Society for Science and Education United Kingdom
TNC Transactions on Networks and Communications
Volume 3, Issue 2
ISSN: 2054 - 7420

Breast Cancer Risk Prediction Using Data Mining Classification Techniques

[1]Peter Adebayo Idowu, [2]Kehinde Oladipo Williams, [3]Jeremiah Ademola Balogun and [4]Adeniran Ishola Oluwaranti
[1, 3, 4]*Department of Computer Science and Engineering, Faculty of Technology, Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria*
[2]*Department of Physical and Computer Sciences, College of Natural Applied Sciences, McPHERSON University, Ajebo, Ogun State, Nigeria*
kehindewilliams@yahoo.com; paidowu1@yahoo.com

ABSTRACT

Breast cancer poses serious threat to the lives of people and it is the second leading cause of death in women today and the most common cancer in women in developing countries in Nigeria where there are no services in place to aid the early detection of breast cancer in Nigerian women. A number of studies have been undertaken in order to understand the prediction of breast cancer risks using data mining techniques. Hence, this study is focused at using two data mining techniques to predict breast cancer risks in Nigerian patients using the naïve bayes' and the J48 decision trees algorithms. The performance of both classification techniques was evaluated in order to determine the most efficient and effective model. The J48 decision trees showed a higher accuracy with lower error rates compared to that of the naïve bayes' method while the evaluation criteria proved the J48 decision trees to be a more effective and efficient classification techniques for the prediction of breast cancer risks among patients of the study location.

Keywords: breast cancer, classification, prediction, risk factors, naïve bayes, J48 decision trees

1 Introduction

According to WHO (2002) cancer has been responsible for the deaths of millions of people worldwide with an estimated increase of 50% for developing countries and for 70% of the total deaths due to cancer. According to Parkin et al (2003) developing nations only possess 5% of global funds for cancer control and very few human and material resources are also available in such countries (Grey et al, 2006).

The American Cancer Society (2008) defines cancer as a generic term for a large group of diseases that can affect any part of the body; other terms are malignant tumors and neoplasm. Breast cancer is a type of cancer which affects the breast tissue which is most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk (Sariego, 2010). Breast cancer is caused by a number of factors called risk factors; they are classified as either modifiable (those that can be controlled like habits, environmental hazards, etc) or non-modifiable factors (those that cannot be controlled like, gender, family history etc). According to the Collaborative Group on Hormonal Factors in Breast Cancer (2002), the primary risk factors for breast cancer are being female and of an older age. Other potential risk factors include: family history of breast cancer, age of menarche (first occurrence of menstruation), age of first birth, age of menopause, body weight (BMI), alcohol

DOI: 10.14738/tnc.32.662
Publication Date: 10th March 2015
URL: http://dx.doi.org/10.14738/tnc.32.662

consumption, exposure to radiation (Poongodi et al, 2011), higher hormonal levels and diet (Yager, 2006).

According to Johnson et al (2009) smoking tobacco appears to increase the risk of breast cancer which is higher depending on how long the person has been smoking. Long term smokers have an increased risk of about 35% to 50% (Santoro, 2009). The risk of breast cancer increases with an increased diet especially for those with fat diet (Blackburn, 2007), alcohol intake (Bofetta et al, 2006) and obesity. Radiation exposure (American Cancer Society, 2005) also increases the chances of breast cancer risk especially for women who have yearly mammogram tests especially between the ages 40 to 80 years face a risk of 225 in every million women screened (Hendrick, 2010). Also, exposure to pesticides, chemicals and organic solvents are believed to increase breast cancer risks also (Ferro, 2012). According to Boris et al (2010) genetics is believed to be the cause of 5% to 10% of breast cancer cases with those with none, one or two affected relatives with breast cancer before the age of 80 has a 2.3%, 4.2% and 7.6% risk respectively (Gage et al, 2012). Those with first degree relative with the disease face double the risk than a normal person.

