STM JOURNALS

Research & Reviews: A Journal of
Bioinformatics

https://journals.stmjournals.com/rrjobi

ISSN: 2393-8722
Volume 12, Issue 3, 2025
September—December
DOI (Journal): 10.37591/RRJoBI

Review RRJoBI

Breast Cancer Detection Using Machine Learning: A Comparative Analysis of Supervised Learning Algorithms

Niral Mahajan[1], Ashwini Kukade[2,*]

Abstract

Globally, breast cancer remains a predominant cause of mortality among women, highlighting the urgent need for timely and precise diagnostic approaches. This research explores the application of machine learning algorithms – including Logistic Regression, SVM, Naive Bayes, KNN, and Random Forest – on the Wisconsin Breast Cancer Dataset for effective tumor classification. Key pre-processing steps, such as missing value handling, feature scaling, and dimensionality reduction, were employed to improve model performance. The study evaluated multiple machine learning models for breast cancer detection using performance metrics such as accuracy, precision, recall, and F1-score. Among the evaluated algorithms, the Naive Bayes classifier demonstrated the highest accuracy and overall reliability, underscoring its potential for medical diagnostic applications. Feature selection played a critical role in improving model efficiency, highlighting its importance in predictive healthcare analytics. This research emphasizes the value of AI-driven techniques in developing scalable, non-invasive diagnostic systems, which can enhance early detection and clinical decision-making. The findings support integrating machine learning approaches into medical workflows to advance precision medicine and improve patient outcomes.

Keywords: Breast cancer, Machine Learning (ML), logistic regression, support vector machine (SVM), Gaussian Naive Bayes, K-Nearest Neighbors (KNN), random forest

INTRODUCTION

Breast cancer is among the most prevalent and life-threatening diseases affecting women worldwide, contributing significantly to cancer-related mortality. According to the World Health Organization (WHO), it is the most diagnosed cancer globally, with millions of new cases reported annually. Due to its biological complexity and heterogeneity, early and accurate diagnosis is vital for improving survival rates and guiding effective treatment strategies. Traditional diagnostic techniques, such as mammography, biopsy, and histopathological analysis, though widely adopted, are often time-consuming, costly, and susceptible to human error. To address these limitations, Artificial Intelligence (AI), and Machine Learning (ML) have emerged as transformative technologies in medical diagnostics, facilitating faster, more accurate, and automated detection processes.

***Author for Correspondence**
Ashwini Kukade
E-mail: ashwinikukade90@gmail.com

[1]Web Developer, Department of AI, G H Raisoni College of Engineering, Nagpur, Maharashtra, India
[2]Professor, Department of AI, G H Raisoni College of Engineering, Nagpur, Maharashtra, India

Received Date: March 26, 2025
Accepted Date: August 31, 2025
Published Date: November 11, 2025

Citation: Niral Mahajan, Ashwini Kukade. Breast Cancer Detection Using Machine Learning: A Comparative Analysis of Supervised Learning Algorithms. Research & Reviews: A Journal of Bioinformatics. 2025; 12(3): 46–52p.

Machine learning techniques have shown considerable promise in detecting and classifying cancerous lesions by analyzing medical imaging, cellular morphology, and structured patient data. These algorithms can process large datasets, uncover complex patterns, and reliably distinguish between malignant and benign tumors. The Wisconsin Breast Cancer Dataset (WBCD) is frequently utilized in machine learning studies due to its comprehensive numerical features describing various tumor characteristics. In this study, multiple supervised learning algorithms – including Logistic Regression, Support Vector Machine (SVM),

Gaussian Naive Bayes, K-Nearest Neighbors (KNN), and Random Forest – were employed to assess their effectiveness in breast cancer detection. The research methodology involves rigorous data preprocessing, including missing value imputation, feature selection, and data normalization, to enhance model accuracy and generalization. Model performance was evaluated using key classification metrics, including accuracy, precision, recall, and F1-score. Among the algorithms tested, Naive Bayes demonstrated superior classification accuracy and robustness, highlighting its strong potential for clinical application. This study contributes to the evolving landscape of AI-driven healthcare by offering a reliable, scalable, and non-invasive solution for early breast cancer detection.

