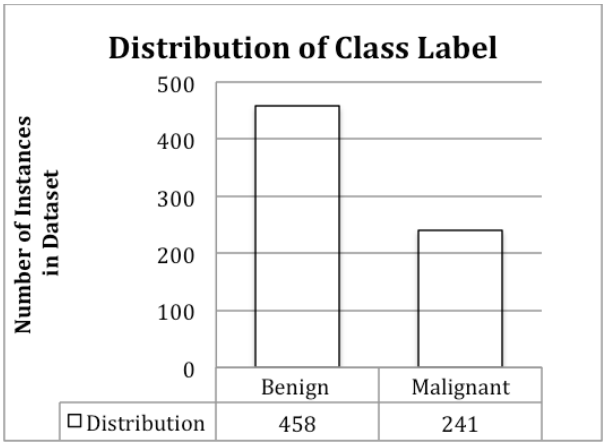

Figure 11. Distribution of Class Label in Dataset

In order to apply the J48 decision tree algorithm to the dataset, pre-processing is required. In this step two attributes of the dataset are changed. Since the sample number is not required in the formation of the model it is removed from the record. J48 requires its class label to be nominal (String) in type. Hence the last attribute of each record in the dataset was changed to either Benign (if attribute value=2) or Malignant (if attribute value=4). Figure 12 shows the unformatted data along with the formatted pre-processed data.

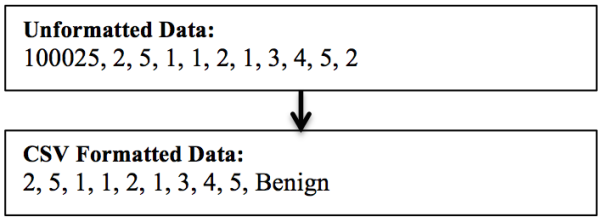

Figure 12. WEKA Formatted Data Input

5.4 Results and Discussion

Feature selection was done on the dataset wherein the sample code number attribute was removed (unbiased predictive model) so as to perform decision tree J48 data mining technique.

Decision tree J48 implements Quinlan's C4.5 algorithm [8] for generating pruned tree. The tree generated by J48 can be used for classification of whether a patient had benign or malignant tumor. The data mining technique uses the concept of information entropy. Each attribute of the data is used to make a decision by splitting the data into smaller modules.

It examines normalized information gain (IG) [11] (difference in entropy) that results from choosing an attribute as a split point. The highest normalized IG is used at the root of the tree. IGs of attributes in the dataset are shown in Table 4. The procedure is repeated until the leaf node is created for the tree specifying the class attribute that is chosen.

Table 4. Information Gain for Dataset Attributes

Information Gain	Attribute
0.675	Uniformity of Cell Size
0.660	Uniformity of Cell Shape
0.564	Bare Nuclei
0.543	Bland Chromatin
0.505	Single Epithelial Cell Size
0.466	Normal Nucleoli
0.459	Clump Thickness
0.443	Marginal Adhesion
0.198	Mitoses

Figure 13 shows the J48 pruned tree (Level-wise Representation) that was generated by the WEKA toolkit when the decision tree data mining technique of J48 was applied on the pre-processed dataset. The tree generated **14 number of leaf nodes** and the **total size of the tree was 24**.

```
Uniformity of Cell Size <= 2
|   Bare Nuclei    <= 3: Benign (405.39/2.0)
|   Bare Nuclei    > 3
|   |   Clump Thickness <= 3: Benign (11.55)
|   |   Clump Thickness > 3
|   |   |   Bland Chromatin <= 2
|   |   |   |   Marginal Adhesion <= 3: Malignant (2.0)
|   |   |   |   Marginal Adhesion > 3: Benign (2.0)
|   |   |   Bland Chromatin > 2: Malignant (8.06/0.06)
Uniformity of Cell Size > 2
|   Uniformity of Cell Shape <= 2
|   |   Clump Thickness <= 5: Benign (19.0/1.0)
|   |   Clump Thickness > 5: Malignant (4.0)
|   Uniformity of Cell Shape > 2
|   |   Uniformity of Cell Size <= 4
|   |   |   Bare Nuclei    <= 2
|   |   |   |   Marginal Adhesion <= 3: Benign (11.41/1.21)
|   |   |   |   Marginal Adhesion > 3: Malignant (3.0)
|   |   |   Bare Nuclei    > 2
|   |   |   |   Clump Thickness <= 6
|   |   |   |   |   Uniformity of Cell Size <= 3: Malignant (13.0/2.0)
|   |   |   |   |   Uniformity of Cell Size > 3
|   |   |   |   |   |   Marginal Adhesion <= 5: Benign (5.79/1.0)
|   |   |   |   |   |   Marginal Adhesion > 5: Malignant (5.0)
|   |   |   |   Clump Thickness > 6: Malignant (31.79/1.0)
|   |   Uniformity of Cell Size > 4: Malignant (177.0/5.0)
```

Figure 13. J48 Pruned Tree – Rules Generated

Table 5 shows the confusion matrix of the J48 Decision tree that was generated. The entries in the confusion matrix have the following meaning in the context of this paper:

- ***a*** : number of correct predictions that an instance is negative,
- ***b*** : number of incorrect predictions that an instance is positive,
- ***c*** : number of incorrect predictions that an instance is negative, and
- ***d*** : number of correct predictions that an instance is positive.

Table 5. Confusion Matrix of J48 Decision Tree

	A - Benign	B-Malignant
A - Benign	**438 (a)**	**20 (b)**
B-Malignant	**18 (c)**	**223 (d)**

The Breast cancer data with 699 tuples and 10 different attributes was analyzed to identify the error rates and accuracy. Table 6 shows the accuracy measures of the result.