Breast cancer risks can be reduced via early detection of the disease; according to the American Cancer Society (2007) early detection of breast cancer risks can help reduce the possibility of mitigating the full growth of tumors. The various ways of detecting breast cancer may include: clinical examination by a physician, self breast examination and mammography. Clinical examination of breast by a physician is one of the effective ways of reducing breast cancer mortality; it is required that a woman goes for clinical examination annually when above 40 years and every 3 years when between 20 and 40 years. Mammography involves the use of x-rays but with lower radiation; it has a breast cancer detection accuracy of 85 to 90% where routing mammogram leads to a 25 to 30% decrease in breast cancer mortality (American Cancer Society, 2007). Self-breast examination involves monthly observation of the breast and underarm by the patient; it allows the patient to be familiar with her breast and easily detect any anomaly she observes during the exercise. Diagnosis is the process of predicting the presence of breast cancer as either benign or malignant cases.

Classification is a data mining technique which involves the use of supervised machine learning techniques which assigns labels or classes to different objects and groups. It involves the process of model construction (analysis of training data for patterns) and model usage where the constructed model is used for classification. Classification accuracy is usually estimated as the percentage of test samples that are correctly classified.

This study aims at using data mining techniques to classify breast cancer risks using datasets of patients' information from LASUTH which contains the risk factors and the cancer classes (unlikely, likely and benign). The J48 decision trees and naïve bayes' classification of breast cancer was performed using the WEKA software.

2 Related Works

A number of papers have been documented and published on the use of data mining techniques in the classification of breast cancer risks. Some of such works are reviewed in the following paragraphs.

According to Rajesh et al (2012) who used SEER dataset for the diagnosis of breast cancer using the C4.5 classification algorithm. The algorithm was used to classify patients into either pre-cancer stage or potential breast cancer cases. Random tests were performed on the dataset which contained information for 1183 patients including the age of diagnosis, regional lymph nodes measures, and sequence number of tumors, dimension of primary tumor and contiguous growth of the primary tumor. The analysis involved the use of three random 500 records form the pre-processed data of 1183 and was used as training data and the lowest error rate achieved was 0.599. During the testing phase, the C4.5 classification rules were applied to a test sample and the algorithm showed had an accuracy of 92.2%, sensitivity of 46.66% and a specificity of 97.4%. Future enhancement of the work will require the improvisation of the C4.5 algorithm to improve classification rate to achieve greater accuracy.

Shajahan et al (2013) worked on the application of data mining techniques to model breast cancer data using decision trees to predict the presence of cancer. Data collected contained 699 instances (patient records) with 10 attributes and the output class as either benign or malignant. Input used contained sample code number, clump thickness, cell size and shape uniformity, cell growth and other results physical examination. The results of the supervised learning algorithm applied showed that the random tree algorithm had the highest accuracy of 100% and error rate of 0 while CART had the lowest accuracy with a value of 92.99% but naïve bayes' had the an accuracy of 97.42% with an error rate of 0.0258.

Mangasarian et al (1995) performed classification on both diagnostic and prognostic breast cancer data. The classification procedure adopted by them for diagnostic data is called Multi Surface Method-Tree (MSM-T) that uses a linear programming model to iteratively place a series of separating planes in the feature space of the examples. If the two sets of points are linearly separable, the first plane will be placed between them. If the sets are not linearly separable, MSM-T will construct a plane which minimizes the average distance of misclassified points to the plane, thus nearly minimizing the number of misclassified points. The procedure is recursively repeated. Moreover they have approached the prognostic data using Recurrence Surface Approximation (RSA) that uses linear programming to determine a linear combination of the input features which accurately predicts the Time-To-Recur (TTR) for a recurrent breast cancer case. The training separation and the prediction accuracy with the MSM-T approach was 97.3% and 97 % respectively whereas the RSA approach was able to give accurate prediction only for each individual patient. Their drawback was the inherent linearity of the predictive models.

Lundin et al (1999) has applied ANN on 951 instances dataset of Turku University Central Hospital and City Hospital of Turku. To evaluate the accuracy of neural networks in predicting 5, 10 and 15 years breast cancer specific survival. From the experiment the values of ROC curve for 5 years was evaluated as 0.909, for 10 years 0.086 and for 15 years 0.883, these values were used as a measure of accuracy of the prediction model. The author compared 82/300 false prediction of logistic regression with 49/300 of ANN for survival estimation and found ANN predicted survival with higher accuracy. It shows that neural networks are valuable tools in cancer survival prediction. In future the study should concentrate on collecting data from a more recent time period and find new potential prognostic factors to be included in a neural network model.

