Available online at http://www.iorajournal.org/index.php/ijgor/index

International Journal of Global Operations Research

e-ISSN: 2722-1016
p-ISSN: 2723-1739

Vol. 6, No. 4, pp. 235-240, 2025

Comparison of Machine Learning Models for Breast Cancer Diagnosis Classification

Riza Ibrahim[1*], Siti Hadiaty Yuningsih[2], Muhammad Iqbal Al-Banna Ismail[3]

[1]*Research Collaboration Community, Bandung, Indonesia*
[2]*Department of Manufacture Engineering, Politeknik Manufaktur Bandung, Bandung, Indonesia*
[3]*School of Mathematical Sciences, Sunway University, No.5, Jalan Universiti, Bandar Sunway, 47500 Subang Jaya, Selangor, Malaysia.*

**Corresponding author email: riza240399@gmail.com*

Abstract

Breast cancer remains one of the most pressing global public health challenges, with approximately 2.3 million women diagnosed worldwide in 2022 and around 670,000 deaths attributed to the disease. Despite the widespread application of machine learning algorithms for breast cancer classification, findings across studies remain highly varied, and there is still no consistent conclusion regarding which algorithm is most superior for breast cancer diagnosis. This study aims to analyze and compare the performance of four machine learning algorithms Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) in predicting breast cancer. The dataset used was the Breast Cancer Wisconsin (Diagnostic) Data Set obtained from Kaggle, containing morphological characteristics of tumor cells. Data preprocessing involved cleaning, label encoding, feature normalization using StandardScaler, and an 80:20 train-test split. Model performance was evaluated using confusion matrix, precision, recall, F1-score, accuracy, and ROC-AUC. The results showed that all four models achieved excellent performance with overall accuracy ranging from 95.61% to 97.37%. SVM emerged as the most accurate model (97.37%) with perfect recall (1.00) for the Benign class. Logistic Regression demonstrated the highest ROC-AUC value (0.9960), indicating excellent discriminative ability. Random Forest and KNN showed slightly lower performance, particularly in detecting Malignant cases with recall of 0.90. These findings confirm that machine learning can serve as an effective tool to support breast cancer diagnosis, with algorithm selection depending on data characteristics and clinical priorities.

Keywords: Breast cancer, classification, machine learning.

1. Introduction

Breast cancer is one of the most urgent public health problems globally, particularly concerning women's health. This disease not only brings a severe impact to individuals diagnosed through pain, emotional burden, and treatment costs but also imposes substantial social and economic burdens on families and healthcare systems (Fortin et al., 2021). Public knowledge and awareness regarding the dangers of breast cancer are often inadequate, causing many new cases to be detected only when the disease has already progressed to an advanced stage (Aga et al., 2024).

According to WHO, in 2022 an estimated 2.3 million women worldwide were diagnosed with breast cancer, and approximately 670,000 deaths occurred due to the disease. Breast cancer is also recorded as the most common type of cancer among women in 157 out of 185 countries, indicating that it represents a widespread global health burden. These data illustrate the extremely high incidence and mortality rates associated with breast cancer worldwide an alarming reality that reinforces the urgency of research, early detection, and effective interventions to mitigate its impact.

The application of machine learning to breast cancer data is important not only as an alternative diagnostic method but also as a way to evaluate how well predictive models can recognize biological patterns associated with cell malignancy. Clinical datasets containing numerical features such as cell size, texture, perimeter, symmetry, and other morphological characteristics enable models to learn subtle differences between benign and malignant tumors. Evaluating various machine learning algorithms provides an overview of each model's ability to capture these complex patterns in terms of accuracy, sensitivity to cancer classes, and prediction consistency (Ansari, 2024; Wang et al., 2024).

Previous studies have indicated that numerous machine learning algorithms have been used for breast cancer classification, and each model has been reported to show different performance depending on the characteristics of the

analyzed data. Among the algorithms most frequently applied in breast cancer diagnostic studies are Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN), as they have demonstrated competitive predictive results on clinical datasets.

Although many studies have employed machine learning for breast cancer classification, findings across studies remain highly varied. Amrane et al. (2018) compared Naive Bayes and K-Nearest Neighbors and reported that KNN achieved the highest accuracy of 97.51%, while Naive Bayes reached 96.19%. Meanwhile, the study by Muntiari and Hanif (2022), which evaluated seven different algorithms, found that Logistic Regression, Decision Tree, Naive Bayes, and K-Nearest Neighbors yielded the same high accuracy of 95.00%. The varying results of these studies indicate that there is still no consistent conclusion regarding which algorithm is the most superior for breast cancer classification.

