

Breast Cancer Prediction and Diagnosis: A Comparative Study Analysis Using Machine Learning Algorithms

Anuj Mangal¹ and Navin Kumar Agrawal²

^{1,2}GLA University / Department of CEA, Mathura, India

Email: mangalanuj@gmail.com, navin.agrawal@gla.ac.in

Abstract— The motive of this article is to discover the algorithms for machine learning for the detection and diagnosis of breast cancer. Through rigorous analysis and comparison, we aim to identify the most effective algorithm in predicting and diagnosing breast cancer. Notably, with an exceptional accuracy of 97.2%, our results show that Support Vector Machine is the best performer. Using Scikit-learn toolkit, in conjunction with the Google Colab and Anaconda environment, this study highlights the critical role that machine learning plays in enhancing cancer management approaches. Apart from attaining remarkable precision levels, the application of these algorithms for cancer prognosis and diagnosis has other noteworthy benefits. This customized approach improves treatment results and problems while lowering the likelihood of adverse effects. By using machine learning approach, this research focus on improving the care for breast cancer data and open the door to a more aggressive and successful strategy for battling this common and deadly illness.

Index Terms— SVM, Random Forest, Logistic Regression, Decision tree, Breast Cancer, Accuracy, Precision.

I. INTRODUCTION

Breast cancer is dangerous syndrome among women worldwide the count of cancer diagnoses has almost doubled in the last 20 years, from an expected ten million in 2000 to approximately 20 million in 2020[1]. This study examines how well various classifiers work, focusing on metrics like recall (sometimes referred to as sensitivity), confusion matrix, accuracy, and precision. These metrics are considered performance measures since they represent high-performance approaches to sickness prediction and diagnosis. Real-time decision making is made easier by these algorithms, which helps save lives and save money on medical expenses. Therefore, incorporating machine learning methods into cancer treatment is essential to improving patient outcomes and reducing the incidence of cancer worldwide. This paper is structured as follows: Section 2 reviews earlier analyses on breast cancer diagnosis methods and results. Section 3 outlines our proposed methodology. Section 4 presents and analyzes the experimental results in detail. Finally, Section 5 concludes the paper.

II. RELATED WORK

Researchers have conducted extensive studies on breast cancer by using different classifier and dataset such as SEER dataset, mammogram images, the Wisconsin Dataset, and datasets from various hospitals [2]. Utilizing these datasets, researchers extract and select various features to conduct their studies.

Among these studies, some significant research. The author A. Melek, S. Fakhry and T. Basha, [3] demonstrates significant advancements in breast cancer risk prediction through the incorporation of deep learning techniques and spatiotemporal analysis of mammography data. In contrast, By using an improved weighting of the particle swarm (WPSO) based on the SSVM for classification, Latchoumiet TP [7] was able to attain a 98.4% classification accuracy. By combining the SVM algorithm with a clustering technique and an effective probabilistic vector support machine, Ahmed Hamza Osman [8] presented a method for identifying Wisconsin breast cancer (WBCD) with an astounding prediction accuracy of 99.10%. The findings indicate that SVM exhibits the most favourable performance among the algorithms assessed.

II. METHODOLOGY

Finding a highly precise and effective approach for breast cancer diagnosis is our main objective. Researchers used machine learning classifiers on the Breast Cancer Wisconsin Diagnostic dataset, such as Support Vector Machine (SVM) [9], Logistic Regression [10], Random Forests [11], and Decision tree (C4.5) [12], to accomplish it. Data cleaning is done during the pre-processing phase to deal with missing values and guarantee data integrity. To improve the models' predictive ability, relevant features are subsequently chosen, the target variable indicating the diagnosis (benign or malignant) is established, and important features are extracted. Data collection is the first phase in the method we use. Next comes pre-processing, which is divided into four stages: feature extraction, target role setting, attribute selection, and data cleaning. Using the information that has been collected, machine learning algorithms are able to forecast breast cancer symptoms for fresh measures.

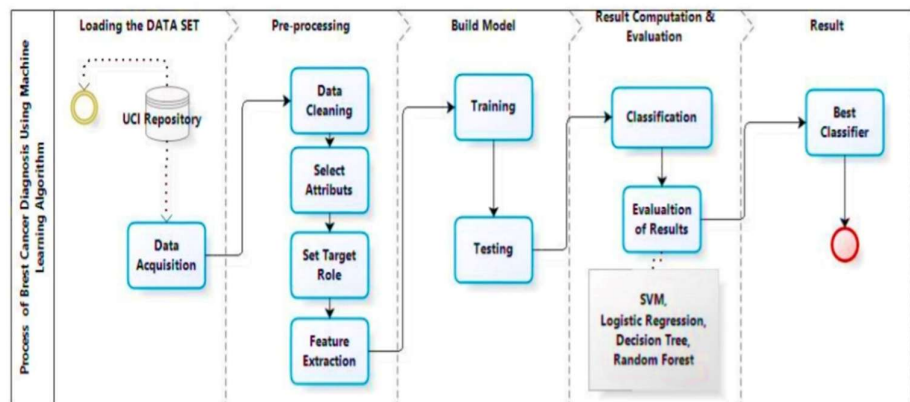


Figure 1. Process Flow diagram

III. MACHINE LEARNING ALGORITHMS

These algorithms have previously proved in early identification and individualized treatment plans by individually enhancing diagnostic accuracy and predictability in breast cancer prediction.

A. Support Vector Machine (SVM)

In order to convert the provided data into a higher-dimensional space where linear division is possible, SVM uses kernel functions. The linear, polynomial, and radial basis function (RBF) kernels are examples of common kernels. Because non-linear correlations are frequently present in complex medical data, the RBF kernel in particular is useful for predicting breast cancer.

SVM is applied in breast cancer prediction to categorize tumor samples according to features taken from genetic profiles, histological data, or medical imaging. The algorithm's strength is its ability to handle data that is both linearly and non-linearly separable by utilizing several kernel functions.

B. Random Forests or Random Decision Forests

Random Forest is an ensemble learning method that has proven highly effective for breast cancer prediction due to its robustness, accuracy, and ability to handle complex datasets. During training, this algorithm creates a large number of decision trees, and it outputs the mean prediction (regression) or the mode of the classes (classification) for each tree.

This is particularly crucial for medical applications, as datasets may be unbalanced, noisy, and contain a disproportionately high proportion of benign than malignant instances. Effective medical decision-making depends on the model remaining robust and performing well on unseen data, which is ensured by the ensemble technique. Random Forest also has the important benefit of being interpretable.

C. Logistic Regression

A basic and popular statistical technique for binary classification tasks, like determining whether breast cancer will manifest or not, is logistic regression. Because it uses less computing power than more intricate algorithms, it is a sensible option for preliminary research and real-time applications. Furthermore, logistic regression offers a strong foundational model that can be improved upon by more advanced methods as needed.

Logistic regression is frequently used in breast cancer prediction to examine data from biopsies, mammograms, and other diagnostic testing. All things considered, logistic regression is a useful tool in the predictive modeling of breast cancer, helping to promote early diagnosis and better patient outcomes due to its simplicity of use, interpretability, and efficacy in handling binary classification tasks.

D. Decision Tree C4.5

It is a well-known decision tree classifier for applications like breast cancer prediction. The C4.5 method, created by Ross Quinlan, incorporates significant improvements that make it more appropriate for use in real-world medical diagnostic applications. The result of C4.5 is a highly interpretable tree structure that facilitates medical practitioners' understanding of the decision-making process. Based on the characteristics of the patient, a series of decisions are made along each path from the root to a leaf node, which results in a prognosis of either benign or malignant breast cancer. In clinical contexts, where it's critical to be able to justify and clarify clinical decisions, this openness is vital.

IV. DATASET ACQUISITION

We used the Breast Cancer Wisconsin Diagnostic dataset, which was provided by the University of Wisconsin Hospitals, Madison, for our study on breast cancer prediction [13]. This dataset includes digital photographs of breast cancer specimens acquired by fine-needle aspiration (FNA), with a focus on aspects of the visible cell nuclei. It has 569 cases, 357 of which are classified as benign and 212 as malignancy. 11 properties with integer values are included in it, including -Id, -Diagnosis, Compactness, smoothness, texture, area, radius, and symmetry, points of concavity, texture, and fractal dimension. In our study, we employ a range of machine learning techniques, including Random Forests, Decision Tree classifiers, and SVM, to predict the breast cancer cases. These algorithms find patterns and correlations that differentiate between benign and malignant tumors by utilizing the full feature set of the dataset.