Delen et al (2005) compared ANN, decision tree and logistic regression techniques for breast cancer prediction analysis. They used the SEER data of twenty variables in the prediction models. From the experiment the author found that the decision tree with 93.6% accuracy and ANN with 91.2% are more superior to logistic regression with 89.2% accuracy. The study is based on multiple prediction models for breast cancer survivability using large datasets along with 10 fold cross validation method. It provides a relative prediction ability of different data mining methods. In future this work is extended by collecting real dataset in the clinical laboratory

2.1 Data Mining Process

Data mining is the process of extracting patterns from data; these patterns may be discovered depending on the data mining tasks that are applied on the dataset. The two basic data mining tasks are: descriptive data mining tasks which help to understand the characteristic properties of dataset and predictive data mining tasks which are used to perform predictions based on available dataset. Predictive data mining is the chosen data mining task for this study.

According to Gupta et al (2011) data mining applications can use different parameters to examine data which includes; association (patterns that define the relationship between data), sequence/pattern analysis (patterns where one event leads to another), classification (identification of new patterns with predefined targets) and clustering (grouping of identical of smaller objects). The basic steps include:

- *Problem definition* is the definition of the goals and objectives and the identification of tools to be used to build the defined model.
- *Data exploration* is the recommendation for useful dataset if the existing dataset does not meet the required need for analysis.
- *Data preparation* is the process of cleaning and transforming data to remove missing and invalid data and validation of data for robust analysis.
- *Modeling* is based on the desired outcomes and data. This involves the use of data mining algorithms (for this study; naïve bayes, decision trees and multi-layer perceptron) in meeting the necessary objectives-which for the purpose of this study is classification.
- *Evaluation and deployment* is the analysis and interpretation of the results of analysis to create recommendations for consideration.

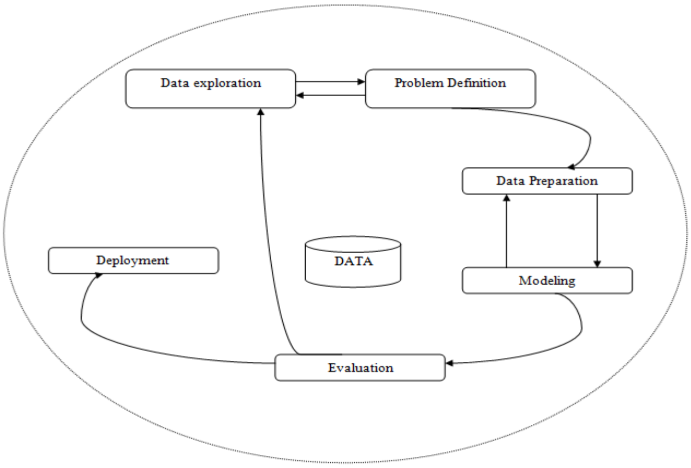


Figure 1: Data Mining Process Representation

3 Materials and Methods

In order to classify the breast cancer data collected form LASUTH with the aim of achieving high accuracy and precision; two supervised learning algorithms i.e., J48 decision trees and the naïve bayes are used. The data preprocessing was performed in order to remove inconsistent data and the data converted into a format that is useful in the simulation environment. WEKA data mining software was the environment used for simulating the breast cancer risk prediction model; which is an open-source data mining software used for academic purposes.

3.1 Training dataset description

LASUTH breast cancer data set was collected from the Cancer Registry of LASUTH, Ikeja in Lagos, Nigeria. The dataset collected contains 69 instances with 17 attributes. The class distribution is framed as unlikely, likely and benign. Hence there are 16 independent variables and 1 dependent variable. The nominal values are set for the independent variables and the dependent variable. The non-modifiable factors are the first 11 variables while the modifiable factors are the next 5 variables while the breast cancer risk is the last variable.