Based on the variations found in previous research, it is evident that the performance of breast cancer classification algorithms has not shown consistency across studies, making it difficult to determine which model is most optimal for clinically based data diagnosis. This inconsistency highlights a research gap, particularly the need for a comparative evaluation testing multiple algorithms on the same dataset to obtain objective results. Therefore, this study was conducted to analyze and compare the performance of four machine learning algorithms Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) in predicting breast cancer.

2. Methodology

2.1. Data Collection

The data used in this study were sourced from the public Breast Cancer Prediction using Machine Learning dataset obtained from the Kaggle platform. This dataset is a digitized version of the Breast Cancer Wisconsin (Diagnostic) Data Set, a benchmark dataset widely used in breast cancer detection research. The data contains clinical information in the form of measurements of tumor cells or nuclei taken during the Fine Needle Aspirate (FNA) procedure, in which cell nuclei are analyzed microscopically and several morphological characteristics (radius, texture, perimeter, area, smoothness, compactness, symmetry, and others) are calculated.

2.2. Preprocessing

This process begins with data cleaning by removing the id and "Unnamed: 32" columns as they contain no diagnostic information. Then, the target label "diagnosis" from the "M" (malignant) and "B" (benign) categories was converted to numeric values 1 and 0 for processing by the classification algorithm. Missing values were checked. The dataset consists of numeric features representing the morphological characteristics of tumor cells, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, each in the form of mean, standard error, and worst-case scenario. All of these features were used as predictor variables (X), while the diagnosis column served as the target (y). Feature normalization was performed using StandardScaler to equalize the range of values between variables, as some algorithms are sensitive to differences in scale. The data was divided into a training set and a testing set with an 80:20 ratio using a stratified train-test split.

2.3. Model Development

This study uses four classification algorithms to develop a breast cancer prediction model, namely Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN).

1. Logistic Regression
   Logistic Regression works by modeling the relationship between predictor features and the probability that a sample belongs to a particular class (Faouzi and Colliot, 2023; Choi et al., 2020). The Logistic Regression formula is as follows:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-z}} \tag{1}$$

   Ware

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{2}$$

2. Support Vector Machine (SVM)
   Support Vector Machine works by finding the optimal separating boundary (hyperplane) that maximizes the margin between two classes in the feature space. The larger the margin formed, the better the model's ability to distinguish between benign and malignant classes (Guido et al., 2024; Sembiring et al., 2024). The general decision function of SVM is expressed as:

$$f(x) = sign\left(\sum_{i=1}^{m} \alpha_i \; y_i \, K(x_i, x) + b\right) \tag{3}$$

3. Random Forest
   Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines their prediction results to obtain the final decision. The prediction formula of Random Forest is expressed as:

$$\hat{y} = mode\{h_1(x),\, h_2(x),\, \ldots,\, h_T(x)\} \tag{4}$$

4. K-Nearest Neighbors (KNN)
   KNN performs classification based on the similarity of a sample to its nearest neighbors in the feature space. The prediction formula of KNN is expressed as:

$$\hat{y} = \arg\max_{c} \sum_{i \in N_k(x)} 1(y_i = c) \tag{5}$$

2.4. Model Performance Evaluation

The performance evaluation of the four models was carried out using the testing set, emphasizing each algorithm's ability to distinguish between benign and malignant classes. The confusion matrix was employed as the primary metric to identify patterns of correct and incorrect predictions, including false positives and false negatives, which are critical considerations in medical diagnosis. In addition, precision, recall, and F1-score were calculated to assess the accuracy and sensitivity of malignant cancer detection, while accuracy reflected the overall proportion of correct predictions (Kenny et al., 2024).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{8}$$

$$F1 - Score = \frac{TP}{TP + FN} \times 100\% \tag{9}$$

The discriminatory capability of each model was further evaluated through the ROC curve and the Area Under the Curve (ROC-AUC) value.

3. Results and Discussion

3.1. Model Performance Evaluation

The results of the model performance evaluation are presented in Table 1.

Table 1: Performance evaluation of machine learning models

Model	Class	Precision	Recall	F1-Score	Support	Accuracy	ROC-AUC
Logistic Regression	Benign	0.96	0.99	0.97	72	0.9649	0.9960
	Malignant	0.97	0.93	0.95	42	0.9649	0.9960
SVM	Benign	0.96	1.00	0.98	72	0.9737	0.9947
	Malignant	1.00	0.93	0.96	42	0.9737	0.9947
Random Forest	Benign	0.95	1.00	0.97	72	0.9649	0.9942
	Malignant	1.00	0.90	0.95	42	0.9649	0.9942
KNN	Benign	0.95	0.99	0.97	72	0.9561	0.9823
	Malignant	0.97	0.90	0.94	42	0.9561	0.9823

The evaluation results show that the four models Logistic Regression, SVM, Random Forest, and KNN generally achieve very high performance in classifying breast cancer diagnoses. Logistic Regression demonstrates a strong balance between precision and recall for both classes, with the highest ROC-AUC value (0.9960), indicating that this linear model can separate the two classes almost perfectly. SVM emerges as the most accurate model (0.9737), achieving perfect recall (1.00) for the Benign class, which reflects its ability to identify all non-cancer samples without error.

Random Forest exhibits a performance pattern similar to SVM, although it performs slightly lower in detecting Malignant cases (recall 0.90), which may occur because the model tends to be more conservative when predicting the positive class. Meanwhile, KNN shows relatively lower performance compared to the other models, particularly in the recall of the Malignant class (0.90), indicating its limitations in detecting cancer cases at certain levels of feature complexity. The occurrence of 100% precision or recall in SVM and Random Forest is due to the characteristics of the dataset, which has highly separable features between Benign and Malignant classes. The features in this dataset exhibit clear distribution patterns, allowing some models to classify all samples in one class without misclassification.

3.2. Confusion Matrix

The results of the confusion matrices are presented in Figure 2.

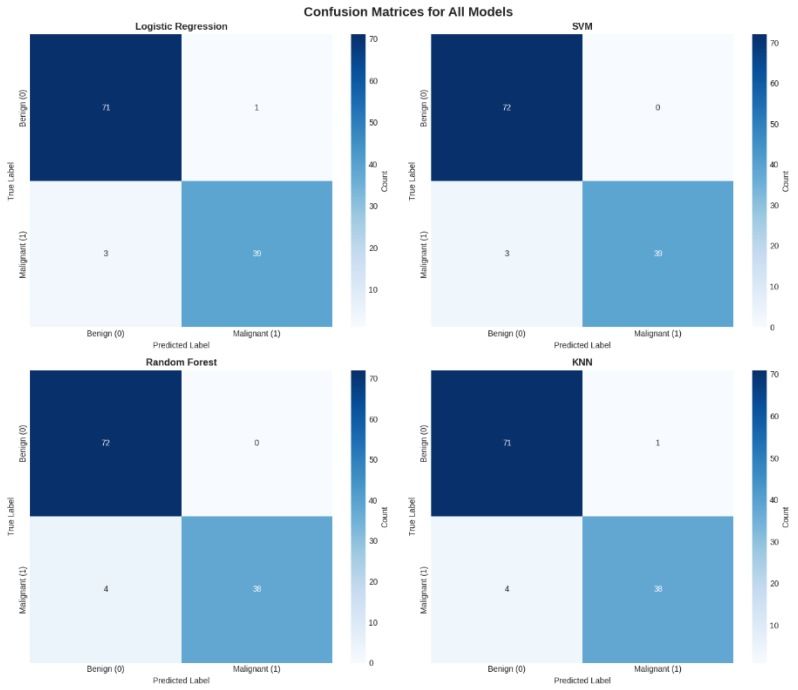

Figure 2: Confusion matrix

Based on the confusion matrices shown in Figure 2, all models demonstrate strong classification capabilities, particularly in identifying the Benign class. Logistic Regression produces only one false positive (71 correct, 1 incorrect), while SVM and Random Forest achieve perfect performance for the Benign class (72 correct, 0 incorrect). KNN shows a similar pattern to Logistic Regression with 71 correct predictions and 1 misclassification. For the Malignant class, Logistic Regression and SVM correctly predict 39 samples with only 3 misclassifications, whereas Random Forest and KNN each correctly identify 38 samples and misclassify 4. These relatively small error patterns indicate that all models are highly effective at separating the two classes, with performance differences primarily reflected in their sensitivity to the Malignant class.

This is consistent with the model evaluation results, which show that Logistic Regression achieves the highest ROC-AUC (0.9960) and a strong balance between precision and recall an outcome reflected in the low number of false negatives in its confusion matrix. SVM emerges as the most accurate model (0.9737), supported by its perfect recall for the Benign class, as indicated by the absence of any misclassification in that category. Random Forest also aligns with the earlier evaluation results, where its slightly lower recall for the Malignant class is confirmed by the presence of four false negatives. KNN follows a similar pattern, demonstrating slightly reduced performance in detecting Malignant samples.

3.3. ROC Curve Analysis

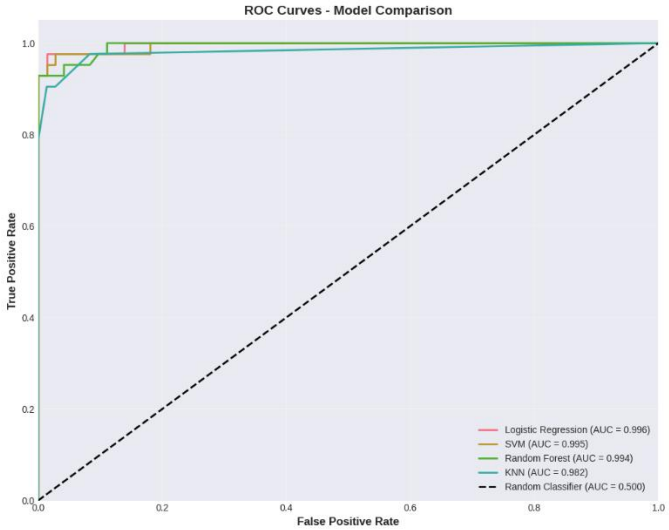

Figure 3: ROC curves

The ROC curves shown in Figure 3 illustrate that all four models demonstrate exceptionally strong discriminative ability, as indicated by their trajectories that rise steeply toward the top-left corner of the plot. Logistic Regression, SVM, and Random Forest show curves that almost perfectly hug the upper boundary, reflecting their extremely high AUC values (0.996, 0.995, and 0.994, respectively). This shape indicates that these models maintain a very high true positive rate even as the false positive rate remains near zero, meaning they can correctly distinguish between benign and malignant cases across various classification thresholds. KNN, although slightly lower with an AUC of 0.982, still forms a strongly convex curve, demonstrating solid performance but with a slightly earlier increase in false positives due to its sensitivity to local feature variations. The smooth and sharply rising curves across all models correspond to the clear separability of the dataset's features, which allows the classifiers to achieve high sensitivity without significantly increasing false alarms. The contrast with the diagonal random classifier line further highlights how far above chance level all models perform.

4. Conclussion

Based on the research conducted, it can be concluded that the four machine learning algorithms tested Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) demonstrated excellent performance in classifying breast cancer diagnoses with overall accuracy ranging from 95.61% to 97.37%. SVM emerged as the most accurate model with an accuracy rate of 97.37% and achieved perfect recall (1.00) for the Benign class, demonstrating its ability to identify all non-cancer samples without error. Logistic Regression showed an excellent balance between precision and recall for both classes with the highest ROC-AUC value of 0.9960, indicating this linear model's ability to separate the two classes almost perfectly. Random Forest displayed a performance pattern similar to SVM, although slightly lower in detecting Malignant cases with a recall of 0.90, which may occur because the model tends to be more conservative in predicting the positive class. Meanwhile, KNN showed relatively lower performance compared to the other models, particularly in the recall of the Malignant class (0.90), indicating its limitations in detecting cancer cases at certain levels of feature complexity.

References

Aga, S. S., Yasmeen, N., Al-Mansour, M., Khan, M. A., Nissar, S., Khawaji, B., ... & Abushouk, A. (2024). Knowledge, awareness and attitude towards breast cancer: Risk factors, signs and screening among Health and Allied students: A prospective study. *Journal of Family Medicine and Primary Care*, *13*(5), 1804-1824.

Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In *2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT)* (pp. 1-4). IEEE.

Ansari, G. A., Bhat, S. S., Ansari, M. D., Ahmad, S., & Abdeljab, H. A. M. (2024). Prediction and diagnosis of breast cancer using machine learning techniques. *Data Metadata*, *3*, 346.

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational vision science & technology*, *9*(2), 14-14.

Faouzi, J., & Colliot, O. (2023). Classic machine learning methods. *Machine learning for brain disorders*, 25-75.

Fortin, J., Leblanc, M., Elgbeili, G., Cordova, M. J., Marin, M. F., & Brunet, A. (2021). The mental health impacts of receiving a breast cancer diagnosis: A meta-analysis. *British Journal of Cancer*, *125*(11), 1582-1592.

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An overview on the advancements of support vector machine models in healthcare applications: a review. *Information*, *15*(4), 235.

Kenny, K., Arisandi, D., & Sutrisno, T. (2024). Evaluasi penilaian kinerja karyawan dengan metode naïve bayes. *Computatio: Journal of Computer Science and Information Systems*, *8*(1), 110-118.

Muntiari, N. R., & Hanif, K. H. (2022). Klasifikasi penyakit kanker payudara menggunakan perbandingan algoritma machine learning. *Jurnal Ilmu Komputer dan Teknologi*, *3*(1), 1-6.

Sembiring, M. A., Saputra, H., Yusda, R. A., Sutarman, S., & Nababan, E. B. (2024). Performance of Robust Support Vector Machine Classification Model on Balanced, Imbalanced and Outliers Datasets. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, *10*(1), 208-215.

Wang, L., Wang, Y., Li, Y., Zhou, L., Liu, S., Cao, Y., ... & Zhu, T. (2024). RETRACTED ARTICLE: A prospective diagnostic model for breast cancer utilizing machine learning to examine the molecular immune infiltrate in HSPB6. *Journal of Cancer Research and Clinical Oncology*, *150*(10), 475.

World Health Organization. (2023). Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer