A Comparative Analysis of Machine Learning Models for Breast Tumor Classification

Arshia Samimi, Amirhossein Raeis Samiee, Sara Nasher-Ahkami

Department of Mathematics and Computer science

Iran University of Science and Technology, Iran

ABSTRACT

This study investigates the performance of various machine learning models in classifying breast tumors as benign or malignant using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. We begin by replicating a previously published logistic regression-based study and extend it by evaluating four additional models including: Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, Random Forest, Linear Regression and K-Nearest Neighbors (KNN). We apply Min-Max normalization and conduct a stratified 80/20 train-test split. We use accuracy, precision, recall, F1 score, Area Under the Curve (AUC), and Root Mean Squared Error (RMSE) as our evaluation metrics similar to the referenced paper for the sake of fair comparison. Cross-validation is performed to compare models average and standard deviation scores. After this phase both for the sake of generalizability check and as final comparison we train models on the whole training set, then we tune the hyperparameters using random search algorithm and verify our results on the test set. These results show that while logistic regression performs reliably, SVM achieves the highest average performance.

Keywords: Logistic Regression, SVM, AUC, KNN, WDBC, Breast Cancer

1. INTRODUCTION

Breast cancer is the most frequently diagnosed cancer among women worldwide and one of the leading causes of cancer-related mortality. In 2022, approximately 2.3 million women were diagnosed with breast cancer globally, resulting in around 670,000 deaths, making it a major global health burden [12]. It accounts for roughly one in every eight cancer diagnoses and about 15% of all cancer deaths in women [13]. While advances in detection and treatment have improved survival rates, especially in high-income countries, significant disparities remain. In low- and middle-income countries, where screening infrastructure is limited and access to care is often delayed, mortality rates are notably higher [14].

Breast cancer typically begins in the cells of the ducts or lobules of the breast. While the exact cause is unknown, risk factors include age, gender, genetics (such as BRCA1 and BRCA2 mutations), hormonal influences, and lifestyle factors like obesity and alcohol consumption [15]. The risk increases with age, peaking in women aged 70 and above, but it is also seen increasingly among younger women, particularly in urban settings [16].

Diagnosis plays a crucial role in reducing mortality. When detected early and localized, the five-year relative survival rate for breast cancer exceeds 90% [17]. However, not all cases are detected early. Mammography, the primary screening tool, has an average sensitivity of 75% to 85%, though this drops significantly in women with dense breast tissue [18]. Other diagnostic tools, such as ultrasound and magnetic resonance imaging (MRI), can supplement mammography but may not be routinely available in all settings due to cost and accessibility [19]. In the United States, for example, the incidence rate of breast cancer is approximately 130.8 cases per 100,000 women, with a mortality rate of 19.2 per 100,000 [20].

Despite improvements in diagnostic methods, the current tools are not perfect. False positives often lead to unnecessary biopsies and anxiety, while false negatives can delay treatment. In recent years, machine learning (ML) and artificial intelligence (AI) have emerged as powerful approaches to enhance diagnostic accuracy and reliability. By learning from large datasets, ML models can detect complex patterns in imaging and clinical data that may be difficult for the human eye to recognize. In a large retrospective study, a deep learning system trained on more than one million mammograms achieved a performance comparable to experienced radiologists, with an area under the curve (AUC) of 0.895. Moreover, when the system's output was combined with radiologist interpretations, diagnostic accuracy improved further [21].

Beyond imaging, ML has been applied to structured datasets like the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, which contains 30 features extracted from digitized images of fine needle aspirates. Logistic regression, support vector machines (SVM), random forest, and neural networks have all been employed to classify tumors as benign or malignant. In a comparative analysis, logistic regression achieved an accuracy of 96.5%, with a precision of 95.7%, recall of 95.6%, AUC of 96.4%, and root mean square error (RMSE) of 0.19 [1]. Other studies report similar or higher performance using more complex models such as deep neural networks and ensemble methods, sometimes reaching accuracies above 98% [22][23].

Despite this impressive performance, many ML studies differ in methodology, preprocessing steps, and evaluation metrics, making it difficult to compare models directly. Furthermore, while complex models may offer marginal gains in accuracy, they often sacrifice interpretability—an important factor in clinical settings where physicians must understand and trust the decision process. This has led to growing interest in explainable AI (XAI), which provides tools like SHAP values to interpret model predictions and support clinical decision-making [24].

In this study, we replicate the methodology of Maulidia et al. [1], who used logistic regression on the WBCD dataset with Min-Max normalization and cross-validation. We extend their work by including additional models such as support vector machine (SVM), random forest, linear regression, and k-nearest neighbors (KNN). Our goal is to evaluate these models fairly using consistent preprocessing and validation techniques. We compare not only accuracy and AUC but also interpretability and robustness across methods. The broader aim of this work is to offer guidance in selecting machine learning models for breast cancer diagnosis, considering both clinical objectives and performance trade-offs. This comparative analysis is intended to bridge the gap between promising academic results and practical clinical application.

2. METHODOLOGY

In this study we used the following flowchart depicted in fig 1. we can divide this flowchart into 4 main parts: dataset, preprocessing, model training and results analysis. In this section we mainly focus on the dataset used, preprocessing, models and chosen metrics. We will discuss the result analysis in the next section.

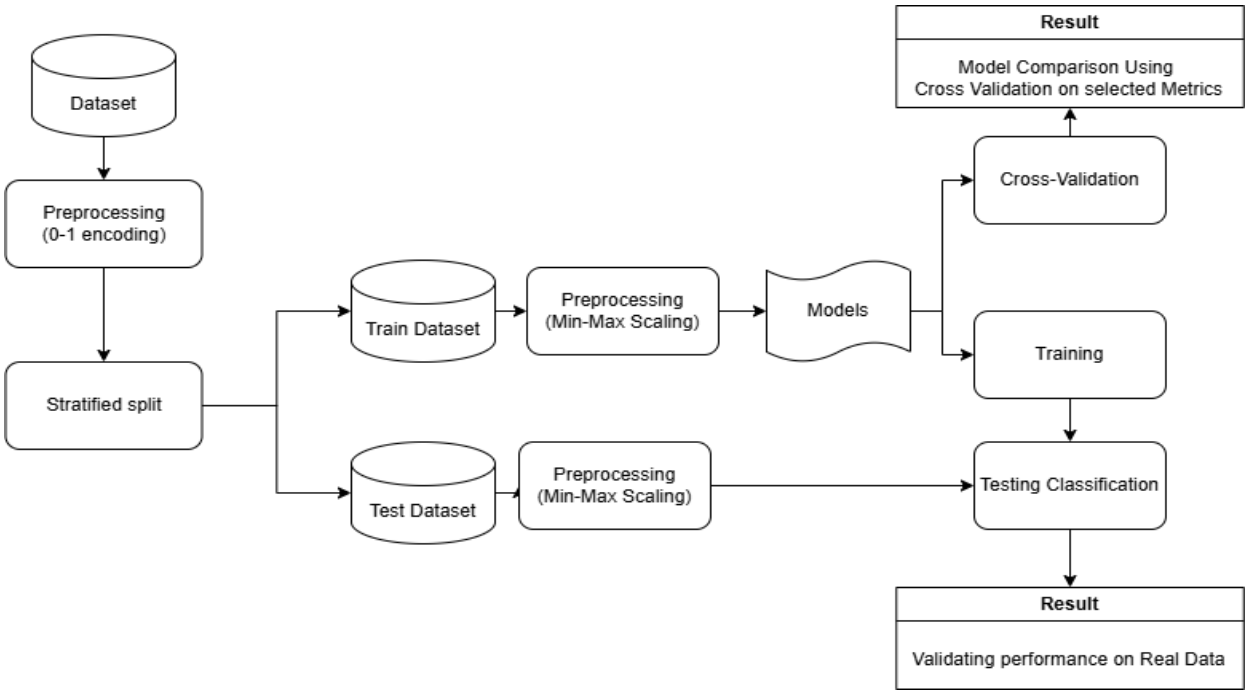

Fig 1. Flow of the Work

2.1 Dataset

In this study, the dataset used is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, a well-established benchmark for evaluating classification algorithms in medical diagnostics. The dataset was originally published by the UCI Machine Learning Repository and accessed via Kaggle for convenience and reproducibility. It was compiled by Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital in Madison, USA. The data collection method employed was Fine Needle Aspiration (FNA), a minimally invasive procedure that extracts cellular material from breast masses using a thin needle. These samples were digitized and analyzed to extract morphological features of the cell nuclei. The digitization process involved converting microscopic images into numerical data, allowing for quantitative analysis of cellular structures.