RELATED WORK

Extensive research has explored the application of various machine learning (ML) techniques for breast cancer detection. One of the most widely used datasets in this domain is the Wisconsin Breast Cancer Diagnostic Dataset (WBCD), made publicly available by the UCI Machine Learning Repository [1]. This dataset has played a pivotal role in training and evaluating ML models for distinguishing malignant and benign tumors.

Early foundational work by Wolberg et al. demonstrated the efficacy of ML techniques in classifying tumors using fine needle aspirate samples, highlighting the potential of supervised learning in medical diagnostics [2]. Their work established a foundation for subsequent advancements in feature extraction and classification methodologies.

Building on these developments, Géron provided a comprehensive guide on implementing machine learning algorithms using modern frameworks, like Scikit-learn**,** Keras, and TensorFlow, which are crucial tools for real-world applications in biomedical domains [3]. These tools played a crucial role in the implementation of various models in our study, including Support Vector Machines (SVM), Random Forest, and K-Nearest Neighbors (KNN).

DATASET AND METHODOLOGY

Dataset Description

The Wisconsin Breast Cancer Dataset (WBCD) is utilized in this study to classify breast tumors as malignant or benign. The dataset comprises 569 instances, each characterized by 30 numerical features extracted from digitized images of fine needle aspirate (FNA) biopsies. These features capture various cellular attributes of breast masses, including radius, texture, perimeter, and smoothness. The target variable, diagnosis, is categorized as Malignant (M) or Benign (B), presenting a binary classification task for machine learning models [4].

Data Preprocessing

To enhance model performance and reliability, the dataset was subjected to the following preprocessing steps [5].

- *Removal of Irrelevant Columns:* Non-informative attributes, such as the ID column, were excluded to eliminate redundancy.
- *Handling Missing Values:* Missing or null entries were addressed through statistical imputation methods, including mean or median substitution.
- *Encoding Categorical Variables:* The target variable is encoded as Malignant = 1 and Benign = 0 for machine learning compatibility.
- *Feature Scaling:* Standardization or normalization techniques are applied to ensure all numerical features contribute equally to the model's learning process.
- *Splitting the Dataset:* The dataset was partitioned into training (80%) and testing (20%) subsets to assess the generalizability of the models.

Machine Learning Models

Machine Learning (ML), a crucial component of Artificial Intelligence (AI), has transformed the field of medical diagnostics by enabling systems to learn from data, detect complex patterns, and make

decisions without the need for explicit programming. ML algorithms are particularly effective in handling large, complex datasets typical in healthcare, where traditional diagnostic methods may be constrained by factors like human bias, cost, or time limitations. In breast cancer detection, ML acts as a powerful tool for differentiating between benign and malignant tumors, assisting healthcare professionals with early diagnosis and treatment planning [6].

This procedure typically entails feeding historical medical data into an algorithm that learns from these examples, allowing it to generalize and make predictions on new, unseen data. In this study, several supervised learning algorithms are employed to classify breast tumors based on features extracted from the Wisconsin Breast Cancer Dataset (WBCD) [7].

Algorithms Used in This Research

This study implements and evaluates the performance of five widely used supervised machine learning algorithms: Logistic Regression, Support Vector Machine, Gaussian Naive Bayes, K-Nearest Neighbors, and Random Forest. A thorough description of each algorithm is presented below [8]:

Logistic Regression (LR)

Logistic Regression is a statistical method used for binary classification tasks, where the dependent variable is categorical (e.g., malignant or benign). It utilizes the logistic (sigmoid) function to estimate the probability that an input belongs to a particular class. Despite being linear in form, Logistic Regression remains popular due to its simplicity, efficiency, and ease of interpretation, particularly when there is a roughly linear relationship between the input variables and the target. It is especially useful in scenarios where the data is well-organized, and the features are not strongly correlated with one another [9].

