IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861.Volume 20, Issue 8 Ser.12 (August. 2021), PP 48-54
www.iosrjournals.org

Predicting Breast Cancer Treatment Using Decision Tree Algorithms and Statistical Metrics

EmanElwerfally[1], Huda Kutrani[2], Saria Eltalhi[3], Naeima Ashleik[4]
[1](Faculty of Information Technology/ University of Benghazi, Libya)
[2](Health Informatics Department, Faculty of Public Health/ University of Benghazi, Libya)
1& 2These authors contributed equally to this work
[3](Computer Science Department, Faculty of Information Technology/ University of Benghazi, Libya)
4(Department of Statistics, Faculty of Science/ University of Benghazi, Libya)
Corresponding Author: Huda Kutrani

Abstract:
Background*: In both developed and developing countries, breast cancer is the most common cancer in women. Also, it is the second main cause of cancer death in women. However, appropriate and successful treatment has a positive effect on the survival rate for a patient with cancer according to WHO's report in 2016. Classification algorithms are frequently used to analyze breast cancer data to predict. The main objective of this research is to identify the best prediction model for breast cancer treatment by using decision tree algorithms.*
Materials and Methods*: Data were collected from the patients' records at the BMC such as patients' ages and stages of the disease. The dataset was 336 patients with malignant and 10 features. Three of the decision tree algorithms were used to develop breast cancer prediction models; the J48, CART, and Random Forest were used as classifiers. This research used WEKA software to build and evaluate the models. Statistical performance metrics were used toevaluate the models such asClassifier Accuracy,Kappa Statistic, and ROC Curves.*
Results*: Experimental results showed that the effectiveness of all models. But the Random Forest classifier has performed better well with the training dataset. Also, the results showed that the sensitivity, specificity, ROC area, and classification accuracy of the Random Forest model has achieved above 91% success for the test dataset. Also it had highest value of Kappa Statistic, and lowest value of Mean absolute error.*
Conclusion: *The research was concluded that the Random Forest algorithm was identified as the best predictive model of training dataset and test dataset in this research.*

Key Word*: Treatment of Breast Cancer, Decision Tree, Prediction Model, Statistical performance metrics Machine Learning, Breast Cancer,*

Date of Submission: 10-08-2021 Date of Acceptance: 25-08-2021

I. Introduction

Breast cancer is defined as an abnormal growth of cells lining the milk ducts or breast lobes, and it spread out into the tissues surrounding the breast. Thus, allowing the opportunity for cancer to spread to the lymph nodes and other body organs such as the liver lungs, and bones in advanced stages. Breast cancer remains one of the principal health concerns affecting women, and a rare malignancy in men [1-2].

The incidence of breast cancer has increased worldwide, but breast cancer mortality seems to be declining, suggesting a benefit from early detection and more effective treatment[3]. In Libya, breast cancer is the most common cancer in females; especially in women less than fifty years of age. Also, Libya is facing the difficult implementation of detection and treatment measures with breast cancer and cancer in general [4].

In recent years, machine learning and data mining have become increasingly important in the medical field; mainly for modeling cancer diagnosis and prognosis [5-8]. Moreover, these techniques can discover and identify patterns and relationships between variables, from complex and huge datasets[7-8].
Machine learning techniques assisted in the development of predictive models; thus obtain effective results to assist for accurate decision making. Artificial Neural Networks, Support Vector Machines, and Decision Trees techniques have been widely applied in cancer research [5-8].

There is a huge amount of data available in medical care organizations. These data are mostly used in the patients' care process. It will be a benefit to use it in prediction to improve healthcare. Especially since the appropriate treatment selection for breast cancer depends on several factors, such as the stage of disease at diagnosis, patients' age, and tumor size. The aim of the research is to use Decision Tree algorithms on real data of cancer patients; data obtained from the Benghazi Medical Center (BMC), to identify the best prediction model for breast cancer treatment.

II. Material and Methods

Dataset Collection

The dataset was collected from patient's records at the Department of Oncology (DoO), the Benghazi Medical Center. The dataset was covering a period of two years (2018 and 2019). The dataset was 336 patients with malignant and 10 features. All features described in Table1.