Each sample in the dataset corresponds to a breast mass biopsy and includes a set of features derived from the digitized images of the FNA samples. The dataset contains 569 instances and 32 columns. The first column is an ID number, followed by the diagnosis label, which is the target variable indicating whether the tumor is malignant (M) or benign (B). There are 357 benign and 212 malignant cases. This distribution is represented in table 1. The remaining 30 columns represent real-valued features that describe various characteristics of the cell nuclei. These features are computed using three statistical

descriptors for each of the 10 base measurements: mean, standard error, and worst (i.e., the largest mean value among the three largest measurements). This results in a total of 30 features that capture the shape, texture, and structural complexity of the nuclei. These descriptors help capture both the central tendency and variability of each feature, offering a more comprehensive view of the cellular morphology.

The dataset is clean and well-structured, with no missing values, making it suitable for machine learning applications. All features are continuous and have been normalized to ensure consistent scaling across models. The diagnosis column is categorical, with values 'M' for malignant and 'B' for benign. The dataset has been widely used in research and practice to evaluate the performance of various classification algorithms, including Support Vector Machines (SVM), Random Forests, K-Nearest Neighbors (KNN), Logistic Regression, and Neural Networks. Its consistent structure and high-quality annotations make it a reliable resource for both academic studies and clinical decision support systems.

The features extracted from the FNA images are critical in distinguishing between malignant and benign tumors. These features reflect the structural irregularities and morphological differences that are often indicative of cancerous cells. The 10 core features used in this dataset are represented in table 2. Each of these features is represented in three forms—mean, standard error, and worst—resulting in a rich and detailed feature set that enables precise classification. The dataset's balanced structure and clear class separation make it ideal for evaluating model performance in terms of accuracy, recall, precision, and F1-score. Its use in breast cancer detection research has contributed significantly to the development of automated diagnostic tools that support early detection and treatment planning. Moreover, the dataset's interpretability and compatibility with various machine learning frameworks make it a valuable asset for both exploratory analysis and production-level deployment.

Additionally, the dataset's relatively modest size allows for rapid experimentation and prototyping, making it ideal for academic projects and algorithm benchmarking. Its binary classification nature simplifies evaluation while still presenting meaningful challenges in terms of feature selection and model tuning. Researchers and practitioners often use this dataset to explore the trade-offs between sensitivity and specificity, especially in clinical contexts where false negatives can have serious consequences.

Table 1. WBCD Dataset Description

Attributes	*Number of Attributes*
Sample total	569
Dimensionality	30
Classes	2
Sample per class	*Benign*: 357 (62.74%) *Malignant*: 212 (37.26%)

Table 2. core features used in WDBC dataset

Feature Name	*Description*
Radius	Mean distance from the center to perimeter points
Texture	Standard deviation of grayscale pixel values
Perimeter	Total boundary length of the nucleus
Area	Size of the nuclues
Smoothness	Variation in radius lengths
Compactness	(Perimeter^2 / Area) - 1.0
Concavity	Degree of concave portions of the nucleus contour
Concave Points	Number of concave segments on the contour
Symmetry	Symmetry of the nucleus shape
Fractal Dimension	Complexity of the nucleus boundary (coastline approximation)

2.2 Preprocessing

First we start the preprocessing by changing the M labels (malignant) to 1 and B labels (benign) to 0. Making these labels numerical gives our models the ability to actually predict them [25]. Then we split the dataset using the same approach as [1], into train data (80% of the dataset) and test data (20% of the dataset). To ensure robust model evaluation, we employed stratified sampling for train-test splitting. This method preserves the original class distribution (benign vs. malignant tumors) in both training and test subsets, By maintaining proportional representation of both classes, stratified sampling prevents evaluation bias that might occur if rare malignant cases were underrepresented in the test set. This approach yields more reliable performance estimates and enhances model generalizability, ensuring our classifier is evaluated under clinically realistic conditions where accurate malignancy detection is critical. For medical diagnostic applications where false negatives carry significant consequences, stratified sampling provides statistically rigorous validation of model robustness [26]. The splitted data is described in table 3 below.

Table 3. Splitting Data

Dataset	*Malignant*	*Benign*	*Total Data*
Training set	170 (37.36%)	285 (62.64%)	455
Testing set	42 (36.84%)	72 (63.16%)	114
Total Data	212 (37.26%)	357 (62.74%)	569

After splitting the data, it's time for scaling it. Data scaling is a critical preprocessing step because many machine learning algorithms assume or perform better when features lie on a similar numerical range. Without scaling, variables with large magnitudes (e.g., "area" or "perimeter" features) can dominate those with smaller ranges (e.g., "smoothness"), causing models—especially those based on gradient optimization (like logistic regression) or distance measures (like KNN, SVM with RBF)—to converge slowly, become numerically unstable, or prioritize the wrong features. By normalizing each feature to a common scale (e.g., Min–Max scaling to [0,1] as in [1], or standardization to zero mean and unit variance), we ensure that no single feature unduly biases the loss function or distance computations.

Moreover, proper scaling improves the conditioning of the underlying matrices in optimization routines, leading to more reliable convergence of solvers (e.g., Newton or gradient-based methods) and more consistent performance across cross-validation folds. In short, scaling aligns the feature distributions, prevents numerical issues, and yields fairer, more stable model training and evaluation. Its important to make sure scaling happens separately on train and test data so the train data wont be affected from the test data. That's why we put scaling after splitting the data [27].

We—for the sake of fair comparison— use the Min-Max scaler similar to what [1] has done. We use the formula below to scale our data [1]:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where:

X' = Attribute to be normalized

X_{min}= The smallest value of the attribute

X_{max}= The largest value of the attribute

After going through the scaling stage, the training data will be processed by each algorithm to carry out the training process using the cross-validation evaluation method.

In the following section, we continue by giving a brief description of each model used.

2.3 Models

2.3.1 Logistic Regression

Logistic Regression is a generalized linear model for binary classification that models the log-odds of the positive class as a linear combination of input features [4]. Denoting the feature vector by X and weights by β, It assumes:

$$Pr(y=1 \mid X) = \sigma(\beta^T X) = \frac{1}{1+e^{-\beta^T X}} \tag{2}$$

Parameters β are estimated by maximizing the conditional likelihood (equivalently minimizing the negative log-likelihood or cross-entropy loss) using numerical optimization methods (e.g., Newton-Raphson or gradient-based solvers). The loss function can be written as:

$$\ell(\beta) = -\sum_{i=1}^{n} [y_i log\sigma(\beta^T x_i) + (1-y_i) log(1-\sigma(\beta^T x_i))] \tag{3}$$

Logistic Regression yields calibrated probability estimates, enabling threshold adjustment when balancing sensitivity vs. specificity—critical in medical diagnosis. Its linear decision boundary makes it interpretable: each coefficient reflects the log-odds change per unit increase in a standardized feature.

Logistic regression is widely used in recent healthcare studies due to its simplicity, interpretability, and effectiveness. In a 2025 systematic review, it was found that logistic regression is still the most commonly applied model for clinical risk prediction, especially when odds ratios and confidence intervals are needed for interpretability [28]. In breast cancer diagnosis, recent studies have reported accuracies above 96%, with some achieving precision and recall over 97% when applied to the WBCD dataset [29]. Hybrid approaches, such as combining logistic regression with AdaBoost, have further improved performance while maintaining transparency [30].

Beyond oncology, logistic regression is actively used in cardiovascular disease prediction and diabetes risk stratification, often embedded within ensemble or hybrid models to balance predictive power and interpretability [31][32]. A 2025 study on gallstone risk prediction, for instance, used logistic regression as the final step in a hybrid framework, leveraging its clinical interpretability after more complex feature selection [33].

These examples show that logistic regression remains a foundational tool in clinical machine learning, often serving as both a baseline and a reliable component in modern hybrid systems. In this study we implemented logistic regression using scikit-learn version 1.7.0's LogisticRegression class within a pipeline. All input features were first normalized using Min-Max scaling to the [0,1] range. This normalization is particularly important for linear models with regularization, ensuring all features are on a comparable scale so that the regularization term penalizes each feature equitably.

The model was trained and evaluated using 10-fold stratified cross-validation, which preserved class balance across splits. This approach provided a stable estimate of generalization performance while guarding against variance introduced by random sampling, especially valuable for moderately sized and slightly imbalanced datasets like WBCD.

To fine-tune the model, we employed RandomizedSearchCV to explore hyperparameter combinations efficiently. You can see the results of this search in table 4. Different Precision and Recall scores are also plotted against different values of hyperparameter C which is arguably the most important hyperparameter of Logistic Regression in fig 2.

Table 4. Logistic regressions hyperparameters table

Hyperparameter	*Description*	*Final value*
C (Regularization)	Inverse of L2 regularization strength	2.481
penalty	Type of regularization norm	L2
fit_intercept	Whether to learn the intercept	True

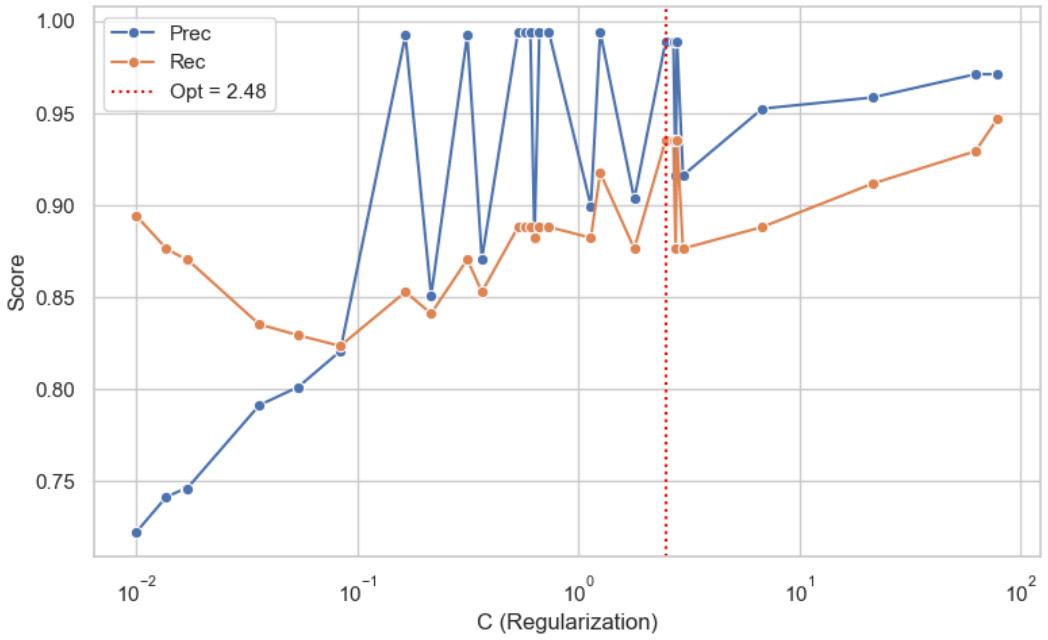

Fig 2. Precision and Recall Scores vs. Regularization Strength (C) in Logistic Regression searched by randomized search cv

Results indicates that moderate regularization performed best on our dataset, preventing overfitting while retaining predictive strength. The model was trained using the liblinear solver, which is efficient and well-suited to smaller datasets and L2-regularized logistic loss. The solver uses a trust-region Newton method, ensuring reliable convergence on convex problems.

2.3.2 Linear Regression Classifier

As a baseline, the Linear Regression classifier fits a continuous-valued linear model to the binary labels (0/1) by ordinary least squares, then we predict by comparing the computed value with the threshold to assign classes. Formally, it minimizes the following loss function [5]:

$$min_{\beta} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 \tag{4}$$

yielding prediction $\hat{y} = \beta^T X$. Although not derived for classification, thresholding with a value less than 0.5 provides a decision rule; continuous outputs can serve as "scores" for AUC after clipping into [0,1]. This approach lacks the probabilistic foundation of logistic regression and can suffer from heteroscedastic residuals and predictions outside [0,1], but it is simple to implement and illustrates the benefit of classification-specific loss. In our workflow, we wrap a linear regression estimator to output binary labels via:

$$predict(X): \hat{y} = \beta^T x;\ class = 1\ if\ \hat{y} >= 0.4\ else\ 0. \tag{5}$$

Although linear regression lacks a probabilistic foundation, it can perform well for classification tasks when applied to structured datasets like the WBCD, where the features of benign and malignant cases are nearly linearly separable [2]. In such settings, even though linear regression minimizes squared error rather than cross-entropy, the resulting decision boundary can correctly classify a high proportion of samples. Its simplicity and interpretability make it a useful baseline model.

Several studies have explored when linear regression is appropriate for classification. For example, Hastie et al. note that linear regression performs adequately when the classes are well-separated and the relationship between inputs and labels is approximately linear [5]. In gene expression studies, linear regression has been used for two-class classification problems with success, especially when paired with dimensionality reduction techniques like principal component analysis or partial least squares [34]. In breast cancer diagnosis, regression-based classification combined with linear programming has also been used to distinguish between benign and malignant tumors with high accuracy [35].

However, the approach suffers from key limitations: predictions are unbounded (not restricted to [0,1]), and it assumes constant variance (homoscedasticity) across all input ranges. These drawbacks are especially problematic in medical contexts, where calibrated probabilities and robust uncertainty estimates are essential. As such, comparing logistic and linear regression highlights the practical advantages of classification-specific loss functions and probabilistic modeling.

We trained the model using scikit-learn 1.7.0's LinearRegression, fitting parameters using ordinary least squares, solved via singular value decomposition (SVD). All features were first normalized using Min-Max scaling to ensure uniform feature contribution and stability in SVD computations.

Predicted outputs were clipped to the [0,1] interval before being interpreted as confidence scores for AUC computation. The model's classification decision rule was based on comparing outputs to the learned threshold (0.4496). This hyperparameter tuning was done using a custom score function which maximizes recall where accuracy > 93%, precision > 90% . The thresholding results are shown in fig 3 bellow.

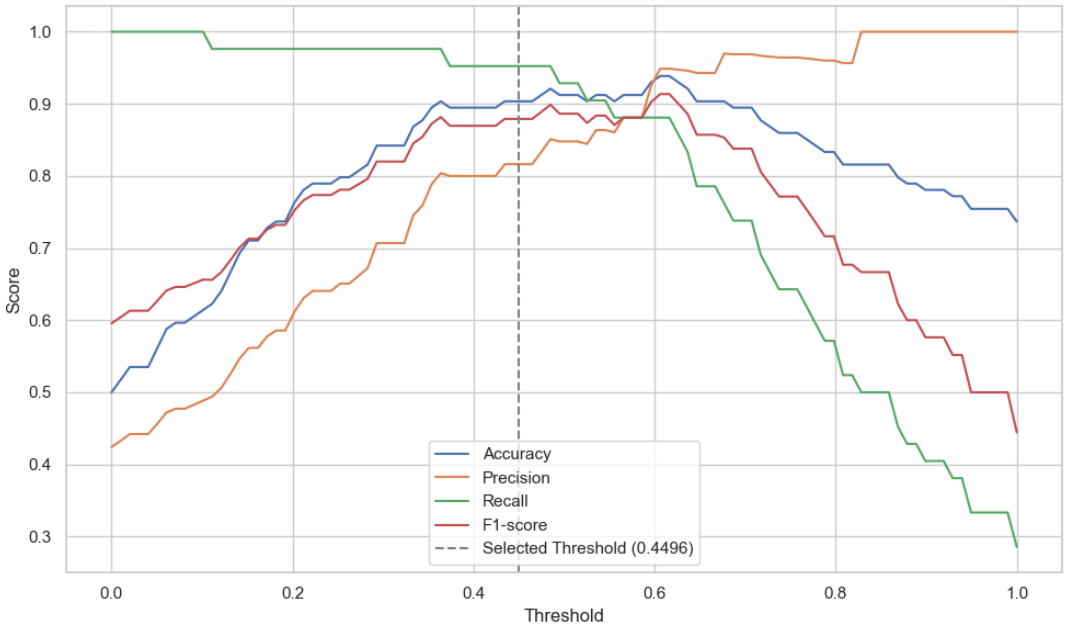

Fig 3. Linear Regression Classifier Performance at Varying Thresholds

Although Linear Regression lacks probabilistic interpretability, this method yields consistent performance on structured data and allows direct comparison with classification-specific methods like logistic regression.

2.3.3 Random Forest

Random Forest is an ensemble of decision trees trained via bootstrap aggregation and random feature selection. Each tree partitions the feature space by binary splits on individual features, capturing nonlinear relationships and interactions without requiring scaling. Training grows each tree on a bootstrap sample of the training data, and at each node a random subset of features is considered for splitting, which decorrelates trees and reduces variance. For classification, each tree votes for a class and the forest predicts by majority vote [6]:

$$\hat{y} = \arg\max_{c \in \{0,1\}} \sum_{t=1}^{T} 1\{h_t(\mathbf{x}) = c\} \tag{6}$$

Random Forest is favored in medical science due to its robustness to overfitting, ability to handle high-dimensional and noisy data, and built-in feature importance estimation, which offers partial interpretability in clinical decision-making [6]. Its ensemble nature allows it to capture complex, nonlinear interactions among features—essential in domains like genomics, imaging, and diagnostics—without requiring extensive preprocessing or distributional assumptions.

Recent applications underscore its versatility across medical fields. Wallace et al. [36] used Random Forests to predict stroke incidence and emphasized the importance of using unbiased variable importance metrics. In a 2024 study, Random Forest was applied to blood biomarkers for COVID-19 diagnosis, achieving AUCs above 0.86 and identifying IL-6 and CRP as key indicators via SHAP values [37]. In

imaging, it has shown performance comparable to deep learning on smaller datasets where neural networks are harder to train effectively [38].

Additionally, Balsters et al. [39] employed Random Forest with SHAP explanations to classify autism spectrum disorder using resting-state fMRI, achieving an AUC of 0.94. In cardiovascular research, SHAP-enhanced Random Forest models helped identify major risk factors with both high predictive power and interpretability [40]. Fascia [41] applied Random Forests to real-time health monitoring, using SHAP to explain how demographic and environmental features affected patient risk scores.

These studies reflect Random Forest's continued value in medical settings where accuracy, resistance to overfitting, and post-hoc interpretability are critical.

In out study Random Forest was implemented using scikit-learn 1.7.0's RandomForestClassifier. We initially trained the model with 100 trees using bootstrap aggregation (bagging) on 80% of the stratified training data. Each decision tree used Gini impurity for splitting and was allowed to grow to full depth without pruning.

To optimize performance and prevent overfitting, we applied RandomizedSearchCV with 10-fold cross-validation, tuning key hyperparameters for F1-score. The best configuration, shown in table 5, used 187 trees (Precision and Recall scores is plotted against different randomized search values of trees in fig 4) with a maximum depth of 20. It selected features per split using the log2 strategy, split nodes with at least 3 samples, and required at least 1 sample per leaf. These settings balanced complexity and generalization, yielding a robust model suited for the moderately sized WBCD dataset.

Table 5. Random forest model hyperparameters

Hyperparameter	*Description*	*Final Value*
n_estimators	Number of trees in the forest	187
max_depth	Maximum depth of individual trees	20
min_samples_split	Minimum samples required to split a node	3
min_samples_leaf	Minimum samples required at a leaf node	1
max_features	Number of features to consider at each split	log2

This tuned configuration not only enhanced predictive accuracy but also maintained robustness and interpretability through feature importance scores. These outcomes confirm Random Forest's suitability for structured medical datasets like WBCD, especially when both performance and reliability are required.

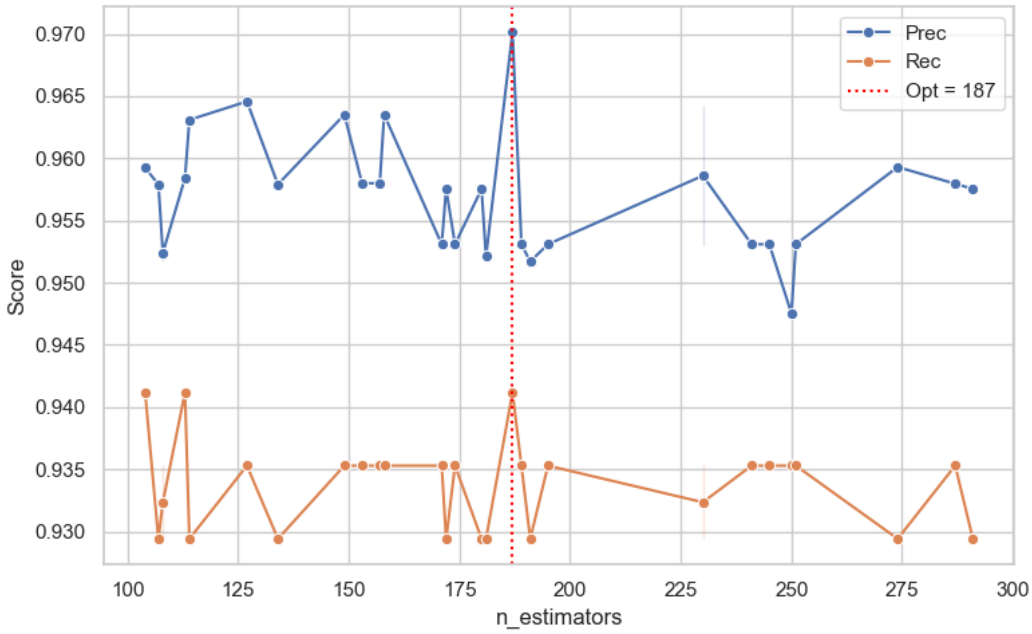

Fig 4. Precision and Recall scores plotted against different values of n_estimators searched by randomized search cv

2.3.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is an instance-based classifier that assigns a label to a query point based on the majority class among its k closest neighbors in feature space using a distance metric (commonly Euclidean). Feature scaling is critical because distance comparisons are sensitive to feature ranges [7]. The decision rule is:

$$y\,(X) = mode\left\{ y_i : i \in N_k(X) \;\; where\; N_k(X)\; are\; the\; indices\; of\;\; the\; k-smallest\; ||x - x_i|| \right\} \quad (7)$$

The hyperparameter k balances bias and variance: small k yields flexible but noisy boundaries, while large k oversmooths distinctions. At prediction time, KNN computes distances to all training points, which can be computationally intensive for large datasets but is acceptable for moderate-sized datasets like WDBC. KNN can model arbitrarily complex boundaries but is sensitive to irrelevant or correlated features.

KNN remains a competitive non-parametric approach in recent clinical research due to its simplicity, intuitive nature, and adaptability to multimodal data. A 2023 study applied KNN to classify breast cancer subtypes using genomic features, achieving over 95% accuracy after dimensionality reduction [42]. In dermatology, KNN has been integrated with feature selection to detect skin cancer in dermoscopic images with promising results [43]. Another 2024 investigation used KNN in a multi-model ensemble for early Alzheimer's detection, benefiting from its low bias and local sensitivity [44]. KNN has also proven effective in real-time health monitoring, where its non-parametric design is useful for adapting to changing data distributions [45].

These works demonstrate that, while simple, KNN continues to serve as a valuable model in medical diagnostics, especially when combined with preprocessing techniques that mitigate its sensitivity to noise and dimensionality.

We implemented the K-Nearest Neighbors classifier using scikit-learn version 1.7.0's KNeighborsClassifier, with the default number of neighbors initially set to 5. Since KNN relies heavily on distance computations, we first applied Min-Max scaling to all input features to ensure each feature contributed equally during Euclidean distance calculation. This normalization step is crucial because unscaled features can distort distance metrics and bias the classifier. Unlike parametric models, KNN makes no assumptions about the underlying data distribution, instead memorizing the training set and deferring computation to prediction time.

To optimize the model's performance, we tuned the hyperparameter k (number of neighbors) using RandomizedSearchCV, allowing us to find the best balance between underfitting and overfitting (see Fig. 4). Smaller values of k result in more flexible decision boundaries but can be sensitive to noise, while larger values increase robustness at the cost of boundary sharpness. Predictions were made by majority vote among the k closest points in the training set, determined by scaled Euclidean distance. We also randomized over additional hyperparameters such as distance weighting (uniform vs. distance) and the Minkowski power parameter (p), enhancing the model's flexibility. The final selected hyperparameters are summarized in table 6.

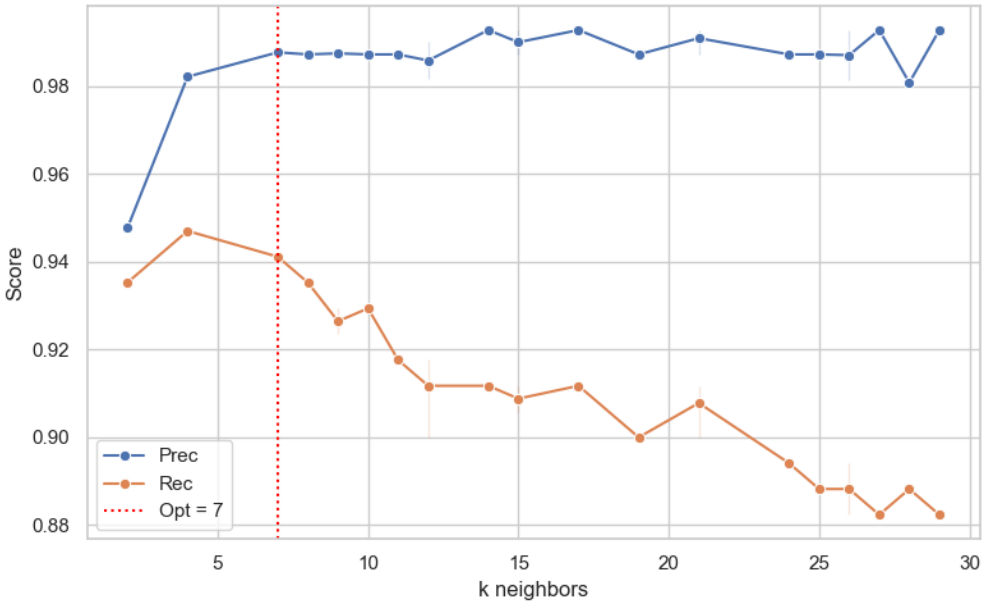

Fig 4. Precision and Recall values plotted against different k values searched by randomized search cv

(unsearched points are interpolated)

Table 6. Final Hyperparameters for K-Nearest Neighbors

Hyperparameter	*Description*	*Final Value*
k_neighbors	Number of nearest neighbors to consider	4
weights	Weight function used in prediction	distance
p	Power parameter for the Minkowski metric	1 (Manhattan distance)

2.3.5 Support Vector Machine (SVM) with RBF Kernel

Support Vector Machine constructs a decision boundary that maximizes the margin between classes in a (possibly transformed) feature space. For nonlinearly separable data, the RBF (Gaussian) kernel maps inputs into a high-dimensional space, allowing complex decision boundaries [8]. The optimization solves:

$$min_{w,b,\xi} \tfrac{1}{2} ||w||^2 + C\sum \xi_i, subject\ to\ y_i(w^T\varphi(X_i) + b) \geq 1 - \xi_i \quad (8)$$

The RBF kernel is defined as:

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2) \quad (9)$$

Key hyperparameters are C (regularization) and γ (kernel width), typically tuned via grid search within cross-validation. SVM is effective in high-dimensional spaces and robust to overfitting when regularized, though sensitive to feature scaling and computationally intensive for large datasets. In breast cancer prediction, SVM with RBF often yields high accuracy and AUC by capturing subtle nonlinear patterns among features.

In breast cancer prediction, SVM with RBF kernel often achieves accuracies above 95% and AUCs over 0.96 by capturing subtle nonlinear feature interactions [47]. Its adaptability makes it a preferred choice for modeling complex clinical data.

Outside oncology, SVM-RBF has been applied to cardiovascular risk prediction and neurological disease classification, leveraging clinical and imaging data to improve early detection and prognosis [48][49]. For example, a 2025 study integrated SVM-RBF with multi-modal features to predict heart failure, enhancing sensitivity without sacrificing specificity [50]. Another recent application used SVM-RBF for early Alzheimer's diagnosis via MRI data, outperforming linear models [51]. The RBF kernel's flexibility allows SVM to handle diverse biomedical data such as genetic expression and radiomics, cementing its role in precision medicine [52].

The Support Vector Machine was implemented using scikit-learn version 1.7.0's SVC class. The inclusion of probability=True allowed for the computation of calibrated class probabilities using Platt scaling, which is particularly important in medical applications where probabilistic confidence in predictions is often required for decision-making, such as threshold tuning or ROC analysis. Training was conducted using the sequential minimal optimization (SMO) algorithm, which is the default solver in scikit-learn's SVC. All input features were preprocessed using Min-Max scaling to normalize their ranges between 0 and 1, since SVMs are highly sensitive to the scale of the input features due to their reliance on

distance-based computations and kernel evaluations. Hyperparameters were selected through randomized search within cross-validation. The regularization parameter C, which controls the penalty for misclassified training examples, was set to approximately 57.28, indicating that the model prioritized minimizing classification errors, potentially at the expense of a narrower margin. The gamma parameter for the RBF kernel, which controls the influence of individual training examples in shaping the decision boundary, was found to be approximately 0.0484. Values of This relatively small gamma resulted in a smoother and more generalizable decision surface. These parameters were chosen to strike an optimal balance between model complexity and generalization performance, based on the highest F1 score observed during the randomized cross-validation search. The final hyperparameters are shown in table 7. Also different

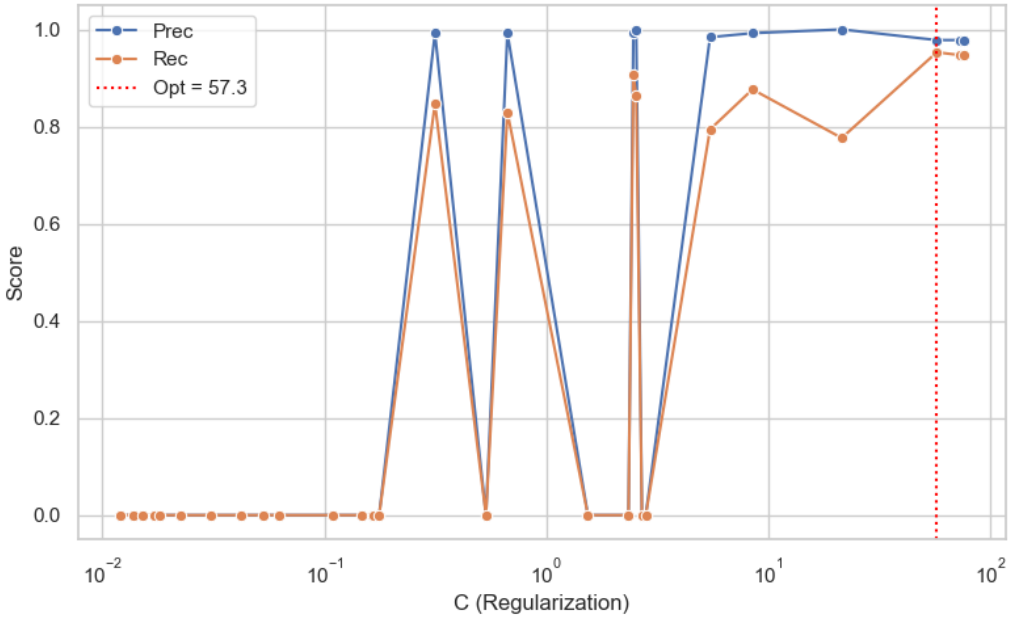

Fig 5. Hyperparameter Tuning Results for SVM (RBF) via randomized search

Table 7. Final Hyperparameters for SVM (support vector machine)

Hyperparameter	*Description*	*Final Value*
C	Regularization parameter (error penalty)	57.279
gamma	RBF kernel coefficient (controls locality)	0.0484
kernel	Type of kernel used	RBF
probability	Enable probability estimates via Platt scaling	True

Now we continue by explaining our metrics. Again for the sake of fair and correct comparison we have chosen the exact same metrics as [1].

2.4 Metrics

2.4.1 Confusion Matrix

A confusion matrix is a performance measurement tool used in machine learning to evaluate the accuracy of a classification model. It provides a detailed breakdown of correct and incorrect predictions by comparing the actual labels with the predicted labels [9]. A typical binary confusion matrix (fig 6) consists of four key components:

- **True Positives (TP):** Cases where the model correctly predicts the positive class.
- **True Negatives (TN):** Cases where the model correctly predicts the negative class.
- **False Positives (FP):** Cases where the model incorrectly predicts the positive class (Type I error).
- **False Negatives (FN):** Cases where the model incorrectly predicts the negative class (Type II error).

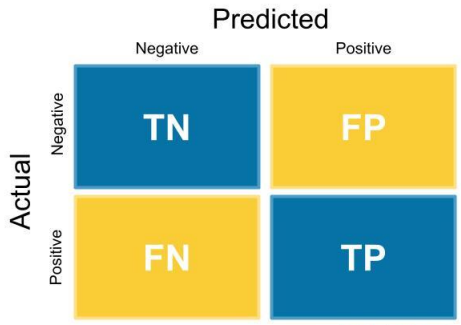

Fig 6. Confusion Matrix

Some key metrics can be derived from the confusion matrix. Below we mentioned the ones that are used in [1].

1) **Accuracy**: Measures the overall correctness of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

2) **Precision**: Indicates the proportion of correctly predicted positive instances out of all predicted positives.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

3) **Recall** (Sensitivity/True Positive Rate): Measures the proportion of actual positives correctly identified by the model.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

4) **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two.

$$F1 = 2 \times \frac{Percision \times Recall}{Percision + Recall} \tag{13}$$

These metrics help assess a model's performance from different perspectives, ensuring a comprehensive evaluation beyond just accuracy

2.4.2 AUC Score

The AUC-ROC (Area Under the Receiver Operating Characteristic Curve) score measures a binary classifier's ability to distinguish between positive and negative classes across all possible decision thresholds. Unlike accuracy, which depends on a single threshold, the AUC evaluates performance by plotting the true positive rate (sensitivity) against the false positive rate (1 − specificity) at various thresholds. The score ranges from 0 to 1, where 1 indicates perfect classification, 0.5 suggests no better than random guessing, and values below 0.5 signal poor model performance. Particularly useful for imbalanced datasets, the AUC provides a threshold-independent assessment of a model's discriminative power. However, it may not always reflect performance on minority classes and is primarily designed for binary classification. A high AUC (e.g., 0.9) implies strong separability, while lower values indicate weaker predictive capability [10]. Again we choose this metric to compare with [1]. An example of AUC-ROC is below at fig 7.

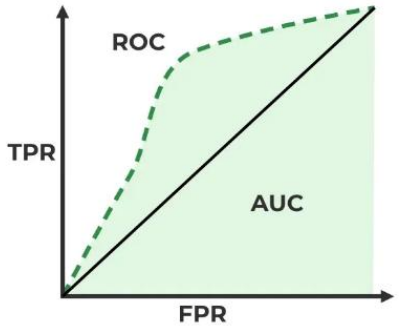

Fig 7. The ROC and AUC Metric

2.4.3 Root-Mean-Squared-Error (RMSE)

Root Mean Square Error (RMSE) is a widely used metric for evaluating the accuracy of regression models by measuring the differences between predicted and actual values [11]. It calculates the square root of the average squared errors, giving more weight to larger deviations due to the squaring operation. RMSE is expressed in the same units as the target variable, making it easy to interpret. A lower RMSE indicates better model performance, with zero representing perfect predictions. While RMSE is sensitive

to outliers, its emphasis on larger errors makes it particularly useful for applications where significant deviations are costly. However, it should be used alongside other metrics like MAE (Mean Absolute Error) to get a comprehensive view of model performance, as RMSE alone doesn't distinguish between over- and under-prediction.