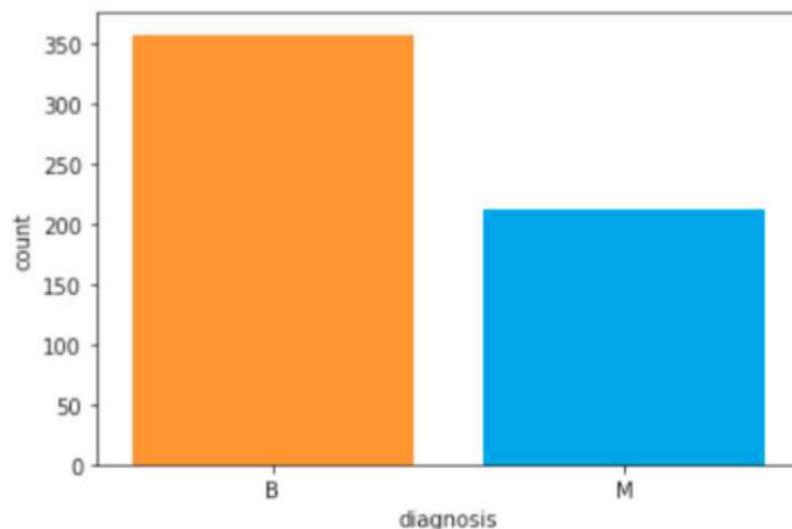


Figure 2. Wisconsin Breast Cancer Diagnostic Data.

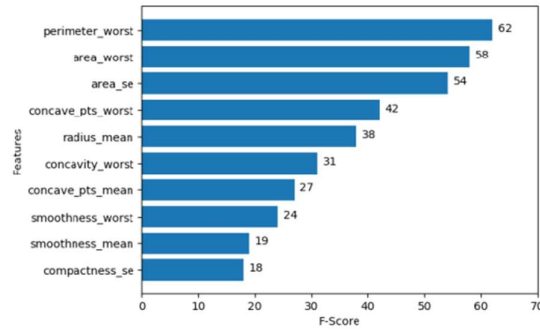


Figure 3. Feature Importance Table For Wisconsin Breast Cancer Dataset.

A. Experiment Environment

For the purpose of predicting breast cancer, four popular machine learning algorithms—Decision Trees, Random Forests, Support Vector Machines (SVM), and Logistic Regression—were selected. This study made use of the vast array of regression and classification techniques that Scikit-learn offers. These tools easily integrate with popular scientific and numerical libraries enabling effective data processing and model implementation.

Among the preprocessing procedures was the imputing of appropriate strategies—such as the mean or median for numerical features—to handle missing values. If any, categorical variables were encoded with methods such as one-hot encoding to make sure the machine learning models worked. For algorithms like SVM that are sensitive to the scale of input data, numerical features were standardized to make sure they were on a same scale.

IV. RESULT AND DISCUSSION

A popular metric called accuracy measures the percentage of accurate forecasts as a percentage of all guesses. Precision is a metric that comes from document retrieval tasks that quantifies the proportion of appropriately identified relevant instances to all instances that the model was able to identify. Sensitivity and recall are synonymous.

The accuracy percentages obtained with the Wisconsin Breast Cancer Diagnostic dataset are shown in Table 1 and Figure 2. The outcomes from the testing and training sets show that different classifiers have different accuracies. Remarkably, SVM outperforms the other classifiers on a regular basis, showing a greater accuracy in the testing set (97.2%).

TABLE I. TRAINING AND TESTING ACCURACY FOR BREAST CANCER DIAGNOSTIC DATASET

Algorithms	Accuracy on Training data (%)	Accuracy on Testing data (%)
SVM	98.4	97.2
Logistic Regression	99.8	96.5
Decision Tree	95.5	95.8
Random Forest	98.8	95.1

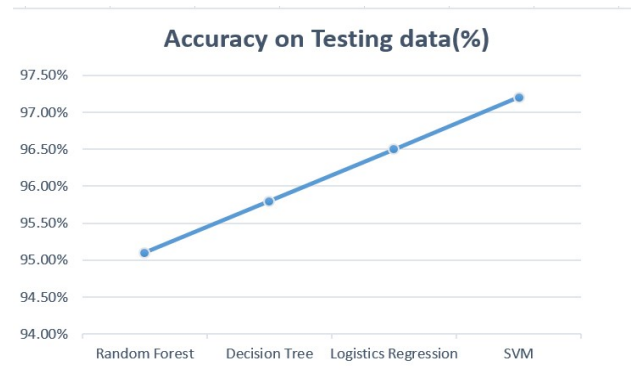


Figure 4. Comparative Graph of different Classifiers

TABLE II. CONFUSION MATRICES

Algorithms	Malignant	Benign	
SVM	201 1	11 356	Malignant Benign
Logistic Regression	111 2	4 54	Malignant Benign
Decision Tree	107 5	1 58	Malignant Benign
Random Forest	196 22	17 335	Malignant Benign

TABLE III. EXPERIMENT RESULTS MEASURE

Algorithms	Precision	Sensitivity	F1-Measure	Class
SVM	0.98	0.94	0.96	Benign
	0.97	0.99	0.98	Malignant
Decision Tree	0.97	0.99	0.98	Benign
	0.94	0.92	0.95	Malignant
Logistic Regression	0.98	0.97	0.96	Benign
	0.93	0.96	0.94	Malignant
Random Forest	0.96	0.94	0.95	Benign
	0.97	0.98	0.97	Malignant

Table 2 uses confusion matrices as a useful tool. Each row represents the rates in an actual class, while each column provides the forecasts for evaluating the classifier. The computed performance metrics of the classification models based on the findings of the confusion matrix, precision, sensitivity, and f1 score for benign(non-cancerous) and malignant(cancerous) conditions are shown in Table 3.

The performance metrics for the various machine learning algorithms applied to the Diagnostic dataset are summarized in Table 3. The F-Measure, Sensitivity, and Precision [14] of each algorithm are given for both the benign and malignant classifications. These metrics show the advantages and disadvantages of each strategy while providing information on how well each model classified the situations.

However, the other classifiers, especially Random Forests, also provide reliable and accurate predictions, making them valuable tools in the diagnostic process [15]. The choice of classifier may ultimately depend on the specific requirements of the diagnostic task, such as the need for interpretability versus raw predictive power.

VI. CONCLUSION

In this research article, we used the Breast Cancer dataset to examine the effectiveness of various machine learning techniques. With a Sensitivity of 0.99 and an F-Measure of 0.98, the SVM performs better than other classifiers with the highest accuracy, and it excels in identifying situations that are malignant in particular. The SVM's efficacy in breast cancer prediction tasks is demonstrated by this high level of performance, which makes it a useful tool for early and precise diagnosis. With a Sensitivity of 0.99, the Decision Tree classifier had exceptional performance as well, particularly in properly predicting benign instances. However, its performance was marginally worse for malignant cases. While both classes had balanced performance from logistic regression, malignant patients showed somewhat lower precision. With strong overall performance and excellent Precision and Sensitivity for both benign and malignant breast cancer prediction, the Random Forests classifier is a dependable option. In conclusion, all of the classifiers examined demonstrated significant promise for breast cancer prediction, even if the SVM seems to be the most successful algorithm for this specific dataset. The final algorithmic decision may be based on particular therapeutic needs, such as the demand for interpretability or the significance of reducing false negatives. Subsequent investigations may concentrate on refining these models even further, incorporating supplementary characteristics, and verifying the outcomes on more extensive and varied datasets to augment the applicability and resilience of breast cancer prediction models. The importance of machine learning in enhancing patient outcomes and developing medical diagnostics is highlighted by this work.

REFERENCES

- [1] 'WHO | Breast cancer', WHO. <http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/> (accessed Jul. 02, 2024).
- [2] Datafloq - Top 10 Data Mining Algorithms, Demystified. <https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144> Accessed December 29, 2015.

- [3] A. Melek, S. Fakhry and T. Basha, "Spatiotemporal Mammography-based Deep Learning Model for Improved Breast Cancer Risk Prediction," 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).
- [4] Mangal, A., Garg, H., & Bhatnagar, C. (2024). A Robust Co-saliency Object Detection Model by Applying CLAHE and Otsu Segmentation Method. *International Journal of Intelligent Systems and Applications in Engineering*, 12(1s), 481-490.
- [5] S. Mishra, M. Angurala and Y. Sharma, "A Next-Gen Approach to Breast Cancer Prediction," 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIHI), Raipur, India, 2023, pp. 1-7.
- [6] G. Shanmugasundaram, S. Balaji, R. Saravanan, V. Malarselvam and S. Yazhini, "SYSTEMATIC ANALYSIS ON BREAST CANCER PREDICTION," 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA), Pondicherry, India, 2018, pp. 1-5.
- [7] Mangal, A., Garg, H., & Bhatnagar, C. (2023). Assessing the performance of Co-Saliency Detection method using various Deep Neural Networks. *Journal of Information Technology Management*, 15(Special Issue: EIntelligent and Security for Communication, Computing Application (ISCCA-2022)), 23-34.
- [8] A. H. Osman, "An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, pp. 158–165, 2017.
- [9] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.
- [10] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215–232.
- [11] Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278–282)*.
- [12] Cherfi, A., Nouira, K., & Ferchichi, A. (2018). Very Fast C4.5 Decision Tree Algorithm. *Applied Artificial Intelligence*, 32(2), 119–137. <https://doi.org/10.1080/08839514.2018.1447479>
- [13] Wolberg, William, Mangasarian, Olvi, Street, Nick, and Street, W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. <https://doi.org/10.24432/C5DW2B>.
- [14] Agrawal, N. K., & Shankhdhar, A. (2022, December). An Enhanced and Effective Approach for Object Detection using Deep Learning Techniques. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 1482-1486). IEEE.
- [15] Shankhdhar, A., Agrawal, N. K., & Srivastava, A. (2021, December). COVID-19 Detection System using Chest X-rays or CT Scans. In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 428-432). IEEE.