Table 1: The Training Dataset Description

S/N	Risk Factor (Points)	Values
1.	Family History of Breast Cancer	Yes, No
2.	Existence of Benign Breast disease	Never, Ever
3.	Mammographically Dense Breast	Never, Ever
4.	Age at First Birth	no, ≤30yrs, >30yrs
5.	Age at Menopause	no, ≤50yrs, >50yrs
6.	Body Mass Index (BMI)	< 25, ≥ 25
7.	Age at Menacre	no, ≤12yrs, >12 yrs
8.	Endogenous Estrogen Levels	Low, High
9.	Waist-Hip Ratio	< 0.81, ≥ 0.81
10.	Age	≤50yrs, >50yrs
11.	Sex	Male, Female
12.	Smoking Frequency	Never, former, current
13.	Alchohol Intake	Never, former, current
14.	Occupational Hazard	No, Yes
15.	Current Oral Contraceptive use	Never, Ever, Current
16.	Breast Feeding	Never, Ever
17.	Breast Cancer Risk	Unlikely, Likely, Benign

3.2 Data mining algorithms used

3.2.1 Naïve Bayes' classifier

Naive Bayes Classifier is a probabilistic model based on Baye's theorem. It is defined as a statistical classifier. It is one of the frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data. Let X is a data sample containing instances, X_i where each instances are the breast cancer risk factors (modifiable and non-modifiable). Let H be a hypothesis that X belongs to class C which contains (unlikely, likely and benign cases). Classification is to determine $P(H_j|X)$, (i.e., posteriori probability): the probability that the hypothesis, H_j (unlikely, benign or likely) holds given the observed data sample X.

- $P(H_j)$ (prior probability): the initial probability of the hypothesis in the class;
- $P(X_i)$: probability that sample data is observed for each attribute, i;
- $P(X_i|H)$ (likelihood): the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X; and
- posteriori probability of a hypothesis H_j (unlikely, likely or benign), $P(H_j|X_i)$, follows the Baye's theorem as follows:

For example, if for a variable X with i attributes (breast cancer risk factors) expressed as:

$X = \{X_1, X_2, X_3, X_4, \ldots\ldots\ldots, X_{1\}}$ and

H_j={unlikely, likely, benign}.

Then,

is the probability of the outcome of a risk factor being under the hypothesis, H_j;

is the probability of the outcome of the risk factor in the training dataset;

is the probability of the outcome of an hypothesis (unlikely, likely, benign i.e. j=3);

is the probability of a variable, X containing risk factors belongs to an hypothesis, H_j;

The breast cancer risk class output = maximum $[P(H_j|X)]$ for j=1, 2, 3

3.2.2 Decision Trees

J48 decision trees classifier is a simple decision learning algorithm, it accepts only categorical data for building a model. The basic idea of ID3 is to construct a decision tree by employing a top down greedy search through the given sets of training data to test each attribute at every node. It uses statistical property known as information gain to select which attribute to test at each node in the tree. Information gain measures how well a given attribute separates the training samples according to their classification.

It is suitable for handling both categorical as well as continuous data. A threshold value is fixed such that all the values above the threshold are not taken into consideration. The initial step is to calculate information gain for each attribute. The attribute with the maximum gain will be preferred as the root node for the decision tree.

Given a set S of breast cancer cases, J48 first grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S;
- Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S_1, S_2,……, S_n for a dataset containing n cases according to the outcome for each case, and apply the same procedure recursively to each subset.

It uses a statistical property known as information gain to select which attribute to test at each node in the tree. It measures how well a given attribute separates the training samples according to their classification.

3.2.3 Performance Evaluation

The performance evaluation criteria allow the measurement of the accuracy of the models developed using the training dataset. The results of the classification are recorded on a confusion matrix. A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the vertical. All correct classifications lie along the diagonal from the north-west corner to the south-east corner also called True Positives (TP) and True Negatives (TN) while other cells are called the False Positives (FP) and False Negatives (FN). If the unlikely case is considered positive then likely and benign are called negatives, if likely is considered as positive then unlikely and benign are considered negatives and the same also applies if benign is called the positive. These values are used to determine the following evaluation criteria.

The error rates of the developed models using both classifiers were also determined alongside with the performance evaluation criteria mentioned above.

4 Experimental Results and Discussions

The experimental results of this study using the two classifiers are discussed using the WEKA software data mining tool. As earlier discussed, breast cancer is classified as either unlikely, likely and benign. The performance evaluation results and the error rates are also discussed as follows.

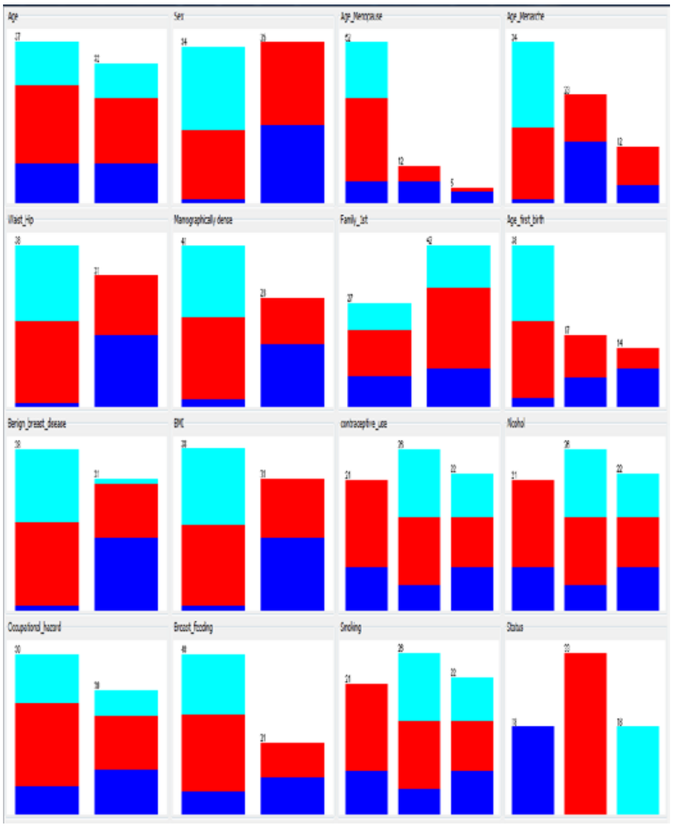
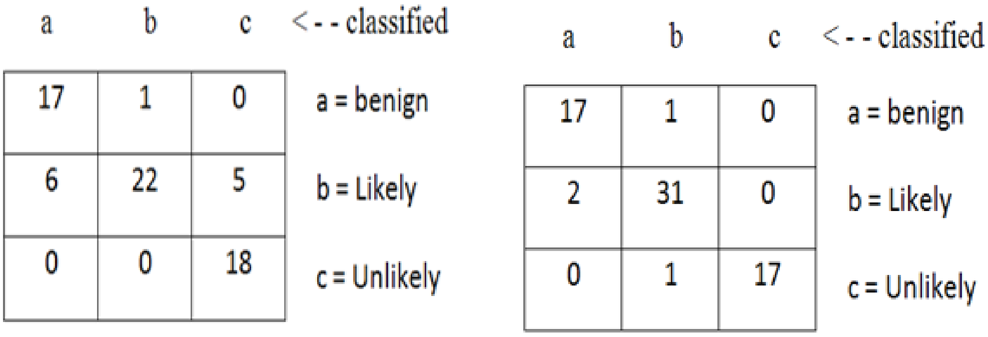


Figure 2: Distribution of the training dataset used



a	b	c	<- - classified
17	1	0	a = benign
6	22	5	b = Likely
0	0	18	c = Unlikely

a	b	c	<- - classified
17	1	0	a = benign
2	31	0	b = Likely
0	1	17	c = Unlikely

Figure 3: Confusion matrix of the results of classification using J48 decision trees (left) and naïve bayes' (right)

From the results of the data mining process for the prediction of Breast Cancer risk using J48 decision trees and Naïve Bayes' classifiers, the confusion matrix of both models can be seen in Figure 3.

The results of the J48 decision trees showed that from the 69 training data collected, out of 18 cases that were benign 17 were correctly classified and 1 incorrectly classified as Likely; out of the 33 cases that were likely 22 were correctly classified with 6 and 5 incorrectly classified as benign and unlikely respectively and form the total 18 cases that were unlikely all were correctly classified with no misclassifications.

The results of the naïve bayes' classifier showed that out of the 69 training data collected; out of 18 cases that were benign 17 were correctly classified with 1 misclassified as likely; out of the 33 cases that were likely 31 were correctly classified with 2 misclassified as benign and out of the 18 cases that were unlikely 17 were correctly classified and 1 misclassified as likely.

From the two confusion matrices, it can be seen that the naïve bayes' model had 57 correct and 12 incorrect classifications giving an accuracy of 82.6% but the J48 decision trees which had correct and incorrect classifications of 65 and 4 respectively had an accuracy of 94.2% (see Table 2 and Figure 4 below).

Table 2: Accuracy of naïve bayes' and J48 decision trees' model

	Naïve Bayes'	J48 Decision Trees
Correct Classification	57	65
Incorrect Classification	12	4
Accuracy (%)	82.6	94.2

From the two models developed for the prediction of breast cancer risk; the confusion matrix developed earlier was used to identify the accuracy of the models; other performance evaluation criteria are as follows. The True Positive (TP) rate/recall which is the percentage of the actual number of positive that were classified as positive cases has an average of 87% and 94% for the naïve bayes' and decision trees respectively. The False Positive (FP) rate which is the percentage actual number of positive cases that were misclassified also called false alarm has an average of 8.1% and 3.1% for the naïve bayes' and decision trees respectively. These results of the TP and FP rate have a value of 96.7% and 99% for the area under the graph of the Receiving Operating Characteristics (ROC) for naïve bayes' and decision trees respectively; this is a good indication of the effectiveness of both models but with the values of the TP rate, FP rate, Area under the ROC and accuracy of the models; the decision tree is a better model with an average precision of 94.4% compared with 82.6% for the naïve bayes' model (see Table 3 below). The error rates of the two models are 0.1396 and 0.058 for the mean absolute error and 0.3242 and 0.1703 for the relative absolute error of the naïve bayes' and the J48 decision trees model respectively (see Figure 5 below).

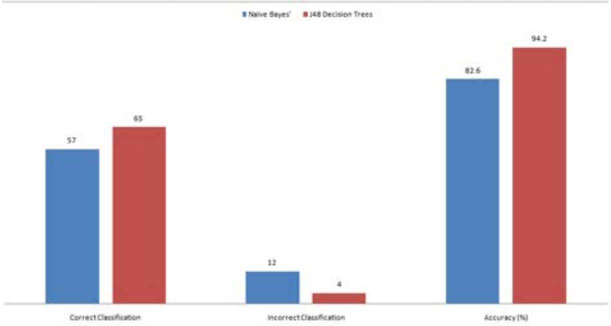


Figure 4: Accuracy, Correct and incorrect classification of Breast Cancer by both models

Table 3: Performance evaluation of both models

Performance criteria	Naïve Bayes'				J48 Decision Trees			
	Unlikely	Likely	Benign	Average	Unlikely	Likely	Benign	Average
TP rate	1	0.667	0.944	**0.870333**	0.944	0.939	0.944	**0.942333**
FP rate	0.098	0.028	0.118	**0.081333**	0	0.056	0.039	**0.031667**
Precision	0.783	0.957	0.739	**0.826333**	1	0.939	0.895	**0.944667**
ROC Area	0.995	0.953	0.953	**0.967**	0.998	0.985	0.987	**0.99**

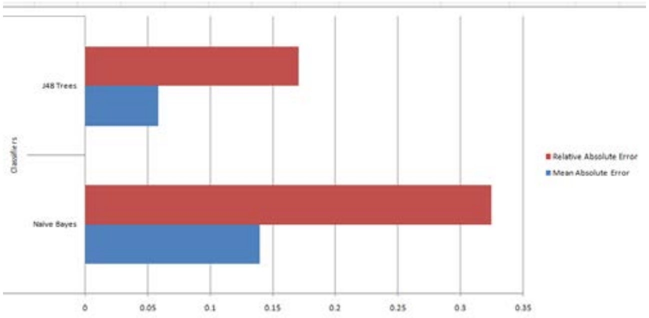
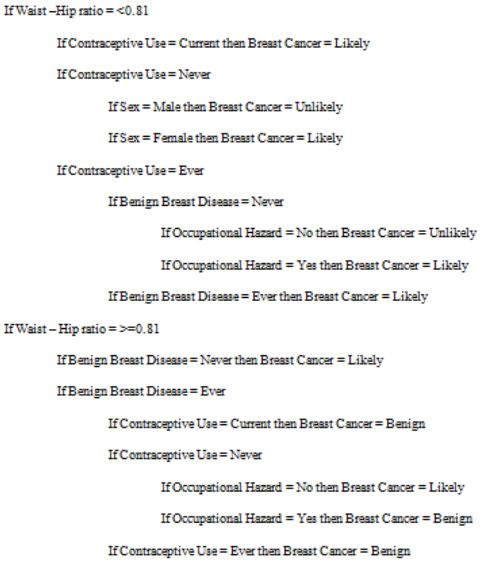


