Acta Scientific MEDICAL SCIENCES (ISSN: 2582-0931)

Volume 10 Issue 5 May 2026

Research Article

Supervised Machine Learning Models for Wisconsin Breast Cancer Diagnostic Dataset Study

Hamza Khan, Kiran Vokkarne and Raji Sundararajan*

School of Engineering Technology, Purdue University, West Lafayette, IN-47907, USA

***Corresponding Author:** Raji Sundararajan, School of Engineering Technology, Purdue University, West Lafayette, IN-47907, USA.

DOI: 10.31080/ASMS.2026.10.2236

Received: March 16, 2026

Published: May 06, 2026

Abstract

With a woman dying every 50 seconds, breast cancer is still a leading cancer of women worldwide, despite all the advanced, millions of dollars research. With early and more accurate diagnosis using efficient supervised machine learning models, it is possible to improve the prognosis of breast cancer patients. Towards this, in this study, six supervised machine learning (ML) models-Logistic Regression (LR), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) were studied using the Wisconsin Breast Cancer Diagnostic dataset. The dataset contains records of 569 patients, 357 benign and 212 malignant, each with thirty quantitative features extracted from digitized fine-needle aspiration cytology images. Models were trained on 80% of the data and evaluated on the remaining 20%. Among all classifiers, the SVM with the radial basis function kernel showed the highest accuracy of 98.25%. This could be because this method finds the clearest possible separation between malignant and benign tumors and can curve that separation when the data is not linearly separable. These findings highlight the strong potential of ML-based systems as diagnostic decision-support tools for better breast cancer detection.

Keywords: Breast Cancer; Machine Learning; Wisconsin Dataset; SVM

Introduction

Breast cancer is one of the most common cancers among women across the globe, with higher mortalities. In 2022, worldwide, there were 2.3 million new cases, with 685,000 mortalities [1], making breast cancer as one of the major causes of cancer-related deaths among women across the globe. In the USA, for 2026, it is estimated that there will be 321,910 new cases with 42,140 deaths. These highlight the importance of early diagnosis and detection, as early diagnosis is strongly related to better treatment outcomes, lower mortality rates, and less invasive treatment modalities, with five-year survival rates exceeding 90 percent in the case of early-stage breast cancer, as opposed to lower survival rates in the case of late-stage cancer [1]. Despite the advantages associated with the diagnosis of breast cancer, the traditional process is often dependent on the interpretation of cytology and imaging data, thus being time-consuming, subjective, and dependent on the expertise of the medical practitioner. Towards this, machine learning (ML) with all its various types of models and algorithms and classifiers, is a boon to early detection.

Machine learning has been emerging as a promising tool in medical diagnostics owing to its capabilities in handling high-dimensional data sets and identifying complex patterns, which might not be easily identified through manual analysis [2]. Machine learning-based models have been widely used in the diagnosis of breast cancer using cytology and imaging data, with previous

studies using supervised models, like Logistic Regression, Naive Bayes, Decision Trees, Random Forests, K Nearest Neighbors, and SVM on the Breast Cancer Wisconsin Diagnostic Data Set and similar data sets and other diseases, such as Alzheimer [3-10]. For example, early detection of Alzhemier's disease with blood plasma proteins using support vector machines was reported by Eke., *et al*. [9]. A machine learning approach for the differential diagnosis of Alzhemier and vascular dementia fed by MRI selected features was presented in [10]. Wisconsin breast cancer data was studied by Sahu., *et al*. [7] and Agarap [6]. They used various supervised models, such as LR, RF, SVM, and feature extraction methods for enhanced prediction of breast cancer. These studies showed high classification accuracy in the diagnosis of breast cancer using cytology data, thus highlighting the importance of shape-based features like concavity radius, perimeter, and area in the diagnosis of malignant and benign tumors [3].

Recent studies applying machine learning to the Wisconsin Breast Cancer Diagnostic dataset have demonstrated strong classification performance, often through feature selection strategies or individual model optimization [6]. While these approaches confirm the effectiveness of classical classifiers, many studies evaluate models in isolation or primarily report overall accuracy without standardized, clinically focused evaluation frameworks [11]. In contrast, in this study, using the same Wisconsin Breast Cancer Diagnostic dataset, we conducted a controlled, model-to-model comparison of six supervised machine learning algorithms under identical preprocessing conditions determine the best model that can be used for the detection of breast cancer and explicitly emphasized malignant-class recall and F1-score, addressing the clinical risk of false negatives. The six supervised machine learning models include logistic regression (LR), naive Bayes (NB), decision tree (DT), random forest (RF), K-nearest neighbors (KNN), and support vector machines (SVM).

