

Performance Evaluation of Machine Learning Algorithms for Breast Cancer Diagnosis using Clinical Data

N. Paranjothi, A. Poongothai

Assistant Professor, Department of Statistics, Annamalai University, Chidambaram.

Research Scholar, Department of Statistics, Annamalai University, Chidambaram

Received: Sep 09, 2026

Accepted: Oct 06, 2026

Published: Oct 13, 2026

Abstract: This study focuses on the application of machine learning techniques for the prediction and classification of breast cancer using clinical data obtained from primary health checkups. The dataset consists of 300 patient records collected from the outpatient department of a private hospital, including socio-demographic details and clinical parameters such as blood pressure, oxygen saturation, and diabetes indicators. The methodology involves data preprocessing, feature selection, and the implementation of four machine learning algorithms, namely Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and XGBoost. These models were trained and tested to evaluate their performance in classifying breast cancer cases. The results were analyzed using performance measures such as sensitivity, specificity, accuracy, True Positive Rate (TPR), False Negative Rate (FNR), and Kappa statistic. Among the models, SVM demonstrated the highest performance with better classification accuracy and reliability, followed by KNN and Naïve Bayes, which showed moderate results. XGBoost exhibited lower performance in comparison to other models. The findings highlight the effectiveness of machine learning techniques in supporting early detection and diagnosis of breast cancer. The study concludes that SVM is the most suitable model for the given dataset and can be effectively used to assist healthcare professionals in decision-making and improving patient outcomes.

Keywords: Machine Learning, Breast Cancer, SVM, Classification, Healthcare Analytics

1.0 INTRODUCTION

Machine learning, an important branch of artificial intelligence, focuses on developing systems that can learn from data and make accurate predictions or decisions. It uses statistical techniques and algorithms to analyze historical or real-time data, identify patterns, and generate meaningful outputs. Over the years, machine learning has gained significant importance due to its ability to handle large and complex datasets efficiently. Its applications have expanded across various fields, particularly in healthcare, where data-driven insights are essential for improving diagnosis, treatment, and patient care.

In India, the growing burden of diseases such as breast cancer highlights the need for effective and timely diagnostic methods. Breast cancer is one of the leading causes of death among women, and early detection remains a major concern due to lack of awareness, limited screening facilities, and delayed diagnosis. In this context, machine learning plays a crucial role by assisting healthcare professionals in analyzing medical data, including imaging and patient records, to identify disease patterns and predict risks. The integration of machine learning techniques into healthcare systems has the potential to enhance the accuracy of breast cancer diagnosis, support clinical decision-making, and improve overall treatment outcomes. This study aims to explore the use of various machine learning approaches and algorithms in the diagnosis and management of breast cancer in India.

2.0 REVIEW OF LITERATURE

The use of machine learning techniques for breast cancer diagnosis has been widely explored with an emphasis on improving classification accuracy and early detection. Lee and Sun [1] demonstrated that supervised learning methods can effectively process clinical datasets to

support diagnosis. Boddu and Jan [2] compared algorithms such as Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree, and Logistic Regression, reporting that SVM achieved higher accuracy in tumor classification. Sharma et al. [3] applied data mining techniques and found that Decision Tree models provided better interpretability, while Reddy et al. [4] showed that KNN performed well with smaller datasets. These studies indicate that traditional machine learning models are effective in handling structured medical data and producing reliable diagnostic outcomes.

With the advancement of computational power, deep learning techniques have been increasingly applied to medical image analysis. Nasser and Yusof [5] used Convolutional Neural Networks (CNN) for breast cancer detection from mammogram images and reported improved accuracy compared to traditional methods. Wang et al. [6] also applied deep learning models and achieved high performance in image-based classification tasks. Zhao et al. [7] highlighted the role of artificial neural networks in identifying complex patterns in large datasets. Similarly, Chen et al. [8] proposed hybrid models combining machine learning and deep learning techniques, resulting in better prediction accuracy. These approaches are particularly useful in extracting features automatically and reducing the need for manual intervention.

Several researchers have focused on prediction and risk assessment using machine learning algorithms. Hussain et al. [9] developed predictive models using Logistic Regression and Random Forest to identify high-risk patients, achieving significant improvements in prediction accuracy. Zuo et al. [10] compared multiple algorithms and concluded that ensemble methods such as Random Forest and Gradient Boosting performed better in recurrence prediction. Khan et al. [11] used machine learning models to predict survival rates, demonstrating the importance of feature selection in improving outcomes. Das et al. [12] applied data-driven approaches in oncology and reported that combining clinical and demographic data enhances prediction performance. These studies highlight the role of machine learning in supporting long-term monitoring and personalized treatment.