Table 1: Description of features in the breast cancer dataset

Variable Name	Description	Value
Place	Place of residence	Benghazi - Out Benghazi
Age at diagnosis	Patient age at diagnosis in years	numeric
Side	The side of breast diagnosed with cancer	(Left - Right - Bilateral)
Cancer stage	Grade of cancer	I - II - III - IV
No. Nodes	Number of Lymph Node involved with cancer	0 - (1-3) - (4-9)
Tumor size	The size of tumor in cm	(<=2) - (>2 - =<5) - (>5)
Distant	Refers to cancer that has spread from the original tumor to distant organs or not	Yes - No
Lymph nodes	Refers to Lymph nodes if involved with cancer or not	Yes - No
duration	time of patient still alive after diagnosing in months	numeric
Class (Treatment type)	The type of treatment underwent by the patient as their first treatment.	Chemotherapy -Mastectomy - Lumpectomy

Training and Testing dataset: Dataset was split into two subsets. The training set was 80% of instances and the testing set was 20% of instances.
The dataset was converted to the arff format, which is the file type used by the Weka tool.

Machine Learning models:
The machine learning model learns from past input data to make a future prediction as output. The machine learning techniques used in analyzing the breast cancer data by Decision Tree. It is a decision support tool that uses a tree-like illustration and It has many algorithms including J48, CART, and Random Forest [9].

J48 algorithm:it selects the best features for the root node using the concept of Information Gain Ratio to build trees. It features Improving computational efficiency [10-11].

CART algorithm: It is Classification and Regression Tree. It uses the Gini index to determine the best properties of data segmentation, and Gini describes the degree of purity. CART is not significantly impacted by outliers in the input variables. Also, CART can use the same variables more than once in different parts of the tree. This capability can uncover complex interdependencies between sets of variables [12-13].

Random Forest: It builds many decision trees when training data and outputting the class by the "bagging" method. It could provide the highest accuracy performance of a model. Also, it is a flexible, and easy-to-use machine learning algorithm [9,14].

Models Evaluation by Statistical Metrics
The models were assessed by statistical performance metrics which were prediction Accuracy, Sensitivity, Specificity, area under the curve ROC, Kappa Statistic, and Mean absolute error. Also, learning time and tree size were used to evaluate the models.

prediction Accuracy:accuracy is essentially measuring how well the model fits the training samples, thus predictive accuracy should be measured based on the difference between the observed values and predicted values [9].

Sensitivity:the ability of a test to correctly identify patients with a disease [9].

Specificity: the ability of a test to correctly identify people without the disease [9].

Area under the curve ROC:is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal) [9, 14].

Kappa Statistic: is used to test interrater reliability. The importance of rater reliability lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured. The Kappa result is interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair,0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement [15].

Mean absolute error: is usually intended to measure average model bias. The lower values of mean absolute error are better [16].

Test dataset: was tested by using three training models of the decision trees.

III. Result & Discussion

Three of the decision tree algorithms were applied to the training datasets with 10-fold cross-validation to preventing overfitting.The class attribute has three values Mastectomy, Chemotherapy, and Lumpectomy.

Statistical Performance Metrics to Evaluate Training Models

Table 2 demonstrates the results of different metrics for the algorithms to predict breast cancer (Original dataset). Comparisons of accuracy, sensitivity, specificity, area under the Receiver Operating Characteristics (ROC) curve, kappa statistic, and mean absolute error were presented in Table 2. Accuracy and ROC also show in Figure 1.

Table 2 : Performance parameters of decision tree algorithms for BMC breast cancer dataset

Classifiers	Accuracy%	Sensitivity	Specificity	ROC Area	kappa statistic	Mean absolute error
J48	72.49	0.725	0.789	0.838	0.51	0.36
CART	72.49	0.725	0.758	0.793	0.48	0.38
Random Forest (100 trees)	86.62	0.866	0.907	0.938	0.67	0.17