Figure 5: Error rate for both model

Figure 6 below gives an expression of the rules developed by the J48 decision trees model for the prediction of breast cancer risk using the dataset collected for cancer patients of LASUTH. It gives a clear picture of understanding better the relationship between each attributes and breast cancer risk.



```
If Waist –Hip ratio = <0.81
    If Contraceptive Use = Current then Breast Cancer = Likely
    If Contraceptive Use = Never
        If Sex = Male then Breast Cancer = Unlikely
        If Sex = Female then Breast Cancer = Likely
    If Contraceptive Use = Ever
        If Benign Breast Disease = Never
            If Occupational Hazard = No then Breast Cancer = Unlikely
            If Occupational Hazard = Yes then Breast Cancer = Likely
        If Benign Breast Disease = Ever then Breast Cancer = Likely
If Waist – Hip ratio = >=0.81
    If Benign Breast Disease = Never then Breast Cancer = Likely
    If Benign Breast Disease = Ever
        If Contraceptive Use = Current then Breast Cancer = Benign
        If Contraceptive Use = Never
            If Occupational Hazard = No then Breast Cancer = Likely
            If Occupational Hazard = Yes then Breast Cancer = Benign
        If Contraceptive Use = Ever then Breast Cancer = Benign
```

Figure 6:Rules created from the dataset using J48 decision trees for predicting cancer risk

From the above results shown, it is very clear that data mining techniques can be used in predicting breast cancer risks and that the J48 decision trees has a better accuracy than the naïve bayes' model which is a statistical tool. This is the rule that was used by the decision trees in testing the model using the test data and the decision trees shows that the best attributes for predicting breast cancer are: Waist-Hip ratio, Contraceptive use, Sex, Benign breast disease and Occupational hazard.

5 Conclusion

In this study two different data mining classification techniques was used for the prediction of breast cancer risk and their performance was compared in order to evaluate the best classifier. Experimental results shows that the J48 decision trees is a better model for the prediction of breast cancer risks for the values of accuracy, recall, precision and error rates recorded for both models. Hence, an efficient and effective classifier for breast cancer risks has been identified while the number of attribute covered by the classifier can be increased by increasing the sample size of the training set and hence the development of a more accurate model.

REFERENCES

[1] American Cancer Society (2005). "Breast Cancer Facts & Figures 2005–2006" (PDF). Archived from the original on 13 June 2007. http://web.archive.org/web/20070613192148/http://www.cancer.org/downloads/STT/CAFF2005BrFacspdf2005.pdf. Retrieved 2013-02-26.

[2] American Cancer Society (2007). "Cancer Facts & Figures 2007" (PDF). Archived from the original on 10 April 2007. http://web.archive.org/web/20070410025934/http://www.cancer.org/downloads/STT/CAFF2007PWSecured.pdf. Retrieved 2012-11-26.

[3] American Cancer Society (2007). "Cancer Facts & Figures 2007" (PDF). Archived from the original on 10 April 2007. http://web.archive.org/web/20070410025934/http://www.cancer.org/downloads/STT/CAFF2007PWSecured.pdf.

[4] Blackburn, GL; Wang, KA (2007). "Dietary fat reduction and breast cancer outcome: results from the Women's Intervention Nutrition Study (WINS)." *The American journal of clinical nutrition* **86** (3): s878-81. PMID 18265482.

