ASEAN Journal of Science and Engineering

Journal homepage: http://ejournal.upi.edu/index.php/AJSE/

Evaluating the Performance of Supervised Machine Learning Algorithms in Breast Cancer Datasets

Obiwusi K.Y. [1], Olatunde Y.O. [1], Afolabi G.K. [1], Oke A. [2], Oyelakin A. M [3], Salami A.[4]*

[1]Department of Mathematics & Computer Science, College of Natural and Applied Sciences Summit University, Offa, Nigeria.

[2] ICT Unit, Summit University, Offa, Nigeria

[3] Department of Computer Science, Faculty of Natural and Applied Sciences Al-Hikmah University, Ilorin, Nigeria

[4] Library Unit, Summit University, Offa, Nigeria

*Correspondence: E-mail: obiwusi.kolawole@summituniversity.edu.ng

ABSTRACT

Breast cancer is the leading cause of mortality globally. Several attempts have been made to use data mining methodology together with machine learning techniques to develop systems that can detect or prevent breast cancer. In line with the reviewed paper; large datasets for illness analysis have been developed. In this study, the results of selected Machine Learning algorithms are compared: Decision Table, J48, SGD, bagging, and Naïve Bayes Updateable on Wisconsin Breast Cancer Original dataset was conducted using weka tools. Exploratory data analysis, pre-processed with supervised attribute selection and class order, was used to identify potential features to aid the performance of the chosen algorithms in classification. The empirical result showed that Decision Table explores greater likelihood (74% correctly classified instances, True Positive Rate of 0.752, False Positive Rate of 0.478, Precision of 0.77, receiver operating characteristic Area of 0.682) in terms of accuracy and efficiency compared with others. This study's comparison technique is thought to aid breast cancer detection.

ARTICLE INFO

Article History:
Submitted/Received 16 Jan 2022
First revised 03 Mar 2022
Accepted 12 May 2022
First available online 23 May 2022
Publication date 01 Sep 2023

Keyword:
Breast cancer classification,
Breast-cancer datasets,
Data mining,
ML algorithms,
Supervised machine learning.

1. INTRODUCTION

Breast cancer is the second leading cause of death worldwide from this most exquisite and internecine disease. The World Cancer Research Fund International (WCRFI) anticipated two million cases in 2018, with 626,679 deaths. Breast cancers come in a variety of forms, depending on the human body system, and are most common in postmenopausal women over 40 (Shah & Shah, 2021).

The goal of classification problems is to identify the characteristics that indicate which group each individual belongs to. This is where the case belongs. This pattern can be used to both understand and predict how data will change in the future. Classification and prediction are two of the most common data mining tasks for knowledge discoveries and a strategy for the future Supervised Machine learning is the term for the classification process. The classification target or class level is already known. There are a variety of techniques that can be used for this (Rajput *et al*., 2011). Asri *et al*. (2016) revealed that data mining approaches, for instance, applied to medical science topics rise rapidly due to their high performance in predicting outcomes, reducing costs of medicine, promoting patients' health, improving healthcare value and quality, and making real-time decisions to save people's lives.

There are numerous algorithms for breast cancer classification and prediction. This work shifted the focus of machine learning classifier comparison studies in breast cancer detection. This research paper provides an overview of a comparison of the performance of five different classifiers: Decision Table, J48 Decision Tree, SGD (stochastic gradient descent), bagging, and NaiveBayesUpdateable. We aim to evaluate the efficiency and effectiveness of those algorithms in terms of accuracy, sensitivity, specificity, and precision. Following that, the algorithms are compared for the datasets.

Using data mining, big data, artificial intelligence, and machine learning techniques, many writers trying to provide solutions for the early and efficient diagnosis of breast cancer have made significant contributions.

Data mining techniques in predicting breast cancer. The algorithms used for the prediction include SVM, KNN, Neural network, tree, and logistic regression. The authors used a clinical-stage dataset for about 130 Libyan women infected with the disease. Accuracy is used as a metric with 10-fold cross-validation, and no data preprocessing, according to the article, the Decision tree has the best performance, with an accuracy of 94.4 percent. These results can be said to be low in view of the need to detect every breast cancer patient as early as possible with the fast prediction techniques. However, the work only focused on the accuracy.

Wu and Hicks (2021) developed a model for used for classification of triple negative breast cancer and non-triple negative breast cancer of patients using gene expression (RNA-Seq) dataset. The algorithms used for evaluation are support vector machine, K-nearest neighbor, Naïve Bayes, and Decision trees. Accuracy and misclassification errors are used as metrics. The paper reported that the support vector machine predicted more accurately (90%) triple-negative and non-triple negative and has fewer misclassification levels. Meanwhile, the authors examined two metrics.

Comparison of the J48, random forest, naïve bayes, NaiveBayes simple, SMO poli-kernel and SMO RBF-kernel classification algorithm. The author used a breast cancer dataset. Precision, accuracy, recall, true positive rate, and false-positive rate were used as metrics. The paper reported that SMO Poly-Kernel + Simple K-Means yielded the best levels of 98.5% of Precision, 98.5% of recall, 98.5% of TPRATE, and 0.2% of FPRATE. However, no preprocessing of data, and the dataset used was not clearly stated.

DOI: http://dx.doi.org/10. 17509/xxxx.xxxx
p- ISSN 2776-6098 e- ISSN 2776-5938

2. METHODS

In this research, supervised machine learning techniques are applied. We also investigated the data collecting, preprocessing, feature selection, and breast classification processes utilizing the defined algorithms.

2.1. Data collection process

The datasets were collected from the UCI Machine Learning repository online using the following link: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29. The dataset descriptions are shown in **Table 1**. To have a better grasp of the nature of the characteristics in the datasets, expository data analysis was performed.

Table 1. Descriptions of the datasets.

S/N	Dataset Author/Year	No Input Features	Number of Instances	Missing values?	Classes	Data type of Features (Input)
1	Asri *et al*., (2016)	9	699	yes	2	The input features are 11 integer type

2.2. Experimental analyses

To carry out our comparative, we examined the behaviors of the selected algorithms; we conducted an experimental analysis that focuses on assessing both the effectiveness and efficiency of the algorithms. These are used to answer our earlier raised research question.

2.3. Simulation environment

The simulation environment in this study is conducted using WEKA. Waikato Environment for knowledge analysis (WEKA) software is a java based open-source tool. It is used as a Machine learning tool. Weka is a collection of a group of different machine learning algorithms which are useful for data preprocessing, classification, clustering, association rules, regression, select attributes, forecast, and C-Python scripting.

2.4. Hardware

The hardware configuration of the system used for the predictive analysis is as follows: core i3 processor, 8GB RAM, and 500 GB Hard Disk Drive but not limited.

2.5. Data preprocessing and feature selection method used

The data pre-processing steps carried out in this study are meant to make the features in the dataset to be in a usable format for the learning algorithms. Although the dataset is for public use that is complete, and consistent but yet there are still some missing values. We begin with the data cleaning followed by data preprocessing using an attribute selection filter, and then a class order filter is used to identify potential features that can help the five Machine Learning algorithms improve classification performance. A supervised attribute filter that can be used to select attributes. This is flexible and allows various search and evaluation methods to be combined. The class order is used to change the order of the classes so that the class values are no longer in the order specified in the header. The values will be in the order specified by the user -- it could be either in ascending/descending order by the class frequency or in random order (see **Table 2**).

DOI: http://dx.doi.org/10.17509/xxxx.xxxx
p- ISSN 2775-6793 e- ISSN 2775-6815

Table 2. The description of datasets after data cleaning and feature extraction.

S/N	No Input Features	Number of Instances	Missing values?	Classes	Data type of Features (Input & Feature)
1	6	286	No	2	The input features are recurrence events and no-recurrence events.

2.6. Experimental results

In this section, the results of the data analysis are reported. To apply our classifiers and evaluate them, we apply the 10-fold stratified cross-validation techniques used in evaluating predictive models that split the original set into a training sample to train the model, and a test set to evaluate it. After applying the pre-processing and preparation methods, we try to analyze the data visually and figure out the output values in terms of effectiveness and efficiency.

2.7. Effectiveness

In This section, we evaluate the effectiveness of all classifiers in terms of correctly classified instances, incorrectly classified instances, time to build the model, and accuracy. The results are shown in **Table 3** and **Figure 1. Table 4** show the performance comparison in terms of efficiency for the selected algorithms.

Table 3. Performance of the classifier.

Algorithms	Classifiers				
	Decision Table	J48	SGD	Bagging	NaiveBayesUpdateable
Correctly classified Instances	75	74	67	70	72
Incorrectly Classified Instances	25	26	33	30	28

Mathematical Formulae used for the performance of the metrics are:

Accuracy= (*TP+TN*)/(*TP+TN+FP+FN*) (1)

Precision=*TP*/(*TP+FP*) (2)

Recall=*TP*/(*TP+FN*) (3)

F1-score=2× (Precision X Recall) / (Precision + Recall) (4)

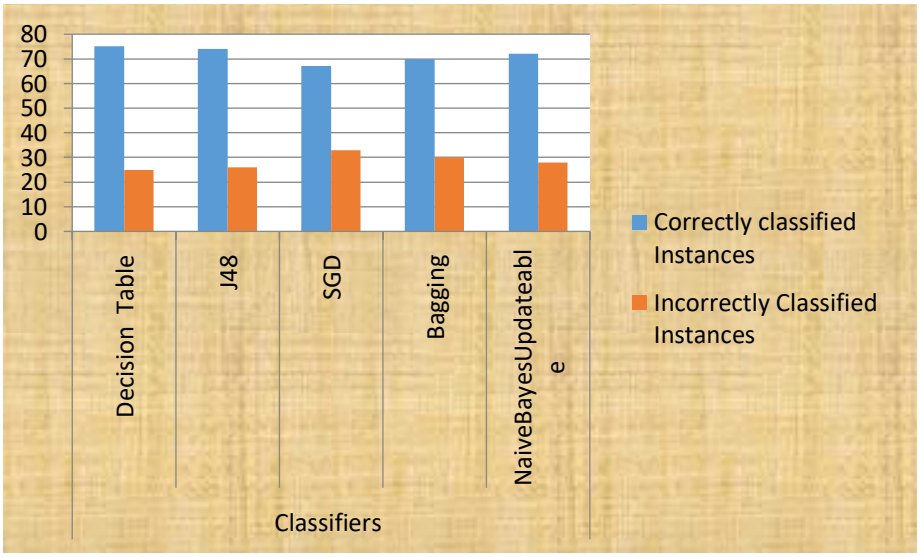

Figure 1. Chart of performance classifiers.

DOI: http://dx.doi.org/10. 17509/xxxx.xxxx
p- ISSN 2776-6098 e- ISSN 2776-5938

Table 4. Performance comparison in terms of efficiency for the selected algorithms.

Algorithms	True Positive Rate	False Positive Rate	Precision	Recall	ROC Area	F-measure	Class	Rank
Decision Table	0.35	0.08	0.65	0.35	0.68	0.46	recurrence-events	1
	0.92	0.65	0.77	0.92	0.68	0.84	no-recurrence-events	
J48	0.26	0.06	0.65	0.26	0.61	0.37	recurrence-events	2
	0.94	0.74	0.75	0.94	0.61	0.83	no-recurrence-events	
SGD	0.31	0.17	0.43	0.31	0.57	0.36	recurrence-events	5
	0.83	0.69	0.74	0.83	0.57	0.78	no-recurrence-events	
Bagging	0.19	0.08	0.50	0.19	0.66	0.27	recurrence-events	4
	0.92	0.81	0.73	0.92	0.66	0.81	no-recurrence-events	
NaiveBayes Updateable	0.46	0.16	0.54	0.46	0.67	0.49	recurrence-events	3
	0.81	0.15	0.76	0.84	0.67	0.81	no-recurrence-events	

N.B: ROC (Receiver operating characteristics curve)

The algorithm used for the classification can be explained as:

(i) Step 1: Load the important libraries
(ii) Step 2: Import the dataset and extract the X variables and Y separately.
(iii) Step 3: Divide the dataset into train and test
(iv) Step 4: Initializing the Decision Table, J48, SGD, Bagging, NaiveBayesUpdateable classifier model
(v) Step 5: Fitting the Decision Table, J48, SGD, Bagging, NaiveBayesUpdateable classifier model classifier model
(vi) Step 6: Coming up with predictions
(vii) Step 7: Visualization of the predictions.

3. RESULTS AND DISCUSSION

Table 1 contains a description of the datasets used in this study. The dataset comprises 9 input characteristics, 699 occurrences, and missing values. Table 2 shows a description of the datasets utilized after preprocessing. In comparison to the original datasets, the number of input features decreases from nine (9) to six (6), and the number of occurrences decreases from 699 to 286 with no missing values. Table 3 compares the performance of the chosen method in terms of successfully categorized instances, erroneously classified instances, and accuracy. The decision table accuracy (77.0 percent) outperforms J48, NaiveBayesUpdateable, Bagging, and SGD in that order. On the other hand, we may state that the decision table has the highest value of properly categorized instances and the lowest value of wrongly classified instances compared to the other classifiers. Table 4 shows that the performances of the five classifiers are promising; with the Decision table method outperforming the other four by a narrow margin in terms of accuracy, ROC area, and f1-

DOI: http://dx.doi.org/10.17509/xxxx.xxxx
p- ISSN 2775-6793 e- ISSN 2775-6815

score. In conclusion, the Decision Table demonstrated its power in terms of efficacy and efficiency based on accuracy and memory. In comparison to a substantial amount of research on Breast-cancer-Wisconsin found in the literature that compares classification accuracies of data mining algorithms, our experimental results achieve the highest value of accuracy (75.0 percent) in classifying breast cancer datasets using stratified 10-fold cross-validation. In identifying breast cancer datasets, the decision table surpasses other classifiers in terms of accuracy, sensitivity, specificity, and precision.

4. CONCLUSION

The performance of the Decision Table, J48, SGD, bagging, and NaiveBayesUpdateable algorithms on publically accessible benchmark datasets was explored in this paper. The five datasets were pre-processed in the same way and used to train and evaluate the specified supervised learning algorithms. In the performance evaluation, parameters such as accuracy, precision, recall, ROC area, and F1 score were employed. According to the study, the Decision table algorithm has the most promising performance across all criteria utilized for evaluation. The technique utilized in this study is thought to have brought more insights into breast cancer research.

5. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

6. REFERENCES

Asri, H., Mousannif, H., Al Moatassime, H., and Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science, 8*3, 1064-1069.

Rajput, A., Aharwal, R. P., Dubey, M., Saxena, S., and Raghuvanshi, M. (2011). J48 and JRIP rules for e-governance data. *International Journal of Computer Science and Security (IJCSS), 5*(2), 201.

Shah, P. J., and Shah, T. (2021). Identification of breast tumor using hybrid approach of independent component analysis and deep neural network. *International Journal of Intelligent Systems and Applications in Engineering, 9*(4), 209-219.

Wu, J., and Hicks, C. (2021). Breast cancer type classification using machine learning. *Journal of Personalized Medicine, 11*(2), 61.

DOI: http://dx.doi.org/10. 17509/xxxx.xxxx
p- ISSN 2776-6098 e- ISSN 2776-5938