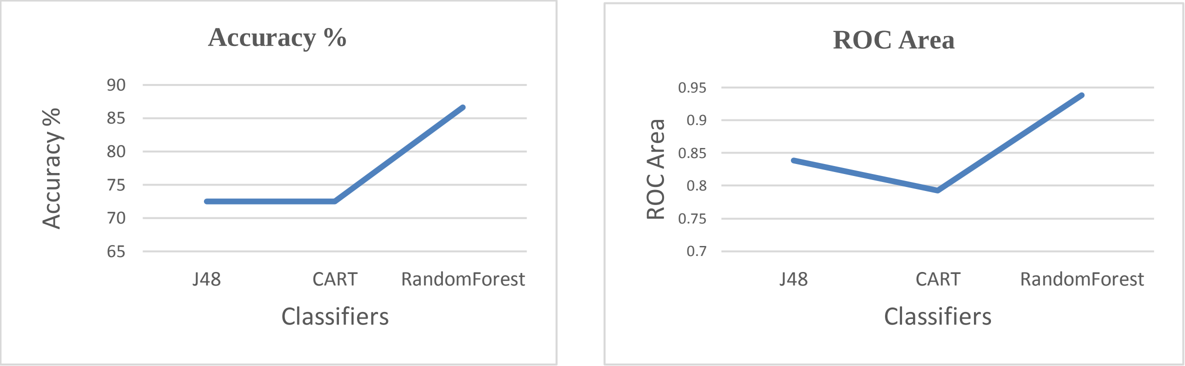

Figure 1: Performance comparison of classifiers using Accuracy and ROC Area

From Table 2 and Figure 1, the results indicated that the Random Forest classifier has performed better well with the training dataset with an accuracy score of 86.62 and a Sensitivity, Specificity, ROC,kappa statistic, and mean absolute error of 0.866, 0.907, 0.938, 0.67, and 0.17 respectively.

In the Random Forest classifier model, test specificity and test sensitivity were high which means the model was able to be correctly classified. Also, the higher ROC area was observed in this model which means the model is able to distinguish among types of treatment and identify the important patterns for each type of breast cancer treatment [17-18]. Moreover, the value of the kappa statistic was substantial which means the data collected in the study are correct representations of the variables measured [15]. Also, the mean absolute error was lowest which means the Random Forest classifier model was the better model [16].

However, the algorithms of the J48 and CART algorithms, are the second successful algorithms with a 72.49 % accuracy rate of each one.

Tree Size

When the size of the decision tree decreases, it reduces the complexity of the algorithm and time consumption [17]. The three decision tree classifiers had different sizes. The Random Forest of 100 trees whose sizes were from 108 to 158. In general, J48 usually generates larger trees [17]. However, J48 and CART provided similar performance accuracy, but the J48 built a larger tree than the CART classifier; as shown in Figure 2.

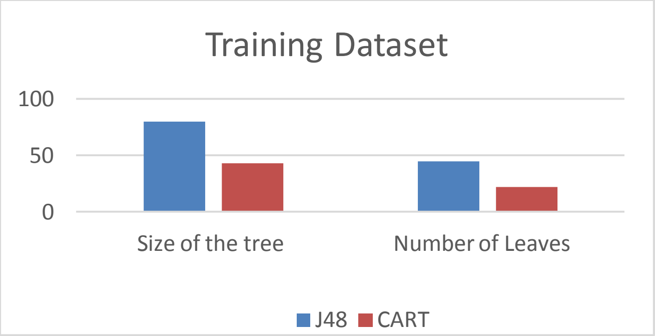

Figure 1: Tree size of decision tree algorithms for Training datasets

Learning Time

The Efficiency of the classifier is concerned with training time. When the classifier can make faster predictions, it has good efficiency [9].

Table 3: Learning Time

Classifiers	J48	CART	Random Forest
Learning Time	0.07	0.29	0.23

Time taken to build a model and learning was presented in Table 3 in seconds. The efficiency of the classifier is concerned with training time. When the classifier can make faster predictions, it has good efficiency [9]. Although Random Forest built 100 trees, time was taken to build the model less than the CART classifier which builds one tree. According to learning time, the J48 classifier had considered high efficiency compared to the Random Forest, but it had less performance accuracy and ROC area than the Random Forest as shown in Table 2. Also, the Random Forest built 100 trees compared to the J48 which built one tree.

Test Dataset

Table 4: Performance parameters of decision tree algorithms for test dataset

Models	Test dataset					
	Accuracy%	Sensitivity	Specificity	ROC Area	kappa statistic	Mean absolute error
J48	80.59	0.806	0.796	0.907	0.6	0.15
CART	71.64	0.716	0.613	0.829	0.37	0.2
Random Forest	91.4	0.914	0.955	0.963	0.83	0.17