Support Vector Machine (SVM)

Support Vector Machines are robust classifiers designed to identify the best hyperplane that separates data into two distinct classes while maximizing the margin between them. Support Vector Machine (SVM) can be extended to manage non-linear data by employing kernel methods such as the Radial Basis Function (RBF) kernel. Thanks to its strong generalization capabilities, even with small datasets, and its proficiency in high-dimensional spaces, SVM is frequently applied in medical fields. In breast cancer detection, it performs effectively, particularly when dealing with complex relationships between features.

Gaussian Naive Bayes (GNB)

This algorithm is grounded in Bayes' Theorem and assumes that each feature independently influences the outcome. In the Gaussian variant, continuous variables are modeled as following a normal (Gaussian) distribution. Despite the simplification of assuming feature independence, Gaussian Naive Bayes is known for its high efficiency and strong performance in classification tasks, making it a reliable choice for real-world applications.

K-Nearest Neighbors (KNN)

KNN is an instance-based learning technique that assigns a class to a data point by analyzing the classifications of its closest "k" neighbors. As a non-parametric algorithm, it doesn't make assumptions about the data distribution. Its simplicity and flexibility make it useful, especially when the dataset has complex structures. However, prediction can be slow when working with large datasets due to the need to compute distances for each query [10].

Random Forest (RF)

Random Forest is an ensemble learning method that builds multiple decision trees during training and aggregates their predictions to enhance classification accuracy and reduce overfitting. Each tree is trained on a randomly selected subset of the data and utilizes a random subset of features, promoting diversity among trees and improving generalization. Random Forest is highly robust to noise and

outliers, scales well with large datasets, and delivers excellent performance across a wide range of classification tasks. Its feature importance scores also help identify which attributes most influence the prediction, making it a valuable tool in medical research.

Model Evaluation

- *Accuracy:* Represents how often the model correctly distinguishes between benign and malignant tumors across all predictions.
- *Confusion Matrix:* Provides a comprehensive summary of prediction outcomes by classifying them into true positives, true negatives, false positives, and false negatives (Figure 1).
- *Precision:* Shows the percentage of cases labeled as positive by the model that are positive.
- *Recall:* Measures the model's ability to accurately identify all true positive instances within the dataset.
- *F1 Score:* The harmonic mean of precision and recall, offering a balanced metric for evaluating model performance when both false positives and false negatives are critical.

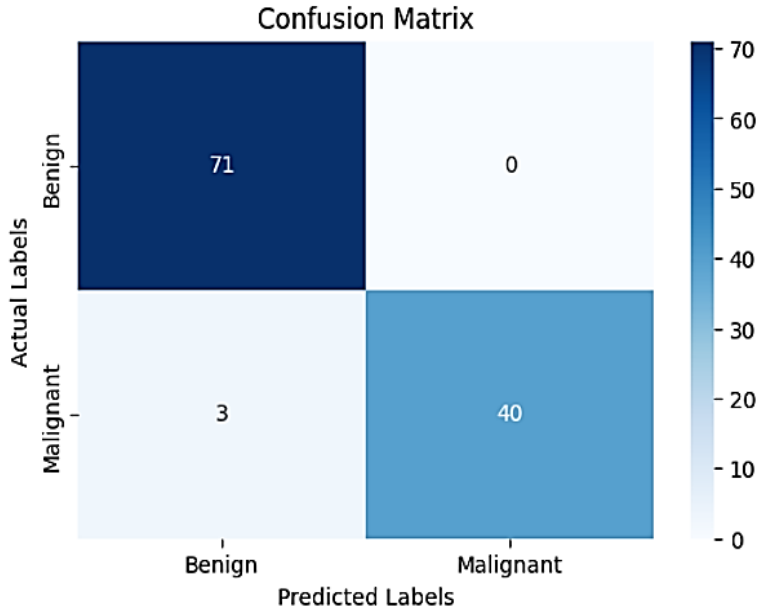

Figure 1. Confusion matrix for breast cancer classification.

RESULTS AND DISCUSSION

The machine learning models were trained and assessed using the pre-processed Wisconsin Breast Cancer Dataset (WBCD). Performance metrics, such as accuracy, precision, recall, and F1-score, were used to compare the effectiveness of each classifier in distinguishing between malignant and benign tumors. The classification accuracy for each model is summarized below:

The performance of five supervised machine learning models – Logistic Regression, Support Vector Machine (SVM), Gaussian Naive Bayes, K-Nearest Neighbors (KNN), and Random Forest – was evaluated using the Wisconsin Breast Cancer Dataset (WBCD). To assess their ability to differentiate between malignant and benign tumors, several performance metrics were employed, including accuracy, precision, recall, and F1-score. The dataset was partitioned into training and testing subsets using a stratified sampling approach to maintain class distribution during model validation. The Gaussian Naive Bayes classifier achieved the highest overall performance, with an accuracy of 97.37%, precision of 97.48%**,** recall of 97.37%**,** and an F1-score of 97.35%**.** This was closely followed by Logistic Regression and Random Forest, both reaching accuracy scores of 96.49%**,** showcasing their effectiveness in classifying benign and malignant tumors.

K-Nearest Neighbors (KNN) demonstrated strong generalization capabilities, attaining an accuracy of 95.61%, while Support Vector Machine (SVM) followed with 94.74%. These results indicate the robustness of these models, especially with proper feature scaling and preprocessing (Table 1).

Table 1. Comparison of evaluation metrics of ML algorithms.

Model	Accuracy	Precision	Recall	F1 Score
Gaussian Naive Bayes	0.9737	0.9748	0.9737	0.9735
Logistic Regression	0.9649	0.9652	0.9649	0.9647
Random Forest	0.9649	0.9652	0.9649	0.9647
K-Nearest Neighbors	0.9561	0.9590	0.9561	0.9555
Support Vector Machine	0.9474	0.9515	0.9474	0.9465

- Gaussian Naive Bayes performs the best overall, with all metrics scoring around 0.9735–0.9748.
- Logistic Regression and Random Forest show nearly identical and strong performance, each with all metrics around 0.9649–0.9652.
- K-Nearest Neighbors (KNN) performs slightly better than SVM, with scores around 0.956–0.9590.
- SVM has the lowest performance, with scores around 0.9474–0.9515, though still relatively high.

Comparison of Different ML algorithms

Figures 2 and 3 present a comparative analysis of machine learning algorithms, illustrating their performance ranking and evaluation metrics for breast cancer classification.