Methodology

Dataset

The Wisconsin Breast Cancer Diagnostic dataset provides thirty quantitative features that describe the morphology of cell nuclei extracted from fine needle aspiration images [3]. It contains 569 patient samples, including 212 (37.26%) malignant and 357 (62.74%) benign cases, offering a well-established benchmark for supervised classification tasks. The distribution of malignant and benign cases in the dataset is shown in Figure 1. These features capture characteristics, such as radius, texture, symmetry, smoothness, and concavity, which have been shown to correlate strongly with malignant cellular behavior [3]. Because these measurements are numerical and multidimensional, they are particularly well suited for supervised machine learning algorithms, which can identify subtle patterns and complex decision boundaries that may not be visually apparent through manual inspection. Figure 2 shows the physical characteristics of benign and malignant cells/tumors, illustrating the variations in the size (radius) [8], shape and form of the cells and the tumors, which were used in this dataset.

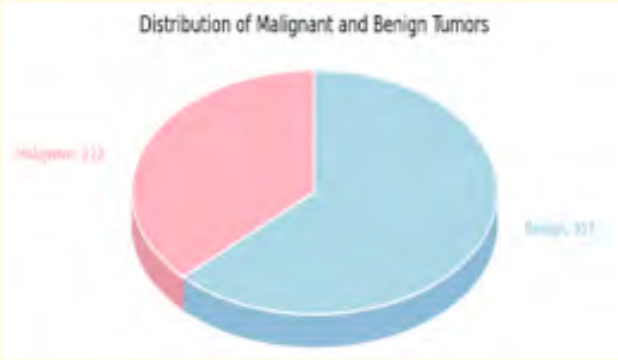

Figure 1: Distribution of benign and malignant cases.

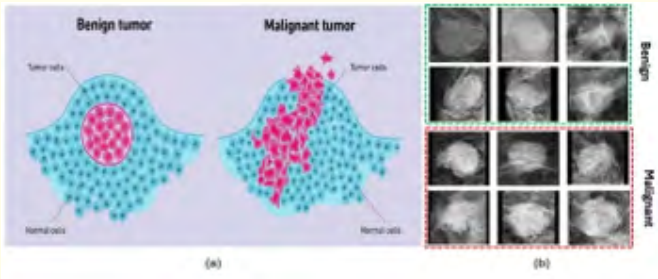

Figure 2: Physical Characteristics of Benign and malignant cells/tumors [8].

The 30 variables for each case that indicate quantitative measurements of tumor cell nuclei extracted using digital images of fine needle aspirates (FNAs). These variables are categorized into three groups: 10 mean variables, 10 standard error variables,

and 10 worst variables. Table 1 provides the listing of variables employed in this analysis, and Table 2 summarizes the feature groupings applied in the dataset.

Serial #	Feature	Description
1	Radius	Mean distance from the center to the boundary of the cell nucleus
2	Texture	Variation in grayscale pixel intensity within the nucleus
3	Perimeter	Total length of the nucleus boundary
4	Area	Total area enclosed by the nucleus boundary
5	Smoothness	Local variation in nucleus boundary radius
6	Compactness	Measure of how compact the nucleus shape is
7	Concavity	Degree of inward curvature of the nucleus boundary
8	Concave Points	Number of concave portions of the nucleus boundary
9	Symmetry	Symmetry of the nucleus shape
10	Fractal Dimension	Complexity of the nucleus boundary shape

Table 1: Tumor features given in the dataset.

Category	Features
Mean	radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave_points_mean, symmetry_mean, fractal_dimension_mean
Standard Error	radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave_points_se, symmetry_se, fractal_dimension_se
Worst	radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave_points_worst, symmetry_worst, fractal_dimension_worst

Table 2: Feature groups.

Machine learning environment

All preprocessing, training, and testing procedures were conducted using the Python programming language within the Google Colab environment [12]. The Wisconsin Breast Cancer Diagnostic dataset was uploaded directly to Colab for analysis, ensuring consistent execution and reproducibility of experiments. Standard machine learning libraries, including NumPy, Pandas, and scikit-learn, were used for data preprocessing, model implementation, and performance evaluation. These libraries provide efficient tools for numerical computation, dataset manipulation, and supervised learning, allowing all models to be trained and evaluated under identical computational conditions. This environment ensured reproducible analysis, consistent implementation, and fair comparison across all evaluated classifiers.

Dataset preprocessing

Data were standardized using z score normalization to improve performance, especially for distance based and margin-based models, such as SVM and KNN [13]. Standardization ensures that features with larger numerical ranges do not dominate features with smaller ranges and allows all variables to contribute equally during model training.

Z score normalization transforms each feature according to:

$$z = \frac{x - \mu}{\sigma}$$

where, represents the original feature value, represents the mean of the feature, and represents the standard deviation. This transformation ensures that each feature has a mean of 0 and a standard deviation of 1.

After normalization, the dataset was divided into 80% training data (455 samples) and 20% testing data (114 samples). This split is commonly used in medical machine learning because it provides a strong balance between model learning capacity and reliable evaluation [14].