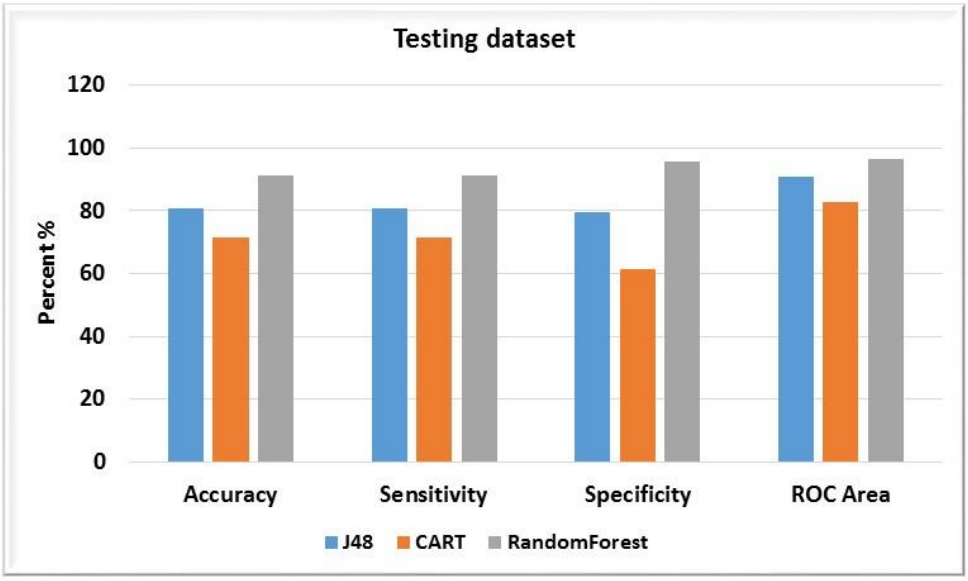

Figure 3: shows the comparison of accuracy, sensitivity, specificity, and ROC area with regard to the three models of the decision tree

Table 4 and Figure 3 show the values of the statistical parameters (sensitivity, specificity, ROC area, kappa statistic, mean absolute error, and total classification accuracy) for the test dataset. The classification algorithms that have an accuracy rate higher than 90% were considered the successful model for detecting the best treatmentof breast cancer [17]. The results showed that the sensitivity, specificity, ROC area, and classification accuracy of the Random Forest model has achieved above 91% success for the test dataset. Also, the kappa statistic was 0.83 which means almost perfect agreement [15]; whereas, the mean absolute error was low which means the model is good [16].

Therefore, the Random Forest model was the best in the accuracy for the testing dataset and training dataset. While 80.59% was the accuracy for the J48 model of the testing dataset but is reduced to 72.49 of the training dataset. Also, the accuracy for the CART model in testing dataset was 71.64% which was the lowest accuracy.

The sensitivity test is used to recognize the degree to which each predictive attribute contributes to the identification of the output class values[14,19-20]. In testing dataset, the sensitivity test of the J48 model was 80.6%, the CART's sensitivity test was 71.6%, and the sensitivity of the Random Forest model was 91.4%. From the previous, Random Forest was the best for the sensitivity of the testing dataset, also in the training dataset.

Specificity is measuring the proportion of negatives that are correctly identified [14, 19-20]. In testing dataset, the specificity of J48 model was 79.6%, and the CART model had the specificity of 61.3%. While, the Random Forest model was 95.5%. The best specificity was the Random Forest model, also it was best in the training dataset.

The ROC area values are measures of the model that have the ability to distinguish between patients with the disease and patients without the disease; in other words, the ROC area has the ability to distinguish among classification class attributes. When the ROC area is higher, it indicates to better the model able to distinguish between patients with the disease and patients without the disease [14-18]. From Table 4 and Figure 3,the ROC area of the J48 model was 90.7%, and the CART's ROC area was 82.9%. While the ROC area of the Random Forest model was 96.3%. From the previous, Random Forest was the best for the ROC area of the testing dataset, also in the training dataset.

Kappa Statistic is used to test interrater reliability; it measures if the data collected in the study are correct representations of the variables measured [15]. In the testing dataset, the Kappa Statistic of the J48 model was 0.6 and was 0.37 in the CART model; while the Kappa Statistic of the Random Forest model was 0.83. From the previous, Random Forest was the best for the Kappa Statistic of the testing dataset, also in the training dataset.

Moreover, the mean absolute error is used to measure average model bias; the lower values of the mean absolute error are better [16]. The Random Forest model had the lowest values of the mean absolute error in both the testing dataset and training dataset, thus it was a better model compared to the J48 and the CART models.

IV. Conclusion

Data mining and machine learning techniques can assist to reduce the number of negative decisions in the medical field. Consequently, data mining and machine learning tools used by researchers in the medical field to identify important patterns and relationships among a large number of medical variables to predict the outcome of a disease or diagnosis, or treatment [21-22].

The experimental study presented in this research was conducted in order to enable us to better understand data mining and machine learning techniques to identify the best prediction model for breast cancer treatment by using decision tree algorithms.

This research used statistical metrics to evaluate the performance and compare the results for three decision tree models J48, CART, and Random Forest. Experimental results showed that the effectiveness of all models. And the Random Forest model was the best in accuracy, sensitivity, specificity, ROC area, kappa statistic, and Mean Absolute Error for the training dataset and testing dataset.

Future Research

Testing the random forest algorithm on more than one dataset. In addition to the possibility of adding variables that affect the choice of the appropriate treatment, but retraining must be done to obtain the highest performance and a test using a different dataset.

Acknowledgement

The authors appreciate the contribution of Benghazi Medical Center as a source of the Home data set. A special thanks to the department of oncology staff for their help.

Ethical approval

The principle of Medical Research Ethics guidelines from the Faculty of Public Health and the Faculty of Medical, University of Benghazi, were observed during this study.

Consent

All respondents who agreed to participate in the study were required to initialize the informed consent form.

Conflict of Interest

The authors declare that they have no conflicts of interest in the research.

References

[1]. National Comprehensive Cancer Network. Breast cancer clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network: JNCCN. 2003 Apr;1(2):148-88.

[2]. Elmore JG, Armstrong K, Lehman CD, Fletcher SW. Screening for breast cancer. Jama. 2005 Mar 9;293(10):1245-56.

[3]. Organisation World Health, "WHO | Breast cancer: prevention and control," WHO. 2016, Accessed: Jan. 11, 2021. [Online]. Available: https://www.who.int/cancer/detection/breast cancer/en/.

[4]. Abulkasim MA. The prevalence of breast cancer in Africa and establishment of The Libyan Breast Cancer Registry (Master's thesis, Faculty of Health Sciences).

[5]. Yadav P, Varshney R, Gupta VK. Diagnosis of breast cancer using decision tree models and SVM. International Research Journal of Engineering and Technology (IRJET) e-ISSN. 2018 Mar:2395-0056.

[6]. Kutrani H, Eltalhi S. Cardiac catheterization procedure prediction using machine learning and data mining techniques. IOSR Journal of Computer Science. 2019;21(1):86-92.

[7]. Eltalhi S, Kutrani H. Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review. IOSR J. Dental Med. Sci.. 2019 Apr;18(4):85-94.

[8]. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Jan 1;13:8-17.

[9]. Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. The Morgan Kaufmann series in data management systems. 3rd ed. Burlington: Morgan Kaufmann Publishing; 2011.

[10]. Sornlertlamvanich V, Potipiti T, Charoenporn T. Automatic corpus-based Thai word extraction with the C4. 5 learning algorithm. InCOLING 2000 Volume 2: The 18th International Conference on Computational Linguistics 2000.

[11]. Jantan H, Hamdan AR, Othman ZA. Human talent prediction in HRM using C4. 5 classification algorithm. International Journal on Computer Science and Engineering. 2010 Dec;2(8):2526-34

[12]. Santhanam T, Sundaram S. Application of CART algorithm in blood donors classification. Journal of computer Science. 2010;6(5):548.

[13]. Zacharis NZ. Classification and regression trees (CART) for predictive modeling in blended learning. IJ Intelligent Systems and Applications. 2018 Mar 1;3:1-9.

[14]. Brownlee J. Bagging and Random Forest Ensemble Algorithms for Machine Learning. [Online]. Last Updated on December 3, 2020; Available from: https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/ [Accessed 19th January 2021].

[15]. McHugh ML. Interrater reliability: the kappa statistic. Biochemiamedica. 2012 Oct 15;22(3):276-82.

[16]. Jeykll J. MAE and RMSE: Which Metric is Better?.Medium publications [Online]. Last Updated on Mar 23, 2016. Available from:https:// medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

[17]. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861–74.

[18]. Narkhede S. Understanding AUC - ROC Curve. Towards Data Science. 2018; 26:220-7

[19]. Almunirawi KM, Maghari AY. A Comparative Study on Serial Decision Tree Classification Algorithms in Text Mining. International Journal of Intelligent Computing Research (IJICR). 2016; 7(4):754–760. doi: 10.20533/ijicr.2042.4655 .2016.0093.

[20]. Hornik K, Stinchcombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural networks. 1990 Jan 1;3(5):551-60. doi: 10.1016/0893-6080(90)90005-6.

[21]. Elsayad AM, Elsalamony HA. Diagnosis of breast cancer using decision tree models and SVM. International Journal of Computer Applications. 2013 Jan 1;83(5): 19-29. doi: 10.5120/14445-2604.

[22]. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013 Apr;4(2):2-4. doi: 10.4172/2157-7420.1000124.

Huda Kutrani;, et. al. "Predicting Breast Cancer Treatment Using Decision Tree Algorithms and Statistical Metrics.'' *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)*, 20(08), 2021, pp. 48-54.