Table 6. Performance of J48 Decision Tree

	Instances	Percentage
Correctly Classified Instances	661	94.5637 %
Wrongly Classified Instances	38	5.4363%

Table 7 presents various other statistics on which the predictive model can be compared with other techniques of data mining.

Table 7. Other Statistics Result

Kappa statistic	0.8799
Mean absolute error	0.0694
Root mean squared error	0.2229
Relative absolute error	15.352 %
Root relative squared error	46.8927 %
Total Number of Instances	699

The results show that J48 classifiers with feature selection is a superior technique that can be applied on breast cancer diagnosis and can further be developed with more training data to accurately predict the same.

6. ALTERNATIVE APPROACHES

6.1 Neural Networks

Neural networks [12] architecture consists of an input layer, a hidden layer and an output layer. The input layer represents the elements of the dataset and the output layer consists of one node. Weights between the layers are adjusted using the training data of the breast cancer database using feed forward neural network and back propagation learning algorithm.

6.2 Digital Mammography

There have been experiments for tumor detection from digital mammography [13]. Figure 14 shows a sample digital mammogram of a normal and a cancerous breast. Association rule mining [14] has been used for anomaly detection and classification. The mammograms are passed through various pre-processing phases where the images are enhanced and cleaned. The features (tissue type, position of breast) are extracted by splitting the image in sixteen parts. The Apriori algorithm is used on the training data with user-defined support and association rules are extracted.

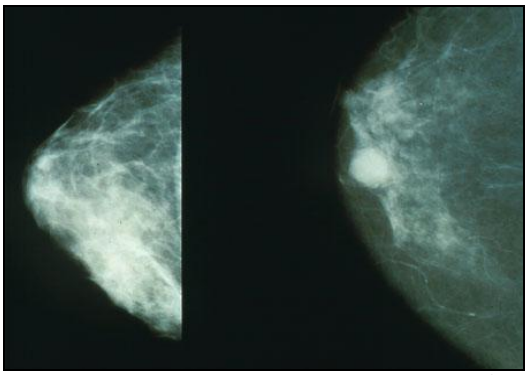

Figure 14. Digital Mammogram – Normal Breast (Left) – Cancerous Breast (Right) [17]

6.3 Naïve Bayes Classifier

In [15] the authors Abdelghani Bellaachia & Erhan Guven have performed analysis of the prediction of survivability rate of breast cancer patients using Naïve Bayes and WEKA toolkit. SEER database (period of 1973-2002 with 482,052 records) have been used with two additional fields Vital Status Recode (VSR) and the Cause of Death (COD).

7. FUTURE FOR BREAST CANCER DIAGNOSIS

In future it is planned to collect the data from various regions across the world and create a more accurate and general predictive model for breast cancer diagnosis. Future study will also concentrate on collecting data from a more recent time period and find new potential prognostic factors to be included in a decision tree. The work can be expanded and enhanced for the automation of Breast cancer diagnosis.

8. CONCLUSION

The automatic diagnosis of Breast cancer is an important real-world medical problem. Detection of breast cancer in its early stages is the key for treatment. This paper shows how decision trees are used to model actual diagnosis of Breast cancer for local and systematic treatment, along with presenting other techniques that can be applied. Experimental results show the effectiveness of the proposed model. The performance of decision tree technique was investigated for the Breast cancer diagnosis problem.

9. ACKNOWLEDGMENTS

The authors wish to thank Dr. B. Vijayakumar and Mrs. S. Rajeshwari for their technical support and helping us in doing a lot of research on this paper.

10. REFERENCES

[1] National Cancer Institute: http://www.cancer.gov/cancertopics/types/breast.

[2] National Cancer Institute Breast Cancer, http://www.cancer.gov/cancertopics/types/breast

[3] Breast Cancer Organization, http://www.breastcancer.org/ symptoms/types

[4] Breast Cancer Organization, http://www.breastcancer.org/ risk/factors/

[5] Breast Cancer Organization, http://www.breastcancer.org/symptoms/

[6] Bellaachia Abdelghani and Erhan Guven,"Predicting Breast Cancer Survivability using Data Mining Techniques", Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining, 06.

[7] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2000.

[8] Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria, Decision Tree Analysis on J48 Algorithm for Data Mining. Proceedings of *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 6, June 2013.

[9] H. Blockeel and J. Struyf. Efficient algorithms for decision tree cross-validation. Proceedings of the Eighteenth *International Conference on Machine Learning* (C. Brodley and A. Danyluk, eds.), Morgan Kaufmann, 2001, pp. 11-18

[10] William H Wolberg, Olvi Mangasarian, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA

[11] White, A.P., Liu, W.Z.: Technical note: Bias in information-based measures in decision tree induction. Machine Learning 15(3), 321–329 (1994)

[12] Chi C.L., Street W.H. and Wolberg W.H., "Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets", *Annual Symposium Proceedings / AMIA Symposium*, 2007.

[13] D. Brazokovic and M. Neskovic. Mammogram screening using multiresolution-based image segmen- tation. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(6):1437–1460, 1993.

[14] R. Agrawal, T. Imielinski, and A. Swami. Min- ing association rules between sets of items in large databases. In *Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data*, pages 207–216, Washington, D.C., May 1993.

[15] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques, "Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining," 2006.

[16] Michael Feld, Dr. Michael Kipp, Dr. Alassane Ndiaye and Dr. Dominik Heckmann "Weka: Practical machine learning tools and techniques with Java implementations"

[17] Wikipedia, http://en.wikipedia.org/wiki/File:Mammo_breast_cancer.jpg

[18] American Cancer Society, http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-diagnosis

[19] NHS Choices, http://www.nhs.uk/Conditions/Cancer-of-the-breast-female/Pages/Treatment.aspx