Let the dataset be represented as:

$$D = D_{train} \cup D_{test}$$

Where,

$$| D_{train} | = 455$$
$$| D_{test} | = 114$$

All preprocessing steps were performed before fitting the models to ensure consistent feature scaling and fair comparison across classifiers. The overall preprocessing and training workflow is shown in Figure 3.

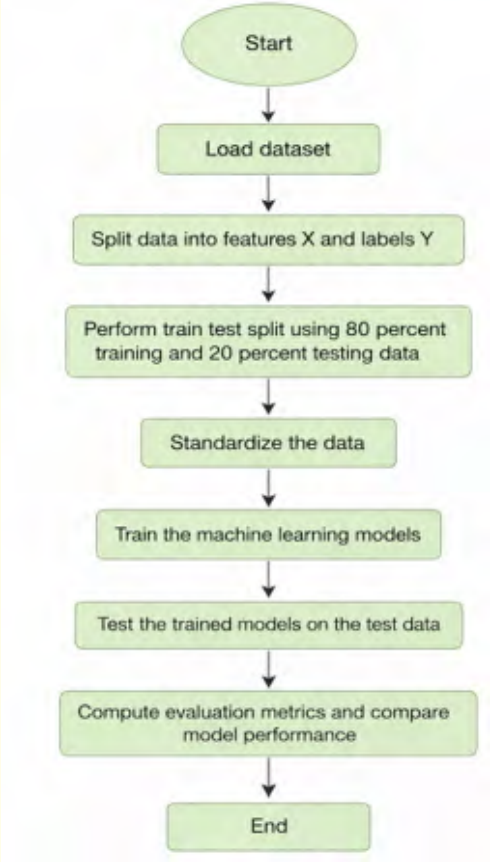

Figure 3: Study workflow.

Machine learning algorithms

Six supervised machine learning classifiers-Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, and K Nearest Neighbors were used to classify tumors as malignant or benign using the Breast Cancer Wisconsin Diagnostic dataset. The mathematical formulations of these classical machine learning algorithms follow standard definitions as described in [15].

Mathematically, the dataset can be represented as,

$$X = \{x_1, x_2, \ldots, x_N\}$$

Where, each sample,

$$x_i \in \mathbb{R}^{30}$$

Represents a 30-dimensional feature vector containing the quantitative tumor measurements described above.

Each sample has a corresponding class label,

$$y_i \in \{0,1\}$$

Where, $y_i = 0$, represents benign tumors and $y_i = 1$, malignant tumors.

Thus, the complete dataset can be represented as,

$$D = \{(x_i, y_i)\}_{i=1}^{N},$$

Where, $N = 569$ represents the total number of samples.

Logistic regression

Logistic Regression (LR) finds the best dividing boundary between malignant and benign tumors. It predicts the probability of a tumor being malignant using the sigmoid function:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}},$$

Where, $P(y = 1 \mid x)$is the probability that the tumor is malignant, is the weight vector, is the feature vecto, is the bias term, and is the exponential constant. This function maps any input value to a probability between 0 and 1, and the class is determined based on a threshold.

Naive bayes

Naive Bayes (NB) uses probability based on tumor measurements to determine whether it is malignant or benign. It applies Bayes' theorem:

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)},$$

Where, $P(y \mid x)$ is the probability of class given the features, $P(x \mid y)$ is the likelihood of the features given the class, $P(y)$ is the prior probability of the class, and $P(x)$ is the total probability of the features. This model predicts the class with the highest probability.

Decision tree

Decision Tree (DT) makes predictions using a series of yes/no questions about tumor features. It splits the dataset using entropy:

$$H(S) = -\sum p\_i \log_2(p\_i)$$

Where, *H (S)* is the entropy and is the probability of each class. Lower entropy results in better separation between classes.

Random forest

Random Forest (RF) involves many decision trees and allows them to vote, making the prediction more stable and precise. The final prediction is determined by majority voting:

$$\hat{y} = mode\big(T_1(x), T_2(x), \ldots, T_n(x)\big),$$

Where, $T_i(x)$ is the prediction from each tree. The final class is the most common prediction. This reduces overfitting and improves accuracy.

Support vector machine

Support Vector Machine (SVM) finds the strongest and cleanest separating line between malignant and benign tumors using margin maximization. The decision boundary is defined as:

$$w^T x + b = 0,$$

Where, w is the weight vector, is the feature vector, and is the bias term.

Since biomedical data are not always linearly separable, the nonlinearity was introduced by the radial basis function (RBF) kernel. The RBF kernel translates the input data in a higher dimensional space whereby a linear separation can be done.

The RBF kernel is defined as:

$$K(x_i, x_j) = e^{-\gamma \| x_i - x_j \|^2}$$

Where, K is the kernel function, controls the shape of the boundary, and $\| x_i - x_j \|^2$ represents the squared Euclidean distance between two samples.

The smaller the values of γ are the smoother the decision boundaries, whereas the larger the values of γ they are the more complex the decision boundaries. RBF kernel has been found to work well in high dimensional biomedical data analyzed when the boundaries between classes are nonlinear and in cases where the interactions among features are complicated.

K nearest neighbors

K Nearest Neighbors (KNN) looks at the most similar tumors and predicts the class based on what it is most similar to. Distance is calculated using Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum (x_i - x_j)^2},$$

Where, d is the distance between samples, where smaller distance means more similarity. The class is determined based on the majority of the nearest neighbors.

The value of k was set to 5 in this study. Selecting a small odd value, such as $k = 5$ helps reduce sensitivity to noise while avoiding ties in binary classification problems. Very small values such as $k = 1$ may lead to overfitting, whereas large values may over smooth the decision boundary. Therefore, $k = 5$ provides a balance between bias and variance in classification performance.

Evaluation metrics

Model performance was evaluated using four standard classification metrics: accuracy, precision, recall, and F1 score [5]. These metrics provide a comprehensive assessment of classification performance, especially in medical diagnosis where correctly identifying malignant tumors is critical.

Accuracy

Accuracy measures the overall proportion of correctly classified samples. It is defined as,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

Where, *TP* True Positives (malignant correctly classified), *N* is True Negatives (benign correctly classified), *FP* is False Positives (benign incorrectly classified as malignant), and *FN* False Negatives (malignant incorrectly classified as benign). Accuracy provides an overall measure of model correctness.

Precision

Precision measures how often tumors predicted as malignant are actually malignant. It is defined as,

$$Precision = \frac{TP}{TP + FP}.$$

Higher precision indicates fewer false positive errors.

Recall (Sensitivity)

Recall measures how many actual malignant tumors were correctly identified. It is defined as,

$$Recall = \frac{TP}{TP + FN}.$$

Recall is especially important in medical diagnosis because missing malignant tumors can have serious consequences.

F1 score

F1 score provides a balanced measure between precision and recall. It is defined as,

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

The F1 score is useful when both false positives and false negatives must be minimized.

Results and Discussion

Figure 4 shows a comparison of the accuracies of the six ML models studied. It is seen that SVM model, with the RBF kernel gave the highest accuracy of 98.25%, followed by LR with 96.49%. DT has the lowest of these 6, at 91.23%.

Table 3 shows a comparison of these model performances on the attributes of the supervised models, such as Precision, Recall and F1-score, along with the accuracy. It shows an overview of the detailed performance measures of each classifier with the accuracy, preciseness, recall, and F1 score of the malignant and benign classes. The highest performance of the Support Vector Machine was observed in all measures of evaluation, and an overall accuracy of 98.25 and F1 score of 98.5. Random Forest, Logistic Regression, and KNN also showed good performance with the accuracy value of more than 95. Comparatively, Naive Bayes and Decision Tree had lower recall of malignant cases which implies that they are less sensitive about recognizing cancerous tumors. These findings demonstrate that SVM is the most reliable and balanced model in terms of classification of this dataset.

Figure 5 shows the comparison of the macro-averaged precision and recall of the classifiers evaluated. Precision and recall are particularly important in the diagnosis of breast cancer, as high precision reduces unnecessary false positives while high recall ensures that malignant tumors are not missed. Again, the SVM model demonstrates the strongest balance between high precision and high recall, indicating its capability to detect cancerous cases reliably with minimal clinically costly errors, particularly for malignant cases, while Naive Bayes and Decision Tree models show lower recall and reduced sensitivity.

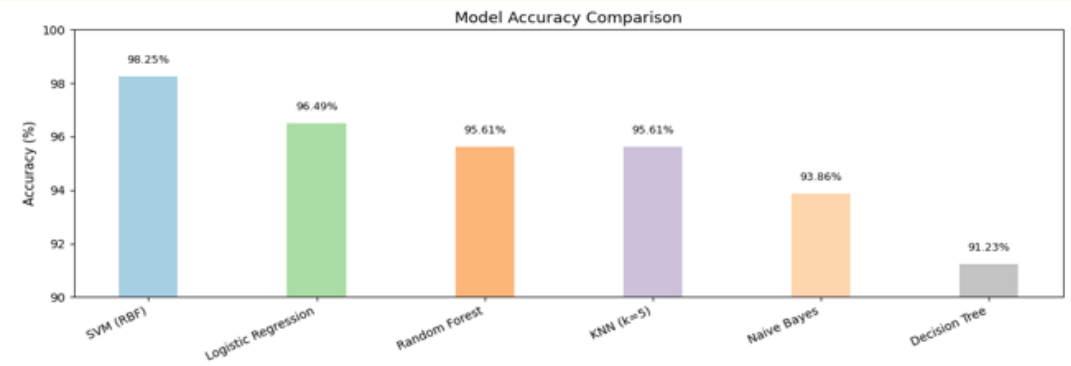

Figure 4: Comparison of the six model accuracies.

Model	Accuracy	Precision			Recall			F1- Score		
		M	B	O	M	B	O	M	B	O
LR	96.49%	97%	96%	96%	93%	99%	96%	95%	97%	96%
NB	93.86%	93%	95%	94%	90%	96%	93%	92%	95%	93.5%
DT	91.23%	85%	96%	90.5%	93%	90%	91.5%	89%	93%	91%
RF	95.61%	95%	96%	95.5%	93%	97%	95%	94%	97%	95.5%
SVM	98.25%	98%	99%	98.5%	98%	99%	98.5%	98%	99%	98.5%
KNN (k = 5)	95.61%	95%	96%	95.5%	93%	97%	95%	94%	97%	95.5%

Table 3: Model performance (M-Malignant, B-Benign, O-overall).

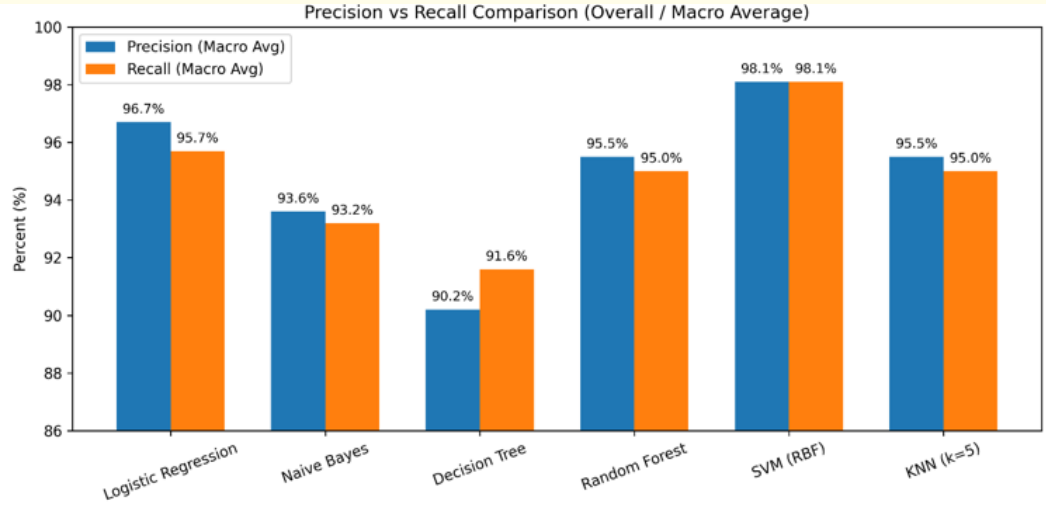

Figure 5: Comparison of precision vs Recall for the six models.

Figure 6 presents the confusion matrices (CMs) for each of the six machine learning classifiers used in this study. Confusion matrices provide a concise overview of classification performance by reporting the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), representing the correct and incorrect predictions of malignant and benign tumors. False negatives are especially to be considered in medical diagnosis issues like detecting breast cancer since the false negative is a malignant tumor that is mistaken as a benign one, thereby delaying diagnosis and treatment. The SVM classifier was the most effective, producing only one false negative and one false positive while correctly classifying 41 malignant and 71 benign tumors. Logistic Regression, Random Forest, and KNN also showed strong results with relatively few misclassifications, whereas Naive Bayes and Decision Tree indicated slightly higher error rates. These findings confirm that the SVM model provides the best overall classification performance for the Breast Cancer Wisconsin Diagnostic dataset.

The details of these CMs are:

Logistic regression

In the case of the Logistic Regression model, the confusion matrix indicates that there are 39 true positives, 3 false negatives, 1 false positive and 71 true negatives. This implies that there were 39 malignant tumors which were rightly classified as malignant, and 3 malignant tumors which were erroneously called benign. Moreover 71 benign tumors were accurately diagnosed as benign,

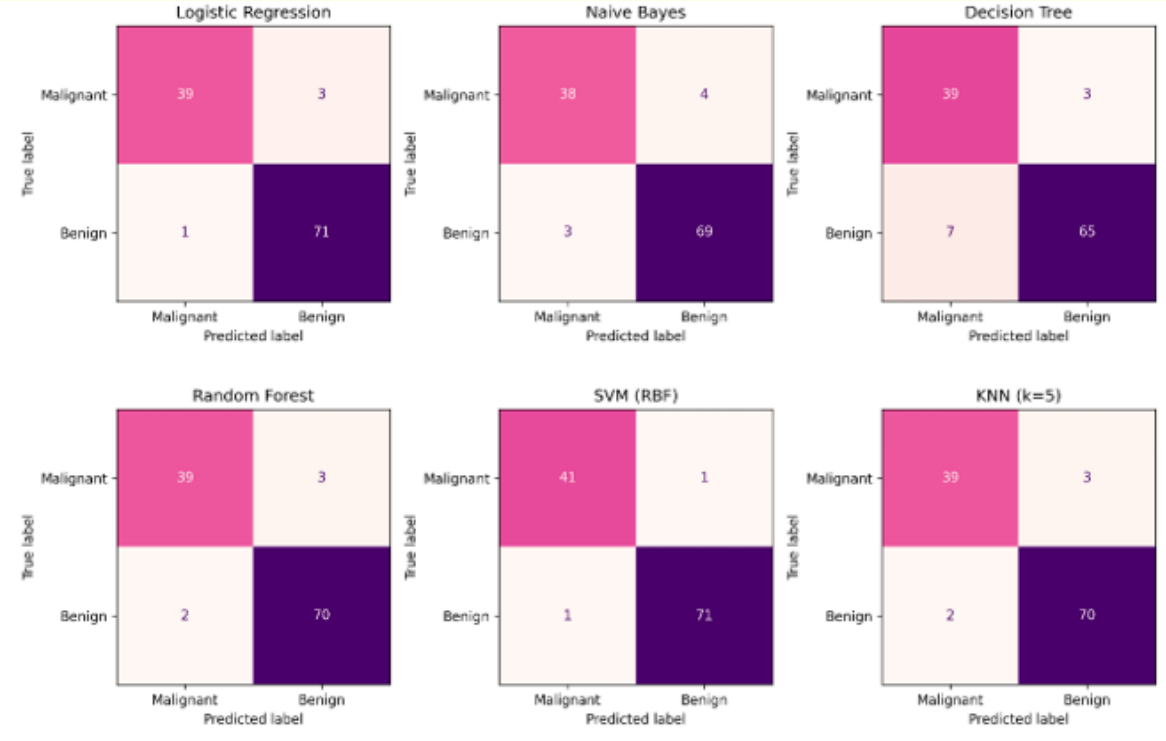

Figure 6: Confusion Matrix of Six Supervised ML Classifiers Evaluated.

and 1 benign tumor was wrongly diagnosed as malignant. Such findings suggest that Logistic Regression is a fairly efficient tool that has only few errors in classification.

Naive bayes

In the case of the Naive Bayes model, the confusion table indicates that there were 38 true positives, 4 false negatives, 3 false positives and 69 true negatives. This implies that there were 38 malignant tumors that were appropriately detected whereas 4 malignant tumors were overlooked and termed as benign. Meanwhile, 69 benign tumors were accurately identified and 3 benign tumors were falsely identified malignant. Naive Bayes has a few more errors, particularly in the detection of malignant tumors, than some of the other classifiers.

Decision tree

In the case of the Decision Tree model, the confusion matrix is 39, 3, 7 and 65 true positives, false negatives, false positives and true negatives respectively. This means that there was a correct classification of 39 malignant tumors as malignant and there were 3 malignant tumors that were wrongly categorized as benign. Besides this, 65 benign tumors were correctly reported yet 7 benign tumors were incorrectly reported as malignant. The false positives are quite high indicating that the Decision Tree will more frequently false classify some benign tumors than the other models.

Random forest

In the case of the Random Forest model, the confusion matrix indicates that there were 39 true positives, 3 false negatives, 2 false positives and 70 true negatives. This implies that 39 malignant tumors were rightly identified and 3 malignant tumors were categorized as benign. On the same note, there were 70 benign tumors that were classified correctly and only 2 benign tumors that were wrongly predicted as malignant. These findings indicate that Random Forest is a high-quality model, and it has quite limited classification errors and a satisfactory sensitivity-specificity ratio.

Support vector machine

In the case of the Support Vector Machine model, the confusion matrix indicates the number of true positives to be 41, false

negative 1, false positive 1 and true negative to be 71. This means that 41 malignant tumors were rightly identified as malignant tumors and only 1 malignant tumor was incorrectly identified as a benign tumor. Also 71 benign tumors had been properly classified as benign, and only 1 benign tumor had been incorrectly characterized as malignant. These findings indicate that the SVM classifier offers the most overall accuracy, and fewest cases of misclassifications of any of the six models.

K nearest neighbors

In the case of the K Nearest Neighbors model, the confusion matrix indicates that there were 39 true positives, 3 false negatives, 2 false positives and 70 true negatives. This implies that 39 malignant tumors were rightly diagnosed, and 3 malignant tumors were not diagnosed and were diagnosed as benign. Meanwhile, 70 benign tumors were rightly identified and 2 benign tumors were wrongly identified as malignant. These outcomes demonstrate that KNN also gives good results and there is not many errors in the classification and its results are almost identical to those of the Random Forest model.

In general, the confusion matrices prove that the Support Vector Machine demonstrated the best performance of the classification, then Random Forest, K Nearest Neighbors, and Logistic Regression. Naive Bayes and Decision Tree had a relatively greater misclassification rate particularly in regard to false negative or false positive and this decreased their overall performance. This is in line with earlier studies which have shown high classification accuracy of machine learning algorithms in breast cancer classification problems. For example, Islam., *et al*. [8] showed high classification accuracy of supervised machine learning algorithms in classification problems involving breast cancer datasets.

Table 4 shows a list of previous studies using the Wisconsin breast cancer dataset and a comparison of their classification accuracies of these along with current study. The study applied a fixed 80-20 train test split of 455 training samples and 114 testing samples as compared to some of the previous studies that employed k-fold cross validation. Cross validation assesses models on more than one data partition, and may also give a slightly more accurate estimate of its validity. Although a hard independent test set was used, the Support Vector Machine scored 98.25, which is within the top range of reported results. This shows that SVM model extrapolates well to unknown data and has high classification capability even when it is applied in realistic evaluation scenarios. These results verify that SVM is among the most effective and reliable methods of classifying breasts cancer using the Wisconsin Diagnostic dataset.

Study	Year	Evaluation Method	Logistic Regression	Naive Bayes	Decision Tree	Random Forest	KNN	SVM
Agarap [16]	2017	70/30 split	96.49%	93.86%	—	—	—	97.66%
Jeyasingh [17]	2017	10 fold CV	96.80%	95.20%	94.10%	97.10%	96.40%	98.00%
Entezari [18]	2018	10 fold CV	96.10%	94.30%	93.50%	96.80%	95.70%	97.90%
Patgiri [19]	2019	80/20 split	95.70%	93.90%	92.60%	96.50%	95.20%	98.10%
This Study	2026	80/20 split	96.49%	93.86%	91.23%	95.61%	95.61%	98.25%

Table 4: Partial list of studies using the Wisconsin data set.

Among the various techniques, the performance of SVMs has been found to be higher in breast cancer classification problems due to their ability to deal with high-dimensional feature spaces and the presence of non-linear decision boundaries through the use of the RBF kernel method [4]. It is established that the performance of SVMs is higher compared to probabilistic models and tree-based

models when the class boundaries overlap, which is a common occurrence in medical datasets based on tumor morphology [4]. However, in the majority of the existing literature, the performance of the models is based on the overall accuracy of the model without considering the performance of the model in terms of the recall and F1 score for malignant cases, where the cost of a false negative is extremely high [20].

The superior performance of SVM could be attributed to several factors:

- **Non-linear decision boundaries**: The RBF kernel allows SVM to classify malignant versus benign tumors even when the decision boundary is non-linear by mapping the data into a higher-dimensional space [4].
- **Margin maximization**: SVM constructs a separating hyperplane that maximizes the margin between classes, enabling more robust decisions when samples overlap or include noise [4].
- **Resistance to overfitting**: Unlike Decision Trees, which may memorize the training data, SVM tends to generalize better, especially when working with high-dimensional biomedical datasets [3].
- **Captures complex tumor features**: Breast cancer morphology includes subtle variations in concavity, smoothness, texture, and symmetry. The RBF kernel captures these multi-dimensional relationships more effectively than linear models or simple distance-based methods [3].

Overall, the SVM model demonstrated the most consistent and reliable performance across the majority of the samples.

Most analyses of the Wisconsin Breast Cancer Diagnostic dataset show that only a few of the thirty features contribute most to distinguishing malignant from benign tumors. Prior research consistently highlights concave_points_worst, radius_worst, perimeter_worst, area_worst, and concavity_mean as top predictors because they reflect nuclear shape irregularity and increased cell size, both of which are strong indicators of malignant progression [3]. These features capture structural abnormalities, such as jagged borders and uneven growth rather than uniform enlargement, making them more discriminative than texture-based measurements.

The significance of these high-variance, shape-related features also aligns with the strong performance of the SVM model in this study. The RBF kernel can model nonlinear boundaries formed by complex geometric differences, allowing it to separate overlapping classes in high-dimensional space [4]. In contrast, Decision Trees may overfit boundary noise, and Naive Bayes assumes feature independence, which limits its ability to capture interactions between size and concavity metrics. Because SVM maximizes class separation and generalizes well with morphological data, it is well suited for datasets where malignancy is expressed through boundary distortion and irregular nuclear structure rather than linearly separable relationships [4].

Conclusion

Wisconsin Breast Cancer Diagnostic Dataset with 569 patients were analyzed using six supervised ML models. SVM with the RBF kernel proved to be the most effective classifier for distinguishing between malignant and benign breast tumors using the dataset. Model performance is evaluated using accuracy, precision, recall, and F1 score, with particular emphasis on malignant class recall due to its clinical importance. Consistent with prior studies in biomedical machine learning, the Support Vector Machine with a radial basis function kernel demonstrated the strongest overall performance, achieving highest accuracy and the highest F1 scores across both classes.

With an accuracy of 98.25%, SVM outperformed Logistic Regression, Random Forest, KNN, Naive Bayes, and Decision Tree models. The strong margin-based separation produced by the RBF kernel allowed SVM to distinguish between malignant and benign tumors even when feature distributions overlapped, a common challenge in medical datasets. This highlights SVM's ability to generalize well and avoid overfitting despite the presence of noisy or correlated features.

Through this project, we implemented the full supervised-learning pipeline from preprocessing and feature scaling to model training, evaluation, and comparison, demonstrating how machine learning can be systematically applied to a real medical dataset. These results indicate that margin-based classifiers are particularly effective when working with high-dimensional biomedical data containing subtle morphological differences.

Beyond overall accuracy, evaluation metrics such as precision, recall, and F1-score further supported the robustness of the SVM model. While other models demonstrated competitive performance, they showed reduced sensitivity in identifying malignant tumors, which is a critical limitation in clinical applications. In contrast, SVM maintained consistently high performance across all evaluation metrics, reinforcing its suitability for diagnostic decision-support systems where minimizing false negatives is essential.

Beyond the technical performance, this research highlights the broader impact of applying machine learning to healthcare. Automated classification systems can help reduce diagnostic variability, support overburdened clinical workflows, and provide rapid second-opinion assessments that may detect malignancy earlier than traditional methods alone. As ML tools continue to advance, they can assist clinicians in making more informed, data-driven decisions, ultimately contributing to improved patient prognosis and better clinical outcomes.

Future work may include hyperparameter optimization, cross-validation on additional and more diverse clinical datasets, and the development of an interactive prediction tool for real-world healthcare environments. With further validation and refinement, SVM-based diagnostic systems may eventually support routine screening, cytology review, and early-detection programs, serving as a valuable complement to existing clinical diagnostic practices.

Bibliography

1. Bray F., *et al*. "Global cancer statistics and trends". *CA: A Cancer Journal for Clinicians* (2023).
2. Topol E. "High-performance medicine: the convergence of AI and healthcare". *Nature Medicine* (2019).
3. Ashraf A., *et al*. "A review of breast cancer classification using FNA cytology and machine learning". *Diagnostics* (2020).
4. Gupta D and Bhavsar H. "SVM-based cancer diagnosis: a comprehensive review". *Computer Methods and Programs in Biomedicine*.
5. Ahmad., *et al*. "Using three machine learning techniques for predicting breast cancer recurrence". *Journal of Health and Medical Informatics* 4 (2013): 2.
6. Uğuz H. "A two-stage feature selection method for breast cancer diagnosis using machine learning". *Computers in Biology and Medicine* (2019).
7. Sahu., *et al*. "Extraction of key features and enhanced prediction framework of breast cancer occurrence". Proceedings of the Sixth International Conference on Trends in Electronics and Informatics (ICOEI 2022). IEEE (2022).
8. Islam., *et al*. "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI". *Scientific Reports* 14 (2024): 8487.
9. Eke., *et al*. "Early detection of Alzheimer's disease with blood plasma proteins using support vector machines". *IEEE Journal of Biomedical and Health Informatics* 25 (2020): 218-226.
10. Castellazzi., *et al*. "A machine leaning approach for the differential diagnosis of Alzheimer and vascular demetia fed by MRI selected features". *Frontiers in Neuroinformation* 14 (2020): 25.
11. Berrar D. "Performance measures for binary classification in medical decision making". *Artificial Intelligence in Medicine* (2021).
12. Google Research. Google Colaboratory (2024).
13. Raschka S and Mirjalili V. "Python Machine Learning". 3rd ed. Packt Publishing (2019).
14. Berrar D. "Cross-validation and train test split in medical machine learning". *Encyclopedia of Bioinformatics and Computational Biology* (2019).
15. Géron A. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow". 3rd ed. O'Reilly Media (2022).
16. Agarap AF. "On breast cancer detection using machine learning". Proceedings of the 2017 International Conference on Machine Learning Applications (ICMLA), IEEE (2017): 1–6.
17. Jeyasingh M. "Breast cancer prediction using supervised machine learning techniques". *International Journal of Engineering Research and Technology* 6.5 (2017): 45–50.
18. Entezari M and Hassanpour H. "Breast cancer diagnosis using machine learning algorithms". *Journal of Medical Systems* 42.6 (2018): Article 112.
19. Patgiri R and Ahmed A. "Machine learning approaches for breast cancer detection". *International Journal of Advanced Computer Science and Applications* 10.3 (2019): 95–102.
20. Chicco D and Jurman G. "The advantages of the Matthews correlation coefficient a over F1 score and accuracy". *BMC Genomics* (2020).