Recent research has also emphasized explainability and real-world implementation of machine learning systems. Arravalli et al. [13] introduced explainable AI techniques to make model predictions more transparent and understandable for clinicians. Verma and Patel [14] discussed the integration of machine learning in clinical decision support systems, showing improvements in diagnostic efficiency. Gupta and Singh [15] focused on early detection models and reported reduced error rates in diagnosis. Iqbal et al. [16] highlighted the importance of healthcare analytics in improving patient care. Rao and Nair [17] studied breast cancer screening practices in India and emphasized the need for technology-driven solutions. Banerjee et al. [18] reviewed recent advancements in machine learning and their impact on healthcare systems. Guo et al. [19] demonstrated that machine learning models can handle complex datasets and improve predictive performance, while Smith and Kumar [20] discussed the broader applications of artificial intelligence in healthcare. Overall, these studies confirm that machine learning algorithms, including SVM, CNN, XGBoost and hybrid models, play a crucial role in improving the accuracy, efficiency, and reliability of breast cancer diagnosis and prediction.

3.0 DATABASE

The data used in this study were obtained from 300 patients who attended the Outpatient Department of a private hospital for primary health checkups related to breast cancer screening. The information was collected directly from outpatient case sheets, ensuring that it reflects real

clinical observations. The dataset contains detailed socio-demographic characteristics such as age and background, along with important clinical measurements including blood pressure, oxygen saturation levels, and diabetes-related indicators. These variables were selected to provide a comprehensive view of the patients' general health condition and their possible association with breast cancer risk and progression.

This study makes use of the collected data to examine patient survival outcomes through different survival analysis methods. The aim is to understand patterns related to disease progression and the influence of various health parameters on survival rates. Primary health checkups play an important role in identifying potential health risks at an early stage, especially in the case of breast cancer, where timely detection can significantly improve treatment success. By applying survival analysis techniques to the dataset, the study attempts to support better clinical decision-making and contribute to more effective patient management practices.

4.0 METHODOLOGY

The methodology of this study is based on the application of supervised machine learning algorithms to classify and predict breast cancer outcomes using clinical data. The dataset is first preprocessed to remove inconsistencies, handle missing values, and normalize the features to ensure uniformity. After preprocessing, relevant features such as blood pressure, oxygen levels, and diabetes indicators are selected for model training. The dataset is then divided into training and testing sets to evaluate the performance of the models. Different algorithms are applied to the same dataset to compare their efficiency in classification and prediction. The performance of each model is assessed using accuracy, precision, and recall, which helps in identifying the most suitable algorithm for the given healthcare problem.

4.1 NAÏVE BAYES ALGORITHM

The Naïve Bayes algorithm is a probabilistic classifier based on Bayes' theorem, which assumes that all input features are independent of each other. This assumption simplifies the computation and makes the algorithm efficient for large datasets. The classification process is based on calculating the posterior probability of each class given the input features. The mathematical model is represented using Bayes' theorem as follows:

$$P(C_k / X) = \frac{P(X/C_k)P(C_k)}{P(X)}, P(X/C_k) = \prod_{i=1}^n P(x_i / C_k)$$

Here, the posterior probability is computed using prior probability and likelihood values. During classification, the algorithm evaluates the probability of the input belonging to each class and assigns it to the class with the highest probability. This method is simple, fast, and effective for medical datasets, especially when the number of features is high.

4.2 SUPPORT VECTOR MACHINE ALGORITHM

Support Vector Machine (SVM) is a supervised learning technique used for classification and regression tasks. In this method, the data points are plotted in an n-dimensional space based on their features, and the algorithm identifies an optimal hyperplane that separates the data into different classes. The main objective of SVM is to maximize the margin between the nearest data points of different classes, which are known as support vectors. The decision boundary created by SVM ensures better generalization and classification accuracy. SVM can handle both linear

and non-linear data by using kernel functions. Although it provides high accuracy, it requires careful parameter tuning and is computationally intensive for large datasets. It is widely used in medical diagnosis due to its ability to handle complex classification problems.

$$\min(w, b) \frac{1}{2} \|w\|^2 \text{ subject to } y_i (w \cdot x_i + b) \geq 1$$

4.3 k-NEAREST NEIGHBOR ALGORITHM

The k-Nearest Neighbor (kNN) algorithm is one of the simplest and most intuitive machine learning methods. It works based on the concept of similarity, where a new data point is classified according to the majority class of its nearest neighbors. The algorithm does not build an explicit model during the training phase; instead, it stores all the training data and performs classification during the testing phase. The distance between data points is calculated using measures such as Euclidean distance. kNN is a non-parametric and lazy learning algorithm, making it flexible for different types of data. However, its performance depends on the choice of the value of K and the distance metric used.

$$d(x, x_i) = \sqrt{\sum_{j=1}^n (x_j - x_{ij})^2}$$

It is particularly useful in healthcare applications where pattern similarity plays an important role in diagnosis.

4.4 GRADIENT BOOSTING ALGORITHM

Gradient Boosting is an advanced ensemble learning technique that builds a strong predictive model by combining multiple weak learners, typically decision trees. The algorithm works in a sequential manner, where each new model is trained to correct the errors made by the previous model. This process is based on minimizing a loss function using gradient descent, which improves the accuracy of the model at each step. Unlike other ensemble methods, Gradient Boosting focuses on reducing prediction errors through continuous refinement. It is highly effective in capturing complex relationships between variables and is widely used for both classification and regression problems.

$$F_m(x) + \gamma_m h_m(x)$$

In healthcare analysis, it provides improved predictive performance and helps in identifying important features that influence disease outcomes.

5.0 RESULTS AND DISCUSSION

The performance of the classification models was evaluated using a confusion matrix, which summarizes the relationship between actual and predicted values. The matrix consists of four components: True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). These values help in understanding how accurately the model classifies both positive and negative cases. The structure of the confusion matrix used in this study is presented in Table 1: Confusion Matrix, which forms the basis for calculating further evaluation metrics.

Table 1: Confusion Matrix

Actual / Predicted	0	1
0	TP	FN
1	FP	TN

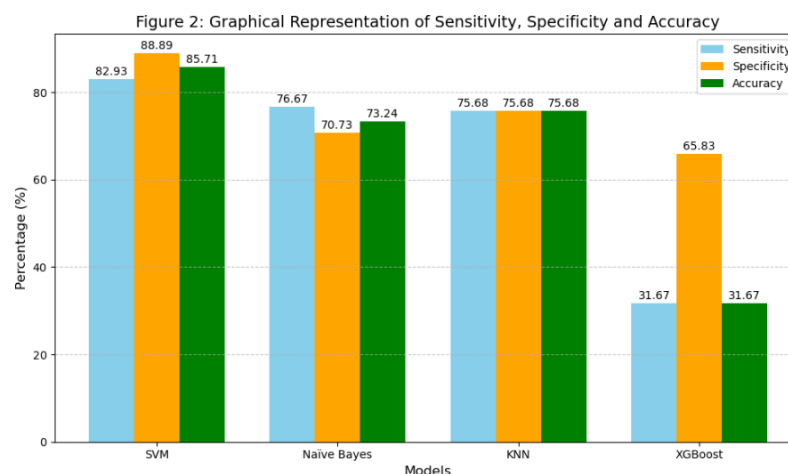
From the confusion matrix, performance measures such as sensitivity, specificity, and accuracy were computed. These measures are important in medical diagnosis, where correct classification of cases is critical. The calculated results for all models are presented in Table 2: Comparison of Sensitivity, Specificity and Accuracy.

Table 2: Comparison of Sensitivity, Specificity and Accuracy

Model	Sensitivity (%)	Specificity (%)	Accuracy (%)
SVM	82.93	88.89	85.71
Naïve Bayes	76.67	70.73	73.24
KNN Algorithm	75.68	75.68	75.68
XGBoost Algorithm	31.67	65.83	31.67

The results in Table 2 show that SVM achieves the highest values across all three parameters, indicating superior classification performance. KNN shows consistent but moderate performance, while Naïve Bayes performs reasonably well but with lower specificity. XGBoost records the lowest values, indicating poor effectiveness for this dataset.

The graphical comparison of these results is shown in Figure 2: Graphical Representation of Sensitivity, Specificity and Accuracy. The figure illustrates that SVM has the highest bars across all metrics, followed by KNN and Naïve Bayes, while XGBoost shows significantly lower values.

Figure 2: Graphical Representation of Sensitivity, Specificity and Accuracy

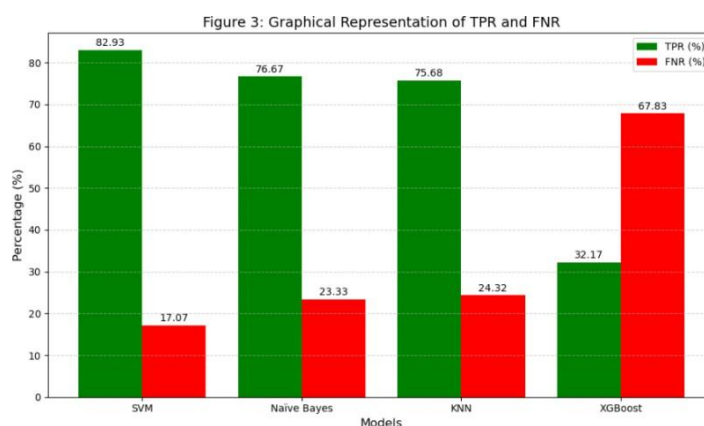
Further evaluation was carried out using True Positive Rate (TPR) and False Negative Rate (FNR), as shown in Table 3: Comparison of True Positive Rate and False Negative Rate.

Table 3: Comparison of True Positive Rate and False Negative Rate

Model	TPR (%)	FNR (%)
SVM	82.93	17.07
Naïve Bayes	76.67	23.33
KNN Algorithm	75.68	24.32
XGBoost Algorithm	32.17	67.83

The results indicate that SVM achieves the highest TPR and lowest FNR, showing its strong ability to correctly identify positive cases. KNN and Naïve Bayes show moderate performance, while XGBoost performs poorly with a high FNR. The graphical representation is shown in Figure 3: Graphical Representation of TPR and FNR.

Figure 3: Graphical Representation of TPR and FNR

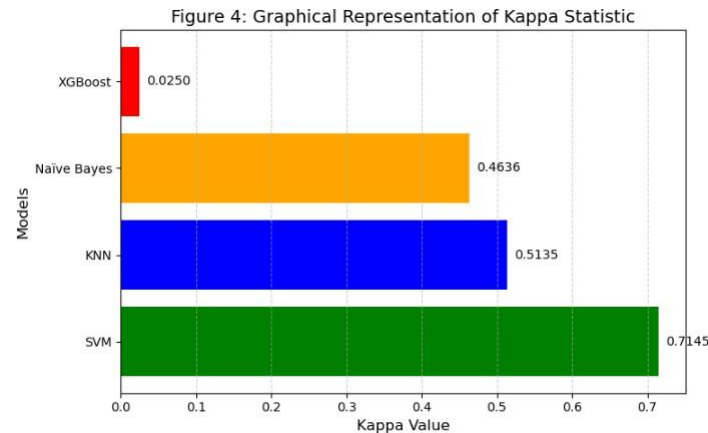


The agreement between predicted and actual classifications was measured using the Kappa statistic, as shown in Table 4: Comparison of Kappa Statistic.

Table 4: Comparison of Kappa Statistic

Algorithm	Kappa Value
SVM	0.7145
Naïve Bayes	0.4636
KNN Algorithm	0.5135
XGBoost Algorithm	0.0250

The results indicate that SVM has the highest agreement level, followed by KNN and Naïve Bayes, while XGBoost shows very poor agreement. The graphical representation is shown in Figure 4: Graphical Representation of Kappa Statistic.



Overall, the numerical tables and graphical figures clearly indicate that SVM provides the best performance in terms of sensitivity, specificity, accuracy, and agreement measures, making it the most reliable model for the given dataset.

6.0 CONCLUSION

The present study examined the performance of different machine learning algorithms for classification and prediction using healthcare data. From the analysis, it is evident that each algorithm shows varying levels of effectiveness depending on the evaluation measures used. Among all the models, the Support Vector Machine (SVM) consistently produced better results in terms of sensitivity, specificity, accuracy, and Kappa statistic. This indicates that SVM is more capable of correctly identifying both positive and negative cases while maintaining a high level of agreement between predicted and actual values. The K-Nearest Neighbor (KNN) and Naïve Bayes algorithms also showed stable and moderate performance, making them suitable for basic classification tasks, although their accuracy was slightly lower compared to SVM. On the other hand, the XGBoost model did not perform well for the given dataset, showing lower sensitivity and accuracy, which affected its overall reliability.

The results clearly highlight the importance of selecting an appropriate machine learning model for healthcare data analysis. Accurate prediction plays a crucial role in early detection and effective treatment, especially in diseases like breast cancer. The study also demonstrates that proper evaluation using multiple performance measures is necessary to understand the strengths and limitations of each model. Overall, the findings suggest that SVM is the most suitable model for this dataset, providing reliable and consistent results that can support clinical decision-making and improve patient outcomes.

SUGGESTIONS

1. Future studies can improve the performance of models by using larger and more diverse datasets, which may help in achieving better generalization and accuracy.
2. Advanced techniques such as hybrid models or deep learning approaches can be explored to enhance prediction accuracy and overcome the limitations observed in some algorithms.

REFERENCES

- [1] M. Lee and Z. Sun, "Machine learning approaches in medical diagnosis," *Journal of Medical Systems*, vol. 46, no. 3, pp. 1–10, 2022.
- [2] A. S. Boddu and A. Jan, "Breast cancer classification using machine learning algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 2, pp. 45–53, 2025.
- [3] P. Sharma, R. Gupta, and S. Verma, "Breast cancer detection using data mining techniques," *International Journal of Data Science and Analytics*, vol. 9, no. 1, pp. 21–30, 2020.
- [4] K. Reddy, M. Rao, and V. Kumar, "Classification of tumors using supervised learning techniques," *Procedia Computer Science*, vol. 165, pp. 567–574, 2019.
- [5] M. Nasser and U. K. Yusof, "Deep learning in breast cancer detection using medical imaging," *Diagnostics*, vol. 13, no. 1, pp. 1–15, 2023.
- [6] L. Wang, H. Zhang, and Y. Li, "Medical image analysis using deep learning methods," *IEEE Access*, vol. 10, pp. 45678–45689, 2022.
- [7] H. Zhao, X. Chen, and Y. Liu, "Neural networks for disease prediction and classification," *Artificial Intelligence in Medicine*, vol. 135, pp. 102–110, 2023.
- [8] D. Chen, J. Lin, and S. Wu, "Hybrid machine learning models for cancer diagnosis," *Journal of Biomedical Informatics*, vol. 103, pp. 103–112, 2020.
- [9] S. Hussain, A. Malik, and F. Khan, "Predictive modeling for breast cancer risk assessment," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, pp. 1–12, 2024.
- [10] D. Zuo, Y. Wang, and L. Zhang, "Comparative analysis of machine learning models for breast cancer recurrence prediction," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, pp. 1–14, 2023.
- [11] M. Khan, R. Iqbal, and S. Ahmad, "AI-based prediction of cancer survival rates," *Computers in Biology and Medicine*, vol. 140, pp. 105–114, 2021.
- [12] N. Das, P. Roy, and S. Banerjee, "Data-driven approaches in oncology using machine learning," *Journal of Healthcare Engineering*, vol. 2023, pp. 1–12, 2023.
- [13] T. Arravalli, S. Prakash, and R. Devi, "Explainable artificial intelligence in breast cancer diagnosis," *IEEE Access*, vol. 13, pp. 23456–23467, 2025.
- [14] A. Verma and S. Patel, "Machine learning in clinical decision support systems," *Journal of Medical Engineering & Technology*, vol. 46, no. 5, pp. 321–330, 2022.

- [15] S. Gupta and M. Singh, “Early detection of breast cancer using machine learning techniques,” *International Journal of Medical Informatics*, vol. 150, pp. 104–112, 2021.
- [16] R. Iqbal, M. Khan, and A. Malik, “Healthcare analytics using artificial intelligence,” *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 345–356, 2021.
- [17] V. Rao and P. Nair, “Breast cancer screening techniques and challenges in India,” *Asian Pacific Journal of Cancer Prevention*, vol. 23, no. 4, pp. 1201–1208, 2022.
- [18] S. Banerjee, N. Das, and P. Roy, “Advances in machine learning for healthcare applications,” *IEEE Transactions on Emerging Topics in Computing*, vol. 12, no. 2, pp. 567–578, 2024.
- [19] Y. Guo, X. Li, and H. Sun, “Machine learning for breast cancer diagnosis and prediction,” *Computer Methods and Programs in Biomedicine*, vol. 240, pp. 107–118, 2024.
- [20] J. Smith and R. Kumar, “Applications of artificial intelligence in healthcare: A review,” *Journal of Healthcare Informatics Research*, vol. 5, no. 2, pp. 89–101, 2021.