Accuracy Assessment of Several Machine Learning Algorithms for Breast Cancer Diagnosis

Shweta Saraswat [#1], Bright Keswani [*2], Ruchi Kulshrestha[*3],

Shashi Sharma[*4], Nootan Verma[*5], Saibee Alam[*6]

Department of Computer Sc. Suresh Gyan Vihar University, Jaipur, Rajasthan, India [#1,* 2]

Department of Computer Sc. JECRC University, Jaipur, Rajasthan, India [*3,*4,*5,*6]

Article Info
Page Number: 12578-12587
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 October 2022
Revised: 28 November 2022
Accepted: 15 December 2022

Abstract: A number of critical metrics, such as confusion matrix, precision, recall, F1-score, support, and accuracy, were used to evaluate and rank several machine learning classification methods. Linear discriminant analysis, logistic regression, decision tree classification, k-nearest neighbors, gaussian naive bayes, support vector machine, and random forest were only few of the statistical methods used to evaluate the Breast Cancer Wisconsin Diagnostic dataset. The accuracy rate of the Logistic Regression classifier is higher than that of its competitors. The assignment was accomplished in the Anaconda environment using the Python programming language and the Scikit-learn package.

Keywords: Benign, Malignant, Classification model.

I. Introduction

Detecting breast cancer using standard methods may be inefficient, expensive, or both, which makes treating the condition even more difficult than it already is. Thanks to the availability of big data, we are no longer required to prioritize quantity over quality [1]. AI algorithms are widely used in the healthcare sector due to their adaptability and capacity for a wide range of tasks, including disease prediction and diagnosis, medication cost reduction, and real-time decision-making. [2] A comparison of seven classification systems (Linear Discriminant Analysis, Logistic Regression, Decision Tree Classifier, K-Nearest Neighbors, Gaussian Naive Bayes, Support Vector Machine, and Random Forest) strengthens the case for supporting funding for cancer research and treatment. To advance health care generally and cancer treatment in particular, artificial intelligence must be used to improve the clinical performance of information and communication technologies [3]. Artificial intelligence's ability to sift through mountains of patient data in search of trends that human doctors might overlook could improve cancer detection and treatment. This might result in patients receiving better, more specialized care. It is possible to diagnose and predict breast cancer using a variety of machine learning techniques. The table below provides a summary of some of the resources we use in this line of work. In order to successfully treat breast cancer, early detection is essential, and a number of methods have shown promise in this regard. If there is a way to improve the diagnostic efficacy and accuracy of breast cancer tests, that needs to be explored further.

Classification Models in Machine Learning	Description	Highlights
Linear Discriminant Analysis (LDA)	reduces data storage dimensions while maintaining categorization accuracy.	It's simple and computationally efficient. It can perform well with more characteristics than training examples. The system can handle multicollinearity.
Logistic Regression (LR)	It predicts categorical dependent variables using specified independent factors.	accurate and descriptive; low computational needs
Decision tree classifier (CART)	Discusses how other factors might forecast the target variable's values.	It generates simple models with little supervision.
K-Nearest Neighbors (KNN)	Thinks the new case or data is similar to existing cases and places it in the category that is most comparable.	Local approximation, no prediction method, heavy processing
Gaussian Naïve Bayes (NB)	believes well-defined that each parameter (feature or predictor) independently predicts the output variable.	Categorical predictors alone are suitable for little train data.
Support Vector Machine (SVM)	identifies the maximal marginal hyperplane (MMH) by clustering related data points.	It works nicely when classes are well defined and have high dimensions.
Random Forest(RF)	The class that the majority of trees choose is the random forest's output.	high prediction accuracy; limited explainability; works well with both continuous and categorical predictors

Table 1: Several machine learning methods with their highlights are used in this article.

The research designs and findings of earlier studies on the diagnosis of breast cancer are presented in Part 2 of this article, and in Section 3 we establish the methodology that will direct our own investigation. We present and discuss the experimental results in Section 4. The conclusions are given in section 5.