[5] Boffetta P, Hashibe M, La Vecchia C, Zatonski W, Rehm J (August 2006). "The burden of cancer attributable to alcohol drinking". *International Journal of Cancer* **119** (4): 884–7. doi:10.1002/ijc.21903. PMID 16557583.

[6] Boris Pasche (2010). *Cancer Genetics (Cancer Treatment and Research)*. Berlin: Springer. pp. 19–20. ISBN 1-4419-6032-5.

[7] Collaborative Group on Hormonal Factors in Breast Cancer (August 2002). "Breast cancer and breastfeeding". *Lancet* **360** (9328): 187–95. doi:10.1016/S0140-6736(02)09454-0. PMID 12133652.

[8] Delen, D., Walker, G., Kadam, A. (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, vol. 34, pp. 113-127, June 2005.

[9] Ferro, Roberto (1 January 2012). "Pesticides and Breast Cancer". *Advances in Breast Cancer Research* **01** (03): 30–35. doi:10.4236/abcr.2012.13005.

[10] Gage, M; Wattendorf, D; Henry, LR (1 April 2012). "Translational advances regarding hereditary breast cancer syndromes". *Journal of surgical oncology* **105** (5): 444–51. doi:10.1002/jso.21856. PMID 22441895.

[11] Grey, N and Sener, S. (2006) Reducing the global cancer burden, http://www.hospitalmanagement.net/features/feature648/ , Date accessed 21 November 2012.

[12] Gupta, S.; Kumar, D., Sharma, A (2011). Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering (IJCSE). Vol. 2 No. 2 pg 198-195, April, 2011. ISSN: 0976-5166.. Accessed on June 24, 2014.

[13] Hendrick, RE (October 2010). "Radiation doses and cancer risks from breast imaging studies.". *Radiology* **257** (1): 246–53. doi:10.1148/radiol.10100570. PMID 20736332.

[14] Johnson KC, Miller, AB, Collishaw, NE, Palmer, JR, Hammond, SK, Salmon, AG, Cantor, KP, Miller, MD, Boyd, NF, Millar, J, Turcotte, F (2009). "Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk (2009).". *Tobacco control* **20** (1): e2. doi:10.1136/tc.2010.035931. PMID 21148114.

[15] Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H.,(1999) "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.

[16] Mangasarian,D.S.;Street, W.N.,Wolberg, W.H (1995). Breast cancer diagnosis and prognosis via linear programming, Operations Research, 43(4), pages 570-577, July-August 1995.

[17] Parkin, D.M., Ferlay, J., Hamdi-Cherif, M., Sitas, F., Thomas, J.O., Wabinga, H., Whelan, S.L. (2003). Cancer in Africa Epidemiology and Prevention, IARC (WHO) Scientific Publications no. 153, IARC Press, Lyon, France.

[18] Poongodi, M., Manjula, L., Pradeepkumar, S., Umadevi, M. *Cancer Prediction Technique Using Fuzzy Logic.* International Journal of Current Research, Vol. 3, Issue 11, pg 333-336, December 12, 2001. http://www.journalera.com. ISSN: 0975-833X. Accessed on June 24, 2014.

[19] Rajesh, K., Anand, S (2012). Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm. International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 2, April 2012. ISSN 2278-1021. http://www.ijarcee.com pg. 72 – 77.

[20] Santoro, E., DeSoto, M., and Hong Lee, J (February 2009). "Hormone Therapy and Menopause". National Research Center for Women & Families. http://www.center4research.org/2010/03/hormone-therapy-and-menopause/.

[21] Sariego J (2010). "Breast cancer in the young patient". *The American surgeon* **76** (12): 1397–1401. PMID 21265355.

[22] Shajahaan, S.S; Shanthi, S., Chitra, V.M. (2013). Application of Data Mining Techniques to model Breast Cancer Data. International Journal of Emerging Technology and Advanced Engineering Vol 3, Issue 11, November 2013. ISSN 2250-2459. http://www.ijetac.com pg 362 – 369.

[23] WHO. (2002). National Cancer Control Programmes; policies and managerial guidelines, 2nd edition.

[24] Yager JD (2006). "Estrogen carcinogenesis in breast cancer". *New Engl J Med* **354** (3): 270–82. doi:10.1056/NEJMra050776. PMID 16421368.