A Comparison of the Diagnosis Performance of Machine Learning Algorithms on the Breast Cancer Wisconsin Dataset

Sreerama Murty Maturi
Department of Computer Science and Engineering, GITAM (Deemed to be University), Hyderabad, India
sreeramsssit@gmail.com

C. Dastagiraiah
Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India
dattu5052172@gmail.com

Ponnuru Sowjanya
Department of Computer Science and Engineering, GITAM (Deemed to be University), Hyderabad, India
sowjanya.ponnuru@gmail.com (corresponding author)

Chanumolu Kiran Kumar
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Green Fields, Vaddeswaram, Andhra Pradesh 522237, India
mounikakiran.138@gmail.com

J. S. V. R. S. Sastry
Department of Computer Science and Engineering, GITAM (Deemed to be University), Hyderabad, India
sridatta2002@gmail.com

Received: 29 July 2025 | Revised: 18 August 2025 and 8 September 2025 | Accepted: 9 September 2025

 DOI: https://doi.org/10.48084/etasr.13706

ABSTRACT

Breast cancer survival rate can be dramatically affected if the disease is diagnosed early and the stage is classified accurately. Model-driven or statistical-driven Machine Learning (ML) techniques are powerful and reliable methods of diagnosing breast cancer, with the ability to utilize, sort, and detect patterns in data, and to handle immense databases. This research work considers the Breast Cancer Wisconsin (BCW) dataset to train several ML models (Random Forest, Support Vector Machines, Gradient Boosting, Logistic Regression, k-Nearest Neighbors) to classify images as either benign or malignant and compares the results. The followed methodology involves data preprocessing and feature extraction. All tested algorithms performed satisfactorily with Random Forest surpassing the others.

Keywords-breast cancer; machine learning; diagnosis; classification; Breast Cancer Wisconsin Dataset

I. INTRODUCTION

Early breast cancer diagnosis can significantly improve survival and cure rates, and reduce mortality. Historically, breast cancer diagnosis occurs through mammography, ultrasound, and biopsy, which are heavily dependent on the physician's clinical judgment and, therefore, results may vary. Machine Learning (ML) is an emerging, powerful tool, often utilized in medical diagnosis. ML algorithms can identify and analyze huge amounts of clinical data to recognize patterns and create predictive models to accurately classify breast tumors. BCW provides a large number of metrics that describe tumor features, making it ideal for developing prediction models.

Authors in [1] proposed a deep neural network model with three hidden layers, acquiring 99.12% accuracy, 99.2% recall,

and 99% F1-score. Authors in [2] used various ML algorithms on the BCW dataset. Authors in [3] examined the significance of prompt detection for optimum treatment of breast cancer. The proposed Lasso Logistic Regression (LR) model performed better in attribute selection than the baseline LR model. Authors in [4] proposed an ensemble ML model using LR, DT, and Support Vector Machines (SVM) to categorize breast cancer working on the WBC dataset. The authors applied feature scaling to the data and balanced the dataset, resulting in accuracy of 0.9708, precision of 0.9821, recall of 0.9482, F1-score of 0.9649, and AUC of 0.9678. Authors in [5] classified medical images as malignant or benign, performing early diagnosis using various ML techniques, and investigated the best technology in terms of accuracy. The authors considered k-Nearest Neighbors (KNN), Random Forest (RF), NB, and SVM. The results demonstrated that RF with 98.11% and SVM with 92.6% accuracy performed best. Authors in [6] tested RF, KNN, and LR on cancer detection, with the RF model outperforming the others.

Authors in [7] proposed an RF model for predicting malignancy, achieving 98% accuracy. Among other methods they used exploratory data analysis in their methodology. Only one classifier was used in this study which is a limitation. In [8], the authors proposed a breast cancer segmentation system that used an improved U-Net 3+ neural network with several optimizations to improve segmentation and localization functionality. This system was tested on the INbreast FFDM dataset against other cutting-edge networks. The proposed model set a new standard for segmentation accuracy with a dice score of 98.47% indicating its potential for use in practical breast cancer detection applications. In [9], the authors proposed a novel method for identifying and categorizing breast cancer from mammography images. Homo Morphic Adaptive Histogram Equalization (HMAHE) was used to preprocess the image to improve contrast and remove noise, and the Canny edge detector was used to identify breast boundaries. Tumors were identified and segmented with the Centroid-based Region Growing Segmentation (CRGS) algorithm and the Chaotic Function-based Black Widow Optimization Algorithm (CBWOA) was utilized to select the relevant features. Six groups of these features were then utilized by the Convolutional Deviation Neural Network Classifier (CSDNN).

This work utilizes five popular ML algorithms that are known to perform exceptionally well on classification tasks: SVM, RF, KNN (k-Nearest Neighbors), Gradient Boosting (GBoost), and (LR).