The standard RMSE is computed using the below equation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{14}$$

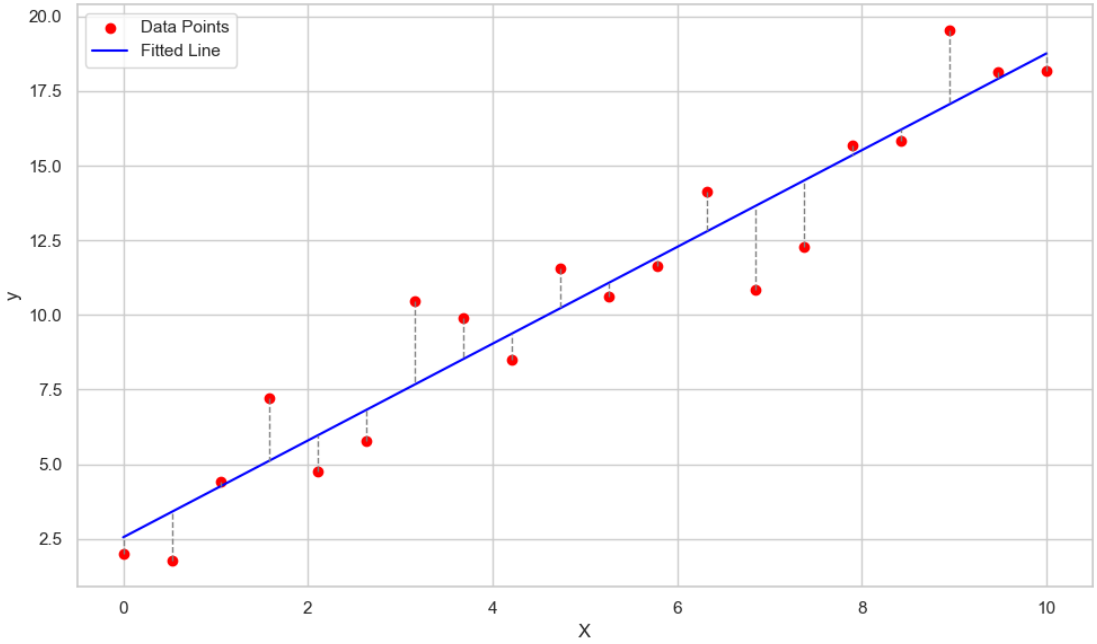

Fig 8. Root Mean Squared Error (RMSE) Visualization (RMSE = 1.47)

3. RESULTS

This section presents the findings from the machine learning model evaluations conducted in this study. Initially, the performance of the replicated logistic regression model is detailed and compared against the previously published results. Subsequently, the performance metrics for all implemented models—Logistic Regression, Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, Random Forest, K-Nearest Neighbors (KNN), and Linear Regression Classifier—are presented, highlighting their effectiveness in classifying breast cancer tumors.

3.1 Comparing Results With Reference

Following the methodology outlined in Fig. 1, the initial phase of this study involved replicating the logistic regression model presented by [1]. The objective was to assess the reproducibility of their results

using a comparable dataset and evaluation pipeline. Table 8 provides a direct comparison between the performance metrics reported in their study and those obtained in our replication on the test set.

Table 8. Comparing Results For Logistic Regression

Model	Accuracy	Precision	Recall	F1	AUC	RMSE
Referenced Study [1]	96.5%	95.7%	95.6%	95.6%	96.4%	0.19
Our Study	98.25%	100.0%	95.24%	97.56%	99.8%	0.13

As shown in Table 8, our replicated logistic regression model produced results that are consistent with those reported in [1], while slightly outperforming them in most metrics. Our model achieved an accuracy of 98.25% and an RMSE of 0.13, compared to the 96.5% accuracy and 0.19 RMSE reported in the original study. These improvements likely stem from refined hyperparameter tuning and implementation differences such as data splitting randomness, preprocessing steps, or specific library versions. Although recall was marginally lower, the results remain within an expected variance range for machine learning experiments and suggest that the core findings of [1] are generally reproducible.

This successful replication establishes a reliable baseline, which we use to conduct a broader comparative evaluation of additional machine learning models. The following sections assess whether other classifiers offer improved robustness or diagnostic performance on the WBCD dataset.

3.2 Cross-Validation Performance

To ensure reliable performance evaluation prior to final testing, each model was assessed using 10-fold stratified cross-validation on the training dataset. This technique partitions the training data into ten equally sized folds, ensuring that the proportion of malignant and benign samples is preserved in each subset. In each iteration, nine folds are used for training while the remaining fold is reserved for validation. This process is repeated ten times, with each fold serving as the validation set once, and the results are averaged. This approach helps mitigate overfitting and provides a more robust estimate of model generalization, especially when working with limited data, as is often the case in medical diagnostics.

table 9 reports the mean scores across all ten folds for each evaluation metric, while table 10 presents the standard deviation (std) of those metrics, providing insight into each model's consistency and robustness across different data splits.

Table 9. Mean Scores using 10-fold Cross-Validation

Model	Accuracy Mean	Precision Mean	Recall Mean	F1 Mean	AUC Mean	RMSE Mean
Logistic Regression	96.27%	99.41%	90.59%	94.59%	99.11%	0.1586
SVM (RBF)	97.14%	97.64%	94.71%	95.97%	99.17%	0.1155
Random Forest	96.26%	96.35%	93.53%	94.84%	99.01%	0.1466
KNN	96.91%	98.57%	92.94%	92.94%	98.50%	0.1157
Linear Regression	95.39%	99.23%	88.24%	93.28%	99.40%	0.1949

The SVM with RBF kernel achieved the best overall performance, with the highest mean accuracy (97.14%) and F1-score (95.97%). It also recorded the top recall (94.71%), meaning it is particularly effective at correctly identifying malignant tumors — a crucial factor in medical contexts where false negatives carry serious risks. Its strong balance between precision (97.64%) and recall highlights its robustness across metrics.

The Logistic Regression model demonstrated the highest precision (99.41%), indicating that when it predicts a tumor as malignant, it is almost always correct. However, its recall (90.59%) was lower than SVM and Random Forest, meaning it is more conservative and risks missing some malignant cases. Despite this, its strong accuracy (96.27%), balanced F1-score (94.59%), and high AUC (99.11%) make it a solid and reliable baseline model.

The Random Forest classifier provided consistently strong performance, achieving 96.26% accuracy with a solid balance between precision (96.35%) and recall (93.53%). Its F1-score (94.84%) reflects this equilibrium. However, its RMSE (0.1466) was slightly higher compared to SVM and KNN, suggesting occasional deviations in prediction confidence. Still, its ensemble-based nature contributes to robustness and generalization.

The K-Nearest Neighbors (KNN) model performed competitively, with accuracy (96.91%) close to SVM. It achieved high precision (98.57%), but its recall (92.94%) was slightly lower, which pulled down its F1-score (92.94%). While strong overall, KNN's sensitivity to data structure and noise could explain its slightly less consistent balance compared to SVM and Random Forest.

The Linear Regression classifier, although not traditionally used for binary classification, showed reasonable results with accuracy (95.39%) and the highest AUC (99.40%), indicating excellent separation ability between classes across thresholds. However, its recall (88.24%) and F1-score (93.28%) were the lowest among the models, suggesting it may miss more malignant cases compared to others. While less competitive in critical performance metrics, its simplicity and interpretability might make it useful in resource-constrained or exploratory scenarios.

Table 10. Std Scores using 10-fold Cross-Validation

Model	Accuracy Std	Precision Std	Recall Std	F1 Std	AUC Std	RMSE Std
Logistic Regression	0.0296	0.0176	0.0798	0.0441	0.0150	0.1102
SVM (RBF)	0.0357	0.0375	0.0809	0.0516	0.0154	0.1233
Random Forest	0.0355	0.0398	0.0668	0.0491	0.0127	0.1262
KNN	0.0467	0.0429	0.0941	0.0680	0.0249	0.1324
Linear Regression	0.0360	0.0231	0.0789	0.0530	0.0086	0.0902

The standard deviation values provide insights into the stability of each model's performance across cross-validation folds. While mean scores highlight overall predictive strength, variability measures reveal how consistently these results can be reproduced under different training–validation splits.

The Logistic Regression model demonstrated strong stability, with the lowest standard deviations in precision (0.0176) and accuracy (0.0296). This indicates it consistently produced reliable results, even when trained on slightly different subsets of data. Its recall (0.0798) and F1 (0.0441) variability were slightly higher, but still lower than most other models. Combined with its strong mean performance, Logistic Regression stands out as both robust and dependable.

The SVM with RBF kernel showed moderate consistency, with standard deviations of 0.0357 in accuracy and 0.0516 in F1-score. While its mean recall was the highest among all models, its recall variability (0.0809) suggests some sensitivity to fold composition. This reflects the model's flexibility: it adapts well to different data partitions but may not always generalize with the same confidence across folds.

The Random Forest classifier had a variability profile similar to SVM. Its accuracy (0.0355) and F1 (0.0491) standard deviations were close to those of SVM, though its recall variability (0.0668) was somewhat lower, pointing to more consistent detection of malignant cases. However, its RMSE standard deviation (0.1262) was relatively high, indicating greater fluctuation in error magnitudes. This balance of moderate consistency and occasional variability is typical of ensemble methods.

The K-Nearest Neighbors (KNN) model was the least stable overall, with the highest standard deviations in recall (0.0941), F1 (0.0680), accuracy (0.0467), and RMSE (0.1324). These values highlight KNN's strong dependence on the specific partitioning of the dataset. Since it relies directly on neighborhood structures, minor differences in training folds can lead to larger fluctuations in predictions, making it less reliable compared to other models.

The Linear Regression classifier, while included for comparison, is not theoretically suitable for binary classification tasks. Its predictions are unbounded and do not represent probabilities, which makes evaluation against classification metrics somewhat misleading. Although it showed the lowest standard deviation in AUC (0.0086) and generally low variability in precision (0.0231), these results should be interpreted with caution. Linear Regression's apparent stability does not translate into practical suitability, and models like Logistic Regression or SVM remain more appropriate for medical classification problems.

In summary, Logistic Regression and SVM demonstrated the best combination of high mean performance and consistent stability. Random Forest followed closely, while KNN showed the highest variability, limiting its reliability. Linear Regression, although stable in some metrics, is fundamentally unsuitable for classification and should not be considered a viable alternative in practice.

3.3 Final Test Set Performance

While cross-validation provides reliable estimates of model performance across different partitions of the training data, evaluation on an independent, unseen test set offers the most objective measure of a model's ability to generalize. After tuning hyperparameters through cross-validation, each model was retrained on the full training dataset and then evaluated on the hold-out test set.

The results on the test set closely mirrored the trends observed during cross-validation. The **Support** Vector Machine (RBF kernel**)** achieved the strongest overall performance, with the highest accuracy (99.12%) and recall (97.62%) alongside a perfect precision score of 100.00%. Its F1-score of 98.80% and AUC of 99.64% confirm that it was not only highly accurate but also particularly effective at minimizing false negatives — a critical consideration in medical applications.

The Logistic Regression model also performed competitively, reaching 98.25% accuracy and 95.24% recall, while maintaining a perfect precision of 100.00%. Importantly, it yielded the highest AUC score (99.80%) among all models, underlining its exceptional ability to distinguish between classes across thresholds. With an F1-score of 97.56%, Logistic Regression demonstrated balanced, reliable performance on unseen data.

The Linear Regression classifier, though not theoretically well-suited for binary classification, was included for comparative purposes. It reported strong numerical results on the test set, achieving 98.25% accuracy, 100.00% precision, 95.24% recall, and an AUC of 99.70%. However, because it does not model probabilities and can produce unbounded predictions, its apparent competitiveness should be interpreted cautiously; Logistic Regression remains the more appropriate linear model for classification.

The Random Forest classifier achieved an accuracy of 97.37%, with perfect precision (100.00%) but a lower recall (92.86%) compared to SVM and Logistic Regression. This indicates that while it reliably identified malignant predictions when made, it missed a few positive cases, reducing its sensitivity slightly.

The K-Nearest Neighbors (KNN) classifier attained 96.49% accuracy, with precision (97.50%) and recall (92.86%) that lagged slightly behind the top-performing models. Its F1-score of 95.12% was the lowest among all classifiers on the test set, suggesting that KNN may generalize less effectively to unseen data, consistent with the variability observed during cross-validation.

Overall, the test set results reinforce the cross-validation findings: SVM with RBF kernel emerged as the most reliable and well-balanced model, followed closely by Logistic Regression. Random Forest offered strong but slightly less sensitive predictions, while KNN showed weaker generalization. Linear Regression appeared competitive in raw metrics but remains conceptually unsuitable for binary classification.

Table 11. Results on The Test Set

Model	Accuracy	Precision	Recall	F1	AUC	RMSE
Logistic Regression	98.25%	100.00%	95.24%	97.56%	99.80%	0.1325
SVM (RBF)	99.12%	100.00%	97.62%	98.80%	99.64%	0.0937
Random Forest	97.37%	100.00%	92.86%	96.30%	99.44%	0.1622
K-Nearest Neighbors	96.49%	97.50%	92.86%	95.12%	98.56%	0.1873
Linear Regression	98.25%	100.00%	95.24%	97.56%	99.70%	0.1325

These test results confirm the generalization behavior of the trained models and serve as a final benchmark for their predictive capability on unseen data.

4. CONCLUSION

This study undertook a detailed comparative evaluation of several widely used machine learning algorithms for breast cancer diagnosis using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The investigation was motivated by the crucial need for accurate, reliable, and interpretable diagnostic tools that can assist healthcare professionals in early detection and decision-making. In alignment with this goal, the research was structured in two phases: first, a faithful reproduction of a previously published logistic regression model for baseline comparison, and second, an extensive benchmarking of alternative models including Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel, Random Forest, K-Nearest Neighbors (KNN), and a custom Linear Regression Classifier.

The initial replication successfully mirrored the original logistic regression implementation, achieving even superior performance on key evaluation metrics such as accuracy, recall, and AUC. This not only reaffirmed the reproducibility of the referenced study's approach but also established a solid foundation for subsequent model comparisons. By adhering closely to the same preprocessing, model design, and evaluation pipeline (including normalization and train-test splitting), we ensured that observed performance differences in later phases could be more confidently attributed to algorithmic strengths rather than inconsistencies in experimental design.

Once the baseline was established, we extended the study by systematically evaluating a broader array of algorithms. These models were carefully tuned using randomized hyperparameter search combined with 10-fold stratified cross-validation. This ensured that each model configuration was optimized for generalization, minimizing overfitting and enhancing fairness in the evaluation process. Evaluation metrics included not only common performance indicators such as accuracy and precision but also more diagnostic-specific measures like recall (which is especially crucial in detecting malignant tumors), F1-score, AUC, and RMSE.

From the results obtained, SVM with RBF kernel consistently emerged as the top performer across multiple dimensions. On the unseen test set, it achieved the highest accuracy (99.12%) and F1-score (98.80%), and importantly, perfect precision (100.00%) alongside the highest recall of all models (97.62%). These metrics indicate its exceptional ability to correctly identify malignant cases while

minimizing false alarms. Logistic Regression also performed impressively, with very close performance to SVM, particularly in AUC (99.80%) and RMSE.

Meanwhile, the Linear Regression Classifier—despite not being a traditional classifier—produced strong numerical results on both test and cross-validation sets (98.25% accuracy, 100.00% precision, 95.24% recall, