Breast Cancer Classification using Decision Tree Algorithms

Omar Tarawneh[1], Mohammed Otair[2], Moath Husni[3]
Hayfa.Y. Abuaddous[4], Monther Tarawneh[5], Malek A Almomani[6]
Software Engineering Department, Amman Arab University, Amman, Jordan[1, 4]
Computer Science Department, Amman Arab University, Amman, Jordan[2]
Software Engineering Department, The World Islamic Sciences and Education University, Amman, Jordan[3, 6]
Computer Science Department, Isra University, Amman, Jordan[5]

***Abstract*—Cancer is a major health issue that affects individuals all over the world. This disease has claimed the lives of many people, and will continue to do so in the future. Breast cancer has recently surpassed cervical cancer as the most frequent cancer among women in both industrialized and developing countries and it is now the second leading cause of cancer mortality among women. A high number of women die each year as a result of this disease. Breast cancer is significantly easier to treat if caught early. This paper introduces a decision tree-based data mining technique for breast cancer early detection with highest accuracy, which helps patients to recover. Breast cancers are classed as benign (unable to penetrate surrounding tissue) or malignant (able to infiltrate adjacent tissue) breast growths. Two tests were included in the review. The primary study uses 10 breast cancer samples from the Kaggle archive, whereas the follow-up study uses 286 breast cancer samples from the same pool. The Decision Tree's accuracy in the first trial was 100%, while it was 97.9% in the follow-up inquiry. These findings justify the use of the proposed machine learning-based Decision Tree classifier in pre-evaluating patients for triage and decision-making prior to the availability of data.**

***Keywords*—Data mining; decision tree; classifier; breast cancer**

I. Introduction

A huge number of cells make up the human body, each with its own unmistakable reason. Cancer is characterized as the uncontrolled proliferation of any of these cells. Cells partition and create uncontrolled, bringing about a cancer, which is an abnormal mass of tissue. Growth cells increase and infiltrate the stomach related, neurological, and circulatory frameworks, interfering with normal body capacities. Although not all growths are malignant [1].

Cancer is classed based on the sort of cell that is afflicted, and there are more than 200 distinct varieties of cancer. Breast cancer is the subject of this review. Breast cancer is the most continuous type of cancer in ladies and it is considered as a major cause of death from all cancers, especially among ladies [2]. There are no established methods for avoiding breast cancer at this time. Figuring out how to recognize breast cancer at an early stage, then again, will assist with peopling who are affected. Breast cancer fixes chances can be enhanced assuming that the disease is diagnosed early. Provided that breast cancer is diagnosed early can it be avoided and safeguarded. Among the several evaluating strategies for classifying breast cancer, digital mammography is the most generally used [3].

Breast cancer survival rates have worked on considerably lately, regarding to the enhanced screening and treatment decisions. As a result of the advancements in data assortment and storage advances, medical organizations and hospitals may now store massive amounts of data associated with their medical records, including meds and disease indications. Data mining is the method involved with obtaining helpful information from large amounts of data utilizing complex algorithms. Medical data have broadened the applicability and potentials of these strategies [4]. Anticipating the result of a disease is a challenging task. Data mining strategies are regularly used to work on the expectation fragment. Large volumes of medical data may now be gathered and made available to medical research bunches using automated approaches. Thus, data mining apparatuses for finding patterns and correlations among countless variables are turning out to be increasingly normal, allowing for the forecast of ailment results based on recently gathered data [5]. However, most of the previous studies suffer from lack of accuracy in diagnose of this disease.

To fill in this gap, recognize and classify breast cancer at an early stage with high accuracy is very important in order to help patients to recover from this disease. Thus, this paper classified the breast cancer dataset by utilizing Decision Tree classification technique in WEKA. The breast cancer dataset is isolated and classified into categories based on features, performance, and different factors. Fig. 1 shows the entire interaction stream of the suggested paradigm.

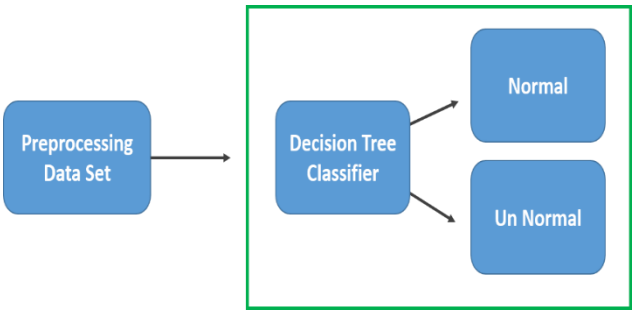

Fig. 1. Block Diagram of the Proposed Approach.

The classifier receives profound characteristics from the completely linked layer. The decision tree classifier makes use of these characteristics. After that, the characterization is complete, and all arrangement properties are defined. The suggested approach might be used to assess images of breast cancer.

The rest of this paper is organized as follows: Section 2 presents the mostly related works. Section 3 explains how decision tree can be used in the proposed paradigm. Section 4 discusses the experimental results. Finally, we conclude the paper in Section 5.

II. Related Work

Poomani and Porkodi conducted a study that aims to get the best classifier, by comparing different learning techniques [6]. The findings of their investigation reveal that the most elevated accuracy is found in the J48 classifier, which returns 0.979 with the least mistake rate of 0.9587 of all the classification methods. Their findings demonstrate that the combination of MLP and J48 classifiers with features determination Prognosis Breast Cancer (PCA) is better than different classifiers. In the Wisconsin Prognosis Breast Cancer WPBC dataset, their discoveries showed that the combination of IBK, J48, SMO, and MLP performs effectively. Saabithet et al. Compare numerous breast cancer data sets in, their research [7]. The extent of exactness with and without features selection techniques for breast cancer data was estimated utilizing classification algorithms like J48, MLP, and Rough. They conclude that the characteristic decision technique is the most reliable signifier process for enhancing the exactness of various categorization techniques, achieving the least Mean Standard Error (MSE) and the most elevated Recipient Operating Characteristics (ROC) to recognize breast cancer disease.

Peter et al. in [8] used two data mining approaches to quantify breast cancer gambles in Nigerian patients using the naïve Bayes algorithm and the J48 decision trees algorithms in their review exertion. The effectiveness of both categorization algorithms was evaluated to figure out which model was the most capable and beneficial. S. Syed Shajahaan, et al. [9] Compare DM approaches to model breast cancer data in a review. They took a gander at how to utilize decision trees to foresee the presence of breast cancer, as well as how to quantify the performance of conservative administered learning algorithms utilizing CART, C4.5, ID3, and Naive Byes. The ID3 random tree was demonstrated to be impressive with the most elevated accuracy in the investigations.

This study [10] compares the performance of the J48, AD Tree, BF Tree, and regression trees (CART) algorithms based on classification accuracy, using several accuracy measurements, for example, FP rate, TP rate, Recall, Precision, ROC Area, and F-measure. Decision trees are common and easy-to-understand structures from which rules may be obtained. Just the numerical values of explicit attributes in the breast cancer data are examined all through the implementation phase. The J48 classifier has the highest accuracy of close to 100%, while the CART algorithm has a 96 percent accuracy, the AD Tree algorithm has a 97 percent accuracy, and the BF Tree algorithm has a 98 percent accuracy. According to the classification findings of the four algorithms, it is obviously shown that J48 beats the other three techniques for the predetermined data set.

In [13], they proposed utilizing gene expression data to classify triple negative and non-triple negative breast cancer patients using a machine learning (ML) technique. The Support Vector Machine approach, out of the four ML algorithms tested, was able to classify breast cancer more accurately into triple negative and non-triple negative breast cancer and had less misclassification errors than the other three.

Epimack et. al [14] presented a computer-aided diagnostic (CAD) system that generates an optimum algorithm automatically which use 13 of the 185 features available to teach machine learning. To distinguish between malignant and benign tumors, researchers deployed five machine learning classifiers. For 10-fold cross-validation, the experimental findings indicated Bayesian optimization with a tree-structured Parzen estimator based on a machine learning classifier.

III. Decision Tree Learning as a Proposed Technique

Breast cancer classification was considered in many recent researches [13][14] using most of machine and deep learning techniques, because such classification is very important to be accurate as much as possible since it is related with most of women lives.

A decision tree is described as a classifier using a recursive split of the instance space. It generates a predictive model that connects node observations to inferences about the desired value of the nodes. The leaf in a tree structure represents the class.

Labels and branches represent the attributes that lead to the class labels. Fig. 2 is an illustrative example of a binary decision tree [11].

The proposed technique comprises constructing a prediction model for detecting whether a tumor is benign or malignant based on various characteristics associated with a given medical record using a decision tree data mining methodology. A decision tree is an excellent tool for classification and prediction in the case of Breast Cancer diagnosis. A number of decision tree approaches are available to categorize the data. Following data pre-processing (in CSV format), WEKA (Java Toolkit for Different Data Mining Techniques) is used to apply the Decision tree algorithm to the dataset, and the data is classed as benign or malignant based on the decision tree's final conclusion. Fig. 3 depicts the flow of the research that was utilized to create the model.

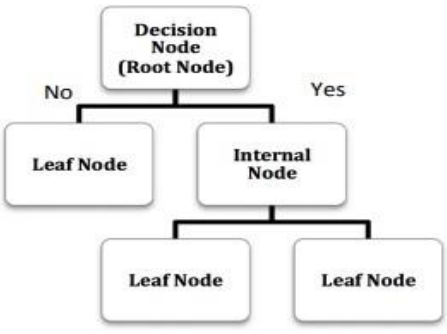

Fig. 2. Illustrated Example of Binary Decision Tree.

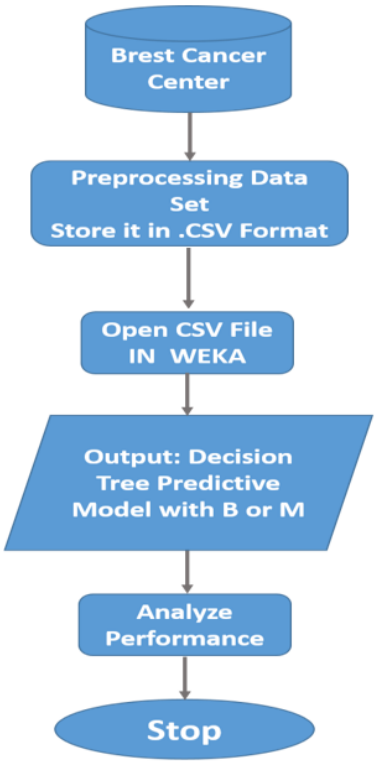

Fig. 3. Example of Binary Decision Tree.

The algorithm is carried out according to the following steps:

1) Collect datasets on breast cancer.

2) Pre-processing data in for performing the Decision Tree data mining approach (Benign, Malignant).

3) Data that has been pre-processed is submitted to the WEKA toolbox for analysis.

4) Implement the Decision Tree algorithm, which generates a decision tree with leaf nodes as class labels (benign and malignant).

5) New patient diagnoses are made by cross-referencing new attribute values in the decision tree and continuing the route until the leaf node is reached, which indicates whether the tumor is benign or malignant.

IV. Experimental Studies

A. Dataset

The CSV (Comma Separated Value) format was utilized in this study. Breast cancer data is supplied in CSV format via sex, age, menopause, tumor- size, inv-nodes, node-caps, deg-malig, breast, and breast-quad, irradiate, Class. These records were prepared in an Excel spreadsheet and saved as a CSV file, then uploaded to WEKA which accepted this format. In breast cancer data two types of properties, benign and malignant. In this study, these data are used and analyzed.

B. Experimental Results

The major goal of this paper is to evaluate the effectiveness of the breast cancer classification algorithms depending on a variety of input factors. They are analyzed using a Decision Tree algorithm. For the performance evaluation, the WEKA application is employed. Two setoff experiments are used to test each classifier. The first experiment has 10 samples. While the second one has 286 samples. This involves categorizing pre-processing based on all of the values of the attributes that have been taken. This paper compares the accuracy of categorization in the Decision Tree algorithm. The analysis of precision, FT Rate, and TP Rate is also performed. The following are the numerous formulas used to calculate various metrics. Precision P is calculated using the Formula (1) where TP = True Positive Rate, which is the proportion of expected positive instances, and FP = False Positive Rate, which is the part of the anticipated false positive cases:

$$Precision\ P = TP/(TP+FP) \quad (1)$$

The proportion of positive instances that were accurately detected has been described as True Positive Rate (TPR), Sensitivity, or Recall. Formula (2) will be used to calculate it:

$$Recall = TP/(TP+FN).\ \text{Where FN=False Negative Rate} \quad (2)$$

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \quad (3)$$

The accuracy in (3) will be computed as the fraction of the total number of correct predictions, where TN = True Negative. The calculation of an average of the recall and precision measures for retrieving information, which called as F-Measure as shown in Formula (4).

$$F = (2 * Recall * Precision)/Precision + Recall) \quad (4)$$

1) First experimental results: The first experiment has 10 samples of breast cancer [12]. Tables I to VI show the results of algorithms and compared with them.

Table I shows the summary of accuracy for proposed technique. Where the number of samples is 10.

Table II illustrates all the measures that used to evaluate the proposed method in details. We can see the proposed technique.

TABLE I. Summary Accuracy of Decision Tree Algorithm

Total Number of Instances	**10**	
Correctly Classified Instances	10	100%
Incorrectly Classified Instances	0	0%
Kappa statistic	1	
Mean absolute error	0.1429	
Root mean squared error	0.1429	
Relative absolute error	28.5714 %	
Root relative squared error	28.5714 %	

TABLE II. METRICS RESULTS ACHIEVED BY DECISION TREE ALGORITHM

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1
Weighted Avg.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	

TABLE III. CONFUSION MATRIX OF DECISION TREE ALGORITHM

a	b	classified as
5	0	a = 0
0	5	b = 1

Table III shows the confusion matrix.

Table IV shows the summary of accuracy in a Random Forest algorithm, where the number of samples is 10.

Table V shows the values of the using measures to evaluate Random Forest algorithm classification.

TABLE IV. SUMMARY ACCURACY OF RANDOM FOREST ALGORITHM

Total Number of Instances	10	
Correctly Classified Instances	5	50 %
Incorrectly Classified Instances	5	50 %
Kappa statistic	0	
Mean absolute error	0.5	
Root mean squared error	0.5	
Relative absolute error	100 %	
Root relative squared error	100 %	

TABLE V. METRICS RESULTS ACHIEVED BY RANDOM FOREST ALGORITHM

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.500	1.000	0.667	0.352	0.500	0.500	0
	0.000	0.000	0.000	0.000	0.200	0.352	0.500	0.500	1
Weighted Avg.	0.500	0.500	0.250	0.500	0.516	0.352	0.500	0.500	

TABLE VI. CONFUSION MATRIX OF RANDOM FOREST ALGORITHM

a	b	classified as
5	0	a = 0
5	0	b = 1

The results that attained from Tables I to VI illustrate the loss and accuracy throughout training and validation. Based on the results, it is clear that the maximum training and validation accuracy is seen in the Decision Tree architecture, with a loss rate of 0%. The Random Forest, on the other hand, achieves the lowest training and validation accuracy of 50% and a loss of 50% at iterations 100. Fig. 4 and Fig. 5 show that the loss values of decision tree equal to zero.

2) Second experimental results: When used large number of samples [12], the accuracy will be change as the Tables VII, VIII and IX. The classification of the attributes (age, menopause, tumor-size, inv-nodes, node- caps, deg-malig, breast, breast-quad, irradiat, Class) can be shown in the Fig. 6.

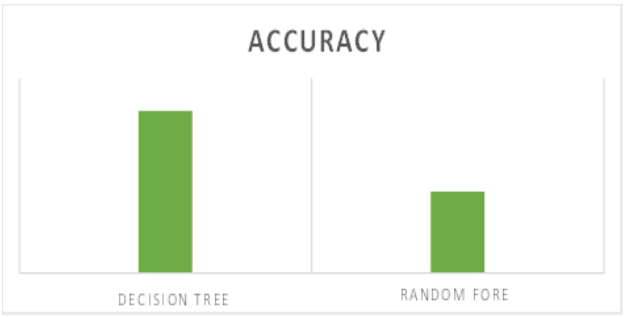

Fig. 4. Accuracy Comparison of Algorithms.

Fig. 5. Loose Comparison of Algorithms.

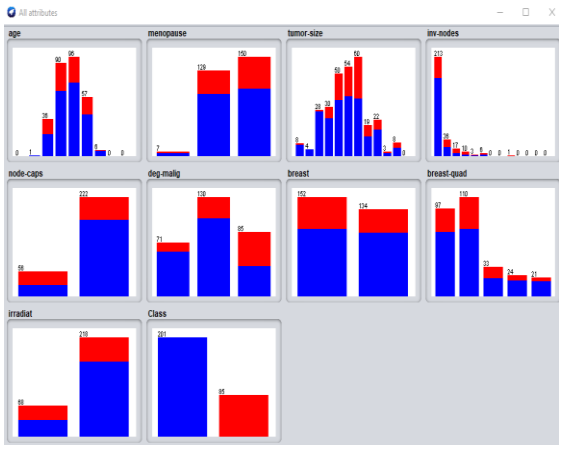

Fig. 6. Attribute Classification.

Table VII shows the accuracy summary in the Decision Tree algorithm, where the number of samples is 286.

Table VIII shows all measures values that were used to evaluate classification in studied methods.

TABLE VII. SUMMARY ACCURACY OF DECISION TREE ALGORITHM

Total Number of Instances	**286**	
Correctly Classified Instances	280	97.9021%
Incorrectly Classified Instances	6	2.0979%
Kappa statistic	0.9491	
Mean absolute error	0.0221	
Root mean squared error	0.1052	
Relative absolute error	5.2937 %	
Root relative squared error	23.0237 %	

TABLE VIII. DETAILED ACCURACY BY CLASS OF DECISION TREE ALGORITHM

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
	0.995	0.059	0.976	0.995	0.985	0.950	0.999	0.999	0
	0.941	0.005	0.988	0.941	0.964	0.950	0.999	0.996	1
Weighted Avg.	0.979	0.043	0.979	0.979	0.979	0.950	0.999	0.998	

TABLE IX. CONFUSION MATRIX OF DECISION TREE ALGORITHM

a	**b**	**classified as**
200	1	a = 0
5	80	b = 1

The results obtained from the Tables VII to IX demonstrate the loss and accuracy during the training and validation stages. The maximum training and validation accuracy are observed for Decision Tree architecture 97.9021% and loss is 2.0979%. These results show that the proposed technique is active in all cases. Fig. 6 illustrates the classification depending on attributes.

V. CONCLUSION

This study demonstrates the use of decision trees to represent actual breast cancer diagnosis for local and systemic treatment, as well as additional strategies that may be employed. The study assesses the performance of the Decision Tree algorithm in terms of classification accuracy, using several accuracy metrics such as the F-measure, ROC Area, Precision, Recall, TP rate, and FP rate. Decision trees are well-known and simple to comprehend structures from that rule may be derived. The studied model's efficacy is demonstrated by experimental findings. For the detection of breast cancer, the effectiveness of the decision tree approach was evaluated and explored. Throughout the implementation phase, only the numerical values of particular breast cancer features are assessed. The experimental findings reveal that the Decision Tree classifier has a 100% accuracy rate, while the Random Forest method only has a 50% accuracy rate. The performance of Decision Tree is superior than the other method for the specified dataset on the basis of the four algorithms' categorization results.

REFERENCES

[1] Syed SS, Shanthi S, Chitra VM. Application of Data Mining techniques to model breast cancer data. International Journal of Emerging Technology and Advanced Engineering. 2018 Nov; 3(11):362–9.

[2] Shiv Shakti S, Sant A, Aharwal RP. An Overview on Data Mining Approach on Breast Cancer data. International Journal of Advanced Computer Research. 2019; 3(13):256–62.

[3] Shweta K. Using data mining techniques for diagnosis and prognosis of cancer disease. International Journal of Computer Science, Engineering and Information Technology. 2019 Apr; 2(2):55–66.

[4] Takiar R, Nadayil D, Nandakumar A. Projections of number of cancer cases in India (2010-2020) by cancer groups. Asian Pac J Cancer Prev. 2017; 11(4):1045– 9.

[5] Vaidehi K, Subashini TS. Breast tissue characterization using combined K-NN classifier. Indian Journal of Science and Technology. 2019 Jan;8(1):23–6.

[6] Salama GI, Abdelhalim MB, Abd-elghany Zei Md. Breast cancer diagnosis on three different datasets using multi-classifiers. International Journal of Computer and Information Technology. 2020 Sep;1(1):36–43.

[7] Gupta S, Kumar D, Sharma A. Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian Journal of Computer Science and Engineering. 2020 Apr-May; (2):188–95.

[8] [Williams, K., Idowu, P. A., Balogun, J. A., & Oluwaranti, A. I. (2015). Breast Cancer Risk Prediction Using Data Mining Classification Techniques. Transactions on Networks and Communications, 3(2), 01. https://doi.org/10.14738/tnc.32.662.

[9] S. Syed Shajahaan, S. Shanthi, & V. ManoChitra (2013), Application of Data Mining Techniques to Model Breast Cancer Data, International Journal of Emerging Technology and Advanced Engineering 3(11): 362-369.

[10] Rajesh K, Anand S. Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm. International Journal of Advanced Research in Computer and Communication Engineering. 2019 Apr; 1(2):72–7.

[11] Saravana Kumar K, Arthanariee AM. Evaluate the multiple breast cancer factors and calculate the risk by software tool breast cancer risk evaluator. Indian Journal of Science and Technology. 2018 Apr;8(S7):686–91.

[12] "breast cancer dataset" accessed on: Jan. 10, 2022[online]. Available: https://www.kaggle.com/search?q=breast+cancer+dataset.

[13] Wu J, Hicks C. Breast Cancer Type Classification Using Machine Learning. J Pers Med. 2021; 11(2):61.

[14] Epimack Michael, He Ma, Hong Li, Shouliang Qi, "An Optimized Framework for Breast Cancer Classification Using Machine Learning", BioMed Research International, vol. 2022, Article ID 8482022, 2022.