II. **Bibliographical Context**

The detection and prognosis of breast cancer might benefit from the use of machine learning methods. Breast cancer knowledge has advanced thanks to the Wisconsin data collection, mammography pictures, the SEER database, and patient records. These methods might help researchers choose characteristics for further study. This kind of study is essential. Using a number of supervised machine learning methods, Sudarshan Nayak [4] identifies 3D breast cancer pictures. He concludes that support vector machines are the way to go. B.M. Gayathri [5] explains why RVM is superior to other machine learning algorithms for diagnosing breast cancer even when the number of variables is reduced and shows that RVM can reach 97% accuracy in its predictions. Latchoumiet TP [6] used weighted particle swarm (WPSO) optimization in tandem with the support vector machine (SSVM) to improve classification accuracy to 98.4%. Ahmed Hamza Osman [7] used a probabilistic vector support machine in conjunction with a clustering approach to provide very accurate predictions (99.10 percent) for Wisconsin breast cancer (WBCD). Using the Breast Cancer Wisconsin (Diagnostic) Dataset, P. P. Sengar, M. J. Gaikwad, and A. S. Nagdive trained and evaluated the Logistic Regression and Decision Tree algorithms for detecting breast cancer. The likes of precision and recollection are examples of performance indicators. Both DT and LR achieved a maximum accuracy of 94.40% in this trial. [8] Some of the methods that are used include support vector machines, graphical regression, multi-layer perceptrons, logical neural networks, and logical paired networks. Accuracy levels of 93.75%, 96.19%, 99.04%, 93.76%, 94.84%, and 96.19% were reached using a system developed by Fred M. Agarap (2018) on the Breast Cancer Wisconsin (Diagnostic) Dataset [9]. We conduct a comprehensive comparison of several machine learning techniques and algorithms for the diagnosis and prognosis of breast cancer.

III.Methodology

The first step is to collect all of the relevant information and organize it so that it can be analyzed. Using the heat map, the correlation matrix can be constructed, which helps in deciding which variables to use. One way to do this is by label encoding. The original dataset is divided into two parts, 75% for training and 25% for testing. The new data set was used to create a set for manual training and testing. Many different machine learning algorithms were trained using the datasets. It is feasible to choose a trustworthy algorithm by comparing the accuracy scores. This method has the potential to be especially helpful in sectors where precise predictions are essential, such as healthcare and the financial sector. However, keep in mind that the method used will vary depending on the nature of the problem and the dataset under consideration.

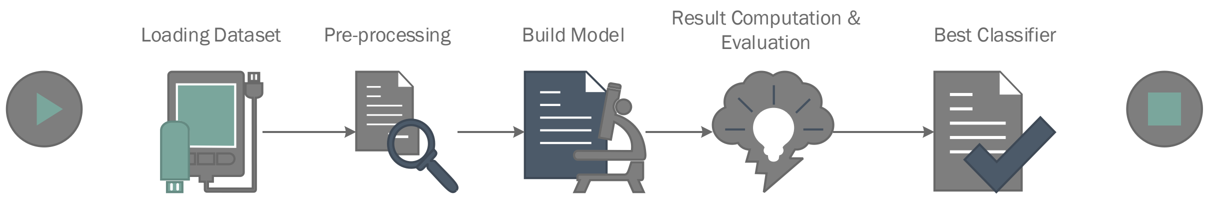

Fig. 1. Process Flow Diagram to find out the best classifier for the given target variable.

Dr. William H. Wolberg of the University of Wisconsin Hospital in Madison, Wisconsin, obtained and organized the data utilized in the research (the Wisconsin Breast Cancer Dataset). The "train-test-split" approach was previously used to this dataset to generate distinct training and testing sets. Several machine learning measures were used to evaluate the efficacy of the studies performed on this dataset. These metrics included the confusion matrix, accuracy, precision, recall, F1-score, and support scores. Making an accurate algorithm to detect breast cancer is our main objective. The Breast Cancer Wisconsin Diagnostic dataset was used to evaluate the efficacy of numerous machine learning frameworks, such as Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LR), Classification and Regression Trees (CART), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Gaussian Nave Bayes. The new building's layout is seen in Figure 1. Our approach quickly executes four preprocessing phases after data collection, including attribute selection, target role creation, data cleaning, and feature extraction. New therapies for breast cancer may be predicted using machine learning algorithms fed with this information.[10] Using newly-tagged data, we assess how well an algorithm performs. The demand for labeled data may be halved by using a train-test split. Using Sklearn's "train-test-split" tool, a dataset may be divided in two for evaluation purposes by selecting appropriate models. While both are used, only one really guarantees a model's correctness. Your data will be automatically organized into digestible chunks. Data is often randomly partitioned in half for use in training and testing. This method is often used since it saves time and effort. If your dataset is too tiny, if you need too many options, or if your dataset isn't dispersed equally, you shouldn't use this approach.

Seventy-five percent of the data was enough for our machine learning model's training. To assess the model's efficacy, we will analyze a subset (about 25%) of the raw data. The purpose of this research is to determine which models have the potential to detect breast cancer at an early stage.

1. **Acquiring Datasets**

Through the breast cancer database at the University of Wisconsin Hospitals and Clinics, you may obtain the Breast Cancer Wisconsin Diagnostic dataset. A digital picture of a breast cancer FNA specimen was used to generate this data set. Due to the large amount of previously published work utilizing this dataset, it was necessary to remove 8 entries to assure reliable findings. There are 355 cases where the form is benign and 206 cases where it is malignant when the ID, radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension are all integers.

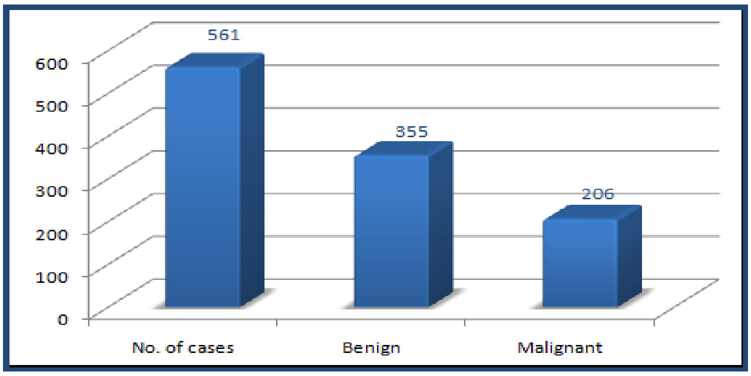

Fig. 2: Diagnostic datasets for breast cancer in Wisconsin

2. Lab Setup for Experiments

Python's Scikit-learn package was used for all of the machine learning experiments in this review. Scikit-learn (also known as sklearn) is a robust machine learning library that is both open source and written in Python. [11] It is pre-installed with numerous classification methods and is compatible with the NumPy and SciPy Python scientific and numerical libraries.

3. Performance Metrices:

Because classification methods produce discrete results, we need a metric that allows us to compare various classes. Classification metrics offer multiple ways to evaluate a model's success at labeling data. The following measures will be discussed for assessing classification models: The area under the curve (AUC) and the F1 score are two examples. Before any measure can be used successfully, its benefits and drawbacks must be understood.

a) Accuracy

The accuracy of a classification system can be expressed as a percentage by dividing the total number of predictions by the number of correct predictions.

b) Confusion Matrix

Evaluation criteria are shown as cells in the confusion matrix. Think about each piece individually: Correctly predicted class sample proportion (TP) is displayed. The frequency with which your model correctly predicted the negative class is reported by the True Negative (TN) statistic. The FP count of negative class samples was not as high as you had predicted in your model. A statistician would call this a Type-I error. This erroneous spot in the ambiguity matrix is highlighted by the null hypothesis. False negatives are examples of the target class for which your model made an incorrect prediction (FN). Error of type 2 statistics. Depending on the null hypothesis, the confusion matrix may incorporate this competing theory.

		Supposed	
		Is Cancer	possesses no cancer
Facts	Is cancer	TP	FP
	possesses no cancer	FN	TN

c) Precision

True positive rates as a percentage of all positive predictions are what we call precision.

$$P = \frac{TP}{TP+FP} = \frac{\text{Cancer patients correctly identified}}{\text{Cancer patients correctly identified+incorrectly labelled cancer patients as non-cancerous}}$$

A large percentage of false positives (precision 0.5) indicates that your classifier is overfitted or has poorly calibrated hyperparameters.

d) Recall/Sensitivity/Hit-Rate

Recall computes the fraction of true positives relative to all true positives in the ground truth. For your classifier to have a recall score below 0.5, the hyperparameters are either overfit or poorly calibrated.

$$R = \frac{TP}{TP+FN} = \frac{\text{Cancer patients correctly identified}}{\text{Cancer patients correctly identified+incorrectly labelled non-cancer patients as cancerous}}$$

e) F1-score

The F1 score considers both accuracy and recall. By taking the mathematical mean of the two scores, we get an F1 grade. Both have a same structure at its core:

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

f) Support

The frequency with which a class appears in the data set serves as a gauge of support for it. While training our model, we first evaluate seven different machine learning classification techniques and compare their performance. Classifier accuracy may suffer if the training data doesn't follow a normal distribution. Instead than rethinking how performance is evaluated, we examine the problems with the status quo.

IV. Results and Discussion

While training our model, we first evaluate seven different machine learning classification techniques and compare their performance.

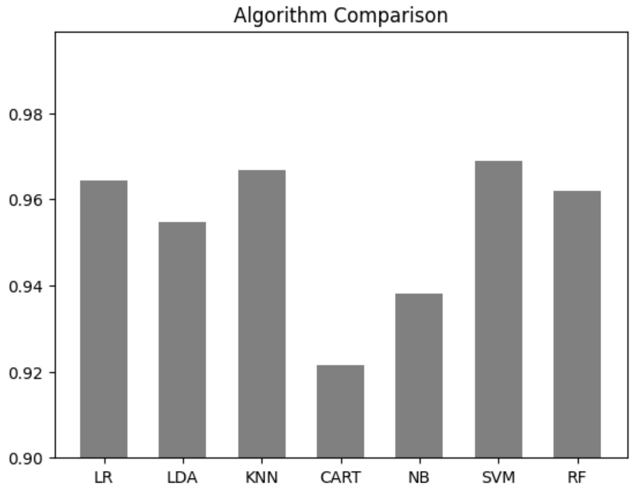

Fig. 3. The accuracy % of training Wisconsin Breast Cancer Diagnostic datasets for various machine learning models.

Researchers in Wisconsin analyzed the state's breast cancer statistics using machine learning techniques. Based on the confusion matrix, accuracy, precision, sensitivity, F1 score, and support measures, we chose the most efficient method. If the solution to a classification problem can be classified in multiple ways, a perplexity matrix can be utilized to evaluate the situation. The standard for evaluating a classification algorithm is its precision. The frequency with which our machine learning model makes accurate predictions could serve as a proxy for its efficacy. The sensitivity of your machine learning model will depend on how frequently it makes accurate predictions. The F1 score takes both precision and sensitivity into consideration.

Algorithm	Accuracy (%)	
	Training set Data shape (448, 31)	Test set Data Shape (113, 31)
Logistic Regression	96.42	99.11
Linear Discriminant Analysis	95.47	94.59
K-Nearest Neighbor	96.66	96.46
Decision Tree Classifier	92.14	95.18
Gaussian NB	93.80	96.36
Support Vector Machine	96.90	96.46
Random Forest	96.19	94.69

Table 2: Comparison of various machine learning algorithms in terms of accuracy

The results in Table 2 show that there is a degree of variation in the classifiers' accuracy. Yet, it is essential to keep in mind that this is not the only consideration when selecting an appropriate machine learning algorithm. The interpretability, scalability, and computing efficiency of a solution are other crucial factors to think about. KNN, SVM, and Random Forest are all quite close to one another in the training phase, however logistic regression consistently outperforms the other classifiers (99.11%) in the testing phase. Table 3 displays the results of the classification models' performance metrics when used with the confusion matrix findings and the precision sensitivity f1 score to distinguish between benign and malignant tumors.

Machine learning model	Cases: Benign:0 Malignan:1	Confusion Matrix		Precision	Recall	F1-score	Support
Logistic Regression	0	70	0	0.99	1.00	0.99	70
	1	1	42	1.00	0.98	0.99	43
Linear Discriminant Analysis	0	69	2	0.97	0.98	0.97	70
	1	1	41	0.96	0.97	0.96	43
K-Nearest Neighbor	0	69	1	0.96	0.96	0.96	113
	1	3	40	0.98	0.96	0.96	113
Decision Tree Classifier	0	69	2	0.97	0.98	0.97	70
	1	2	40	0.97	0.96	0.96	43
Gaussian NB	0	68	2	0.97	0.94	0.96	70
	1	4	38	0.96	0.93	0.94	43
Support Vector Machine	0	69	1	0.97	0.96	0.96	113
	1	3	40	0.96	0.96	0.96	113
Random Forest	0	69	1	0.93	0.99	0.96	70
	1	5	38	0.97	0.88	0.93	43

Table 3: Comparative Evaluation of Various Machine Learning Algorithms, Regardless of Performance Metrics

V. Conclusion

On the Wisconsin Breast Cancer Diagnostic Dataset (WBCD), we used a variety of machine learning and statistics techniques, including SVMs, RFs, LRs, LR-DTrees, NBayes, and KNNs. We built a confusion matrix using the criteria accuracy, sensitivity, precision, F1-score, and support to find the best machine learning approach. Each method was developed using the Python programming language and the Anaconda Python development environment, namely the scikit-learn package. Logistic Regression was the most helpful in terms of efficiency, accuracy, and recall. The best way to diagnose and predict the course of breast cancer is through logistic regression. We need to do a thorough examination of the WBCD database to grasp the implications of our findings. The database's findings must be verified by doing the same analysis on other data sets. We plan on employing our own and other researchers' machine-learning methods to apply these fine-tuned parameters to ever-larger data sets covering an ever-expanding range of illnesses. This will help us confirm that our first findings are not specific to the data we used to generate them. This may lead to discoveries concerning disease mechanisms and improvements in patient treatment.

References

1. Hanis, T. M., Islam, M. A., & Musa, K. I. (2022). Diagnostic Accuracy of Machine Learning Models on Mammography in Breast Cancer Classification: A Meta-Analysis. Diagnostics, 12(7), 1643.
2. Turcu-Stiolica, A., Bogdan, M., Dumitrescu, E. A., Zob, D. L., Gheorman, V., Aldea, M.,... & Lungulescu, C. V. (2022). Diagnostic Accuracy of Machine-Learning Models in Predicting Chemo-Brain in Breast Cancer Survivors Previously Treated with Chemotherapy: A Meta-Analysis International Journal of Environmental Research and Public Health, 19(24), 16832
3. Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis Procedia Computer Science, 191, 487–492.
4. Nayak, S., & Gope, D. (2017, June). Comparison of supervised learning algorithms for RF-based breast cancer detection in the 2017 Computing and Electromagnetics International Workshop (CEM) (pp. 13–14). IEEE.
5. Preethi, N., & Jaisingh, W. (2022). Analysis of Fine Needle Aspiration Images by Using Hybrid Feature Selection and Various Machine Learning Classifiers In Data Science and Security: Proceedings of IDSCS 2022 (pp. 383–392) Singapore: Springer Nature Singapore.
6. Latchoumi, T. P., & Parthiban, L. (2017) Abnormality detection using weighted particle swarm optimization and smooth support vector machines Biomedical Research, 28(11), 4749–4751.
7. Osman, A. H. (2017). An enhanced breast cancer diagnosis scheme based on the two-step SVM technique International Journal of Advanced Computer Science and Applications, 8(4)
8. Sengar, P. P., Gaikwad, M. J., & Nagdive, A. S. (2020, August). A comparative study of machine learning algorithms for breast cancer prediction In 2020, the Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 796–801) IEEE.

9. Agarap, A. F. M. (2018, February). On breast cancer detection: an application of machine learning algorithms to the Wisconsin diagnostic dataset In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (pp. 5–9).
10. Shweta Saraswat, Bright Keswani, Vandna Sharma, Vrishit Saraswat, & Monica Lamba. (2022). Mammograms-Based Breast Cancer Detection Using AI Image Processing Techniques. Journal of Coastal Life Medicine, 11(1)
11. Fabian Pedregosa and all (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830