II. THE PROPOSED METHODOLOGY FOR BREAST CANCER PREDICTION

The proposed methodology for diagnosing breast cancer over ML algorithms is a structured sequence in which data are pre-processed, features are extracted, the models are trained, and then validated (Figure 1). By exploiting the benefits of several ML prototypes, this methodical approach guarantees enhanced diagnostic precision and dependability. This method improves the primary identification and diagnosis of breast cancer by combining sophisticated feature selection approaches with various classification algorithms.

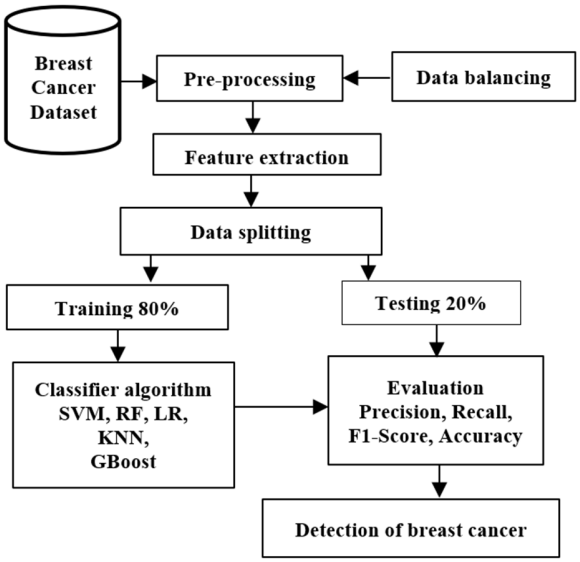

Fig. 1. Proposed system architecture.

The steps of the proposed methodology are:

- Data Collection: The BCW dataset from the UCI Machine Learning repository was utilized. This dataset contains features that have been derived from breast images (specifically digitized images) of masses, labelled as benign or malignant. The dataset includes 569 cases with 30 numerical characteristics measured from digital photos of breast tumors [10].
- Data pre-processing: Any missing values were dealt with by appropriate imputation. The data were normalized or scaled to keet consistency across features. Label encoding was used to encode the categorical labels (malignant and benign). The dataset was divided into training (80%) and testing (20%) parts.
- Feature Selection: Methods such as Correlation Matrix, PCA or Select Best were used to determine the most significant features.
- Model Selection: SVM, RF, KNN, LR, ang GBoost models were considered.
- Model Evaluation: performance metrics were used to assess the models' performance. Cross-validation was used to establish a reliable assessment.
- Training Accuracy Verification: Each model was trained using a 10-fold cross-validation technique to confirm the reported training accuracies. The dataset was split into 10 equal parts, with one part being used for validation and nine parts for training. This process was repeated 10 times. To guarantee stability and avoid overfitting, the mean training accuracy was calculated across folds. All models had a standard deviation across folds of less than 0.5%, demonstrating the consistency and dependability of the claimed training accuracies.

III. IMPLEMENTATION

A. Implementation Steps

- Data Preparation: Python libraries Pandas and NumPy were utilized. Missing data values were corrected and features were scaled using Standard Scalar.
- Model Building: The different ML models were trained utilizing Scikit-Learn. Either Grid Search or Random Search were used for hyper parameter tuning to optimize the models.
- Evaluation: To visualize model performance, Matplotlib and Seaborn were used.
- Testing and Monitoring: Through different test inputs, the application was thoroughly tested. A monitoring method for tracking model performance over time was implemented. This end-to-end implementation provides accurate and streamlined diagnosis of breast cancer by ML algorithms.

B. Performance Metrics

Accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC) were considered. Their definitions are:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

$$\text{F1- score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

where TP, FP, TN, and FN represent the total True Positive, False Positive, True Negative, and False Negative predictions, respectively. The following Confusion Matrices are given in the $\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$ format.

IV. RESULTS AND DISCUSSION

A. Logistic Regression

The LR model was trained in the Scikit-Learn toolkit with 98.90% training and 96.49% testing accuracy. The Confusion Matrix is: $\begin{bmatrix} 66 & 1 \\ 3 & 44 \end{bmatrix}$, whereas the testing set results can be seen in Table I.

TABLE I. LR EVALUATION METRICS

Accuracy (%)	96.49
Precision (%)	98.51
Recall (%)	95.65
F1-Score (%)	97.06

B. K-Nearest Neighbors Model

Many values of k were used to train the KNN classifier. The best results were obtained with k = 5 using the Euclidean distance. The model established trustworthy but marginally worse performance due to its sensitivity to noisy data, with training accuracy of 96.70% and testing accuracy of 95.61%.

Confusion Matrix: $\begin{bmatrix} 66 & 1 \\ 4 & 43 \end{bmatrix}$

TABLE II. KNN EVALUATION METRICS

Accuracy (%)	95.61
Precision (%)	98.51
Recall (%)	94.29
F1-Score (%)	96.35

C. Support Vector Machine

An SVM classifier with a Radial Basis Function (RBF) kernel was considered. We conducted a grid search over some values of the kernel coefficient (gamma) and the penalty parameter (C) in order to find the optimal hyper-parameters. The values:

C = {0.01, 0.05, 0.5, 0.1, 1, 10, 15, 20} and gamma = {0.0001, 0.001, 0.01, 0.1} were considered.

The acquired optimal values were C = 15 and gamma = 0.01, as they bring about a balance between accuracy and complication. The SVM gained testing and training accuracy of 98.25% and 98.90% with this outline. Outstanding overall performance was shown as can be seen in Table III.

Confusion Matrix: $\begin{bmatrix} 67 & 0 \\ 2 & 45 \end{bmatrix}$

TABLE III. SVM EVALUATION METRICS

Accuracy (%)	98.25
Precision (%)	100
Recall (%)	97.10
F1-Score (%)	98.53

D. Random Forest

A total of 130 decision trees (n_estimators = 130), maximum depth of 10, and entropy as the splitting criterion were used to fine-tune the RF model. Max_features = 0.5, min_samples_split = 3, and min_samples_leaf = 2 were additional options. With 99.12% and 99.56 testing and training accuracy, the optimized model performed the best.

Confusion Matrix: $\begin{bmatrix} 66 & 1 \\ 0 & 47 \end{bmatrix}$

TABLE IV. RF EVALUATION METRICS

Accuracy (%)	99.12
Precision (%)	98.51
Recall (%)	100
F1-Score (%)	99.25

E. Gradient Boosting

GBoost was used with the exponential loss function, n_estimators = 180, and a learning rate of 0.1. With this

configuration, the model demonstrated competitive performance with ensemble learning, with a testing accuracy of 97.37% and 100% training accuracy.

Confusion Matrix: $\begin{bmatrix} 65 & 2 \\ 1 & 46 \end{bmatrix}$

TABLE V. GBOOST EVALUATION METRICS

Accuracy (%)	97.37
Precision (%)	97.01
Recall (%)	98.48
F1-Score (%)	97.74

V. DISCUSSION

The RF classifier obtained the highest accuracy across all models and shows solid generalization and capacity in dealing with diverse features. SVM had very good performance, exhibiting robustness with an acceptable balance of precision and recall, after the hyperparameter tuning. GBoost, showed a strong performance leveraging ensemble learning with boosting to reduce bias and variance. LR is a simple and effective model with decent performance, as fitting the data can be simply determined through the linear decision boundaries. The accuracy of KNN was a bit poorer compared to the other models, probably as a consequence of noisy data scaling. Table VI exhibits the model comparison, Figure 2 shows the ROC-AUC comparison, and Figure 3 summarizes the performance metrics for all models.

TABLE VI. MODEL COMPARISON

Model	Accuracy	F1 Score
RF	99.12	99.25
SVM	98.25	98.53
GBoost	97.37	97.74
LR	96.49	97.06
KNN	95.61	96.35

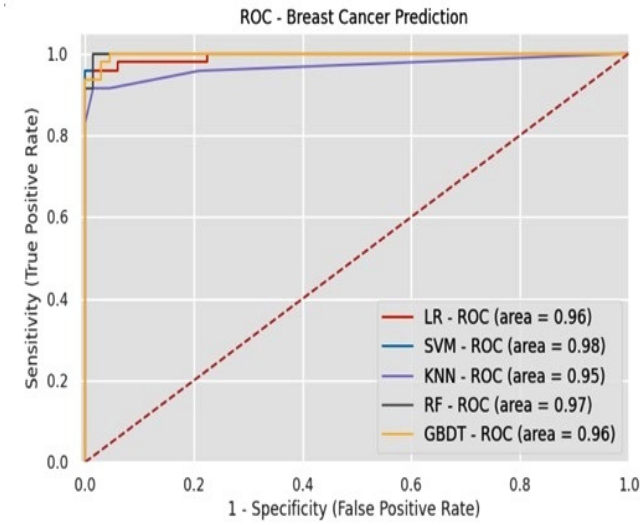

Fig. 2. ROC-AUC comparison.

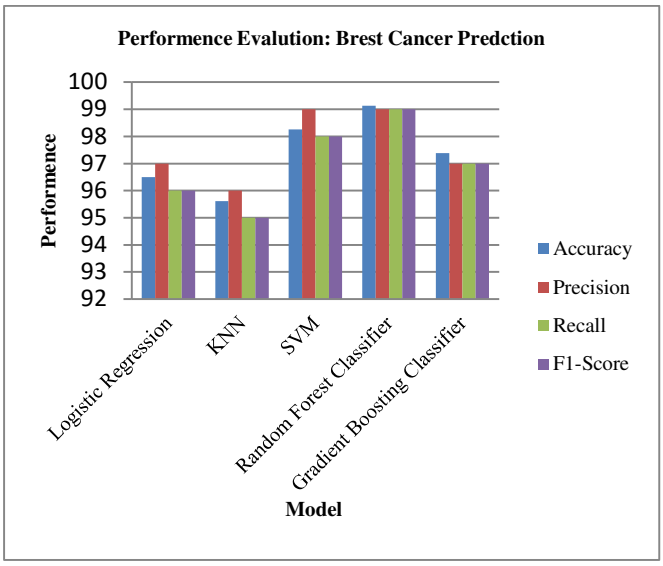

Fig. 3. Model performance summary.

VI. CONCLUSION AND FUTURE SCOPE

In this paper, machine learning techniques were utilized for breast cancer detection. Data pre-processing dealt with missing data values. Highly correlated features were omitted to remove multicollinearity. The data were scaled using Standard Scalar to improve model performance. The performance evaluation was carried out through the use of confusion matrices and classification metrics such as accuracy, recall, F1-score, precision, and ROC.

The novelty of this paper is the ability of breast cancer prediction while integrating optimized attributes with the computationally most efficient algorithms to reach the highest diagnostic accuracy while using minimized model complexity, making the system suitable for integration into healthcare decision support systems.

Future enhancement of this work can be conducted in various sectors, such as interactive web applications, data visualization, real-time analysis and reporting, model comparison dashboard, mobile responsiveness, API integration, authentication, and security. Patient history can be stored, allowing the ability to track the therapy outcomes over time. Therapy schedules and line charts can be used to visualize patient progress. By implementing these aspects, the project can extend into a complete breast cancer diagnosis and monitoring system, valuable and usable for medical practitioners regardless of location.

REFERENCES

[1] A. Bekkouche, M. Merzoug, M. Hadjila, and W. Ferhi, "Towards Early Breast Cancer Detection: A Deep Learning Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17517–17523, Oct. 2024, https://doi.org/10.48084/etasr.8634.

[2] R. Dwivedi, K. Ramdev, Er. M. Jain, and N. Agrawal, "Classification of Breast Cancer Diagnosis Using Machine Learning Algorithms," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, Jan. 2023, pp. 1–8, https://doi.org/10.1109/ICCCI56745.2023.10128478.

[3] R. Rekha and K. L. Vinoci, "Wisconsin Breast Cancer Detection Using L1 Logistic Regression," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, Coimbatore, India, Oct. 2023, pp. 166–169, https://doi.org/10.1109/ICISCoIS56541.2023.10100387.

[4] A. Singh and K. S. Kaswan, "Breast Cancer Diagnosis using Soft Voting Classifier Approach," in *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*, Sonipat, India, Feb. 2024, pp. 292–297, https://doi.org/10.1109/INNOCOMP63224.2024.00055.

[5] K. G. Devi, K. Balasubramanian, and L. A. Ngoc, Eds., *Machine Learning and Deep Learning Techniques for Medical Science*. Boca Raton, FL, USA: CRC Press, 2022.

[6] N. N. Caleb, S. Zwalnan, and C. Pahalson, "Breast Cancer Diagnosis using Machine Learning Approach," *International Journal of Advanced Research in Science, Communication and Technology*, vol. 8, no. 1, pp. 459–466, Aug. 2021, https://doi.org/10.48175/IJARSCT-1880.

[7] Jamal, J. H. Antor, R. Kumar, and P. Rani, "Breast Cancer Prediction Using Machine Learning Classifiers," in *2022 5th International Conference on Advances in Science and Technology (ICAST)*, Mumbai, India, Sep. 2022, pp. 456–459, https://doi.org/10.1109/ICAST55766.2022.10039656.

[8] S. M. Shaaban, M. Nawaz, Y. Said, and M. Barr, "An Efficient Breast Cancer Segmentation System based on Deep Learning Techniques," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12415–12422, Dec. 2023, https://doi.org/10.48084/etasr.6518.

[9] N. Behar and M. Shrivastava, "A Novel Model for Breast Cancer Detection and Classification," *Engineering, Technology & Applied Science Research*, vol. 12, no. 6, pp. 9496–9502, Dec. 2022, https://doi.org/10.48084/etasr.5115.

[10] W. Street, W. Wolberg, and O. Mangasarian, "Breast Cancer Wisconsin (Diagnostic)." UCI Machine Learning Repository, 1993, https://doi.org/10.24432/C5DW2B.