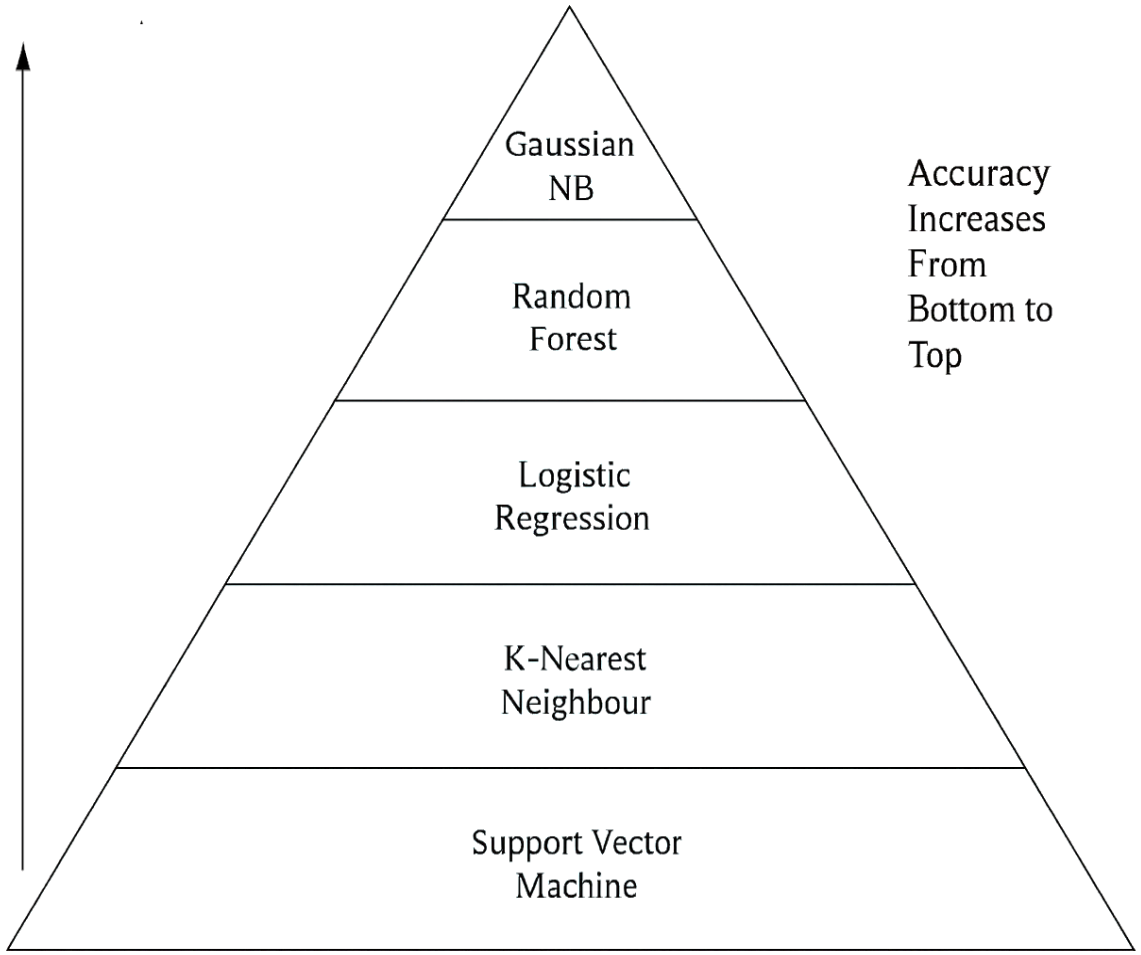

Figure 2. Accuracy ranking of machine learning algorithms for breast cancer classification.

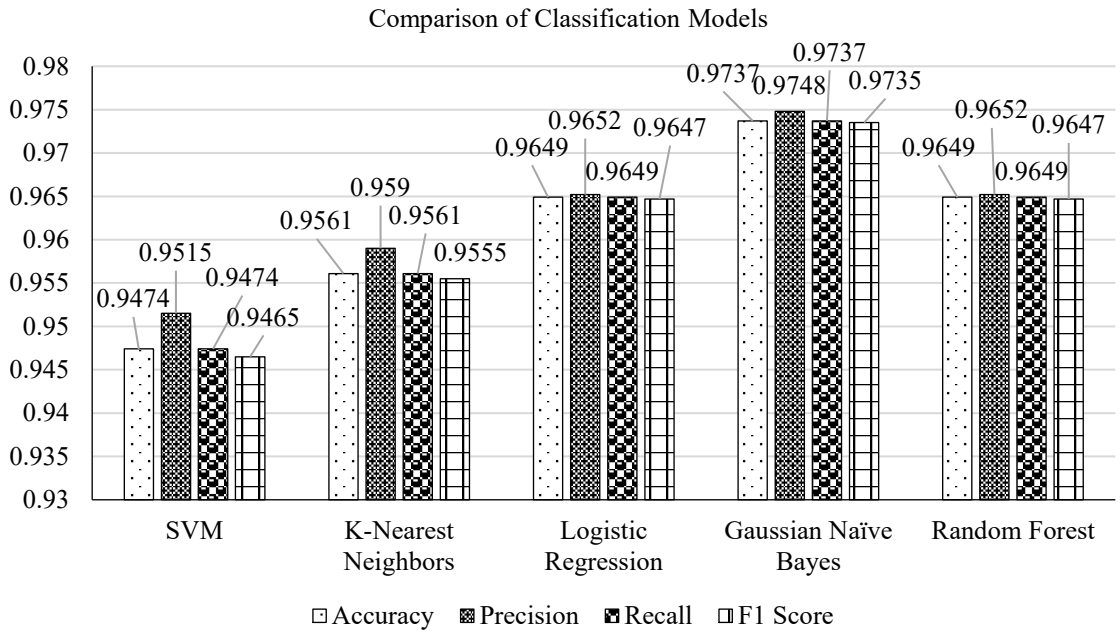

Figure 3. Performance comparison of machine learning algorithms based on accuracy, precision, recall, and F1-score.

CONCLUSION AND FUTURE WORK

- The experimental findings revealed that among the tested models, Gaussian Naive Bayes delivered the most accurate classification, achieving an accuracy of 97.37%. This was closely followed by Logistic Regression and Random Forest, both scoring 96.49%, then K-Nearest Neighbors with 95.61%, and finally Support Vector Machine at 94.74%. These outcomes emphasize the potential of machine learning techniques in effectively identifying breast cancer, offering valuable support to healthcare professionals in diagnostic processes. The outstanding performance of Gaussian Naive Bayes is likely due to its strength in handling high-dimensional data and effectively separating classes, making it particularly suitable for medical data analysis.
- Integrating deep learning architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to enhance classification performance.
- Optimizing hyperparameters of existing machine learning models to further improve predictive accuracy.

In conclusion, machine learning proves to be a powerful and scalable approach for early breast cancer detection, offering substantial benefits in terms of speed, accuracy, and cost-effectiveness.

REFERENCES

1. Dua D, Graff C. UCI machine learning repository: Breast Cancer Wisconsin (Diagnostic) dataset. Irvine (CA): University of California, School of Information and Computer Science; 2019 [cited 2025 Sep 4]. Available from: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).
2. Wolberg WH, Street WN, Mangasarian OL. Using machine learning methods to diagnose breast cancer from fine needle aspirates. Cancer Lett. 1995 Jul;77(2–3):163–71.
3. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed. Sebastopol (CA): O'Reilly Media; 2019.
4. Chaurasia V, Pal S. Data mining techniques for breast cancer detection: A new approach. Int J Innov Res Comput Commun Eng. 2017 Jan;5(1):1206–12.
5. Alyass A, Turcotte M, Meyre D. Challenges and opportunities in big data analysis for personalized medicine. BMC Med Genomics. 2015 Jul;8(1):33.

6. Kapila R, Saleti S. Breast cancer detection using an ensemble-based machine learning approach. Procedia Comput Sci. 2023;218:1483–92.
7. Fatima A, Shabbir A, Janjua JI, Ramay SA, Bhatty RA, Irfan M, et al. Breast cancer detection: A review of machine learning and deep learning approaches. J Comput Biomed Inform. 2024;7(2):45–57.
8. Singh SJ, Agrawal P. A machine learning approach for ternary-class prediction in breast cancer diagnosis. In: Proceedings of the 2024 International Conference on Machine Learning and Bioinformatics. ResearchGate; 2024.
9. Sharma A, Pati N, Gourisaria MK, Pattnayak P. A comparative evaluation of machine learning methods for detecting breast cancer. In: Proceedings of the 2024 International Conference on Computational Intelligence in Biomedical Applications. ResearchGate; 2024.
10. Patra A, Biswas P, Behera SK, Nanthaamornphong A. Advanced AI techniques in image-based breast cancer detection and severity assessment. Biomed Signal Process Control. 2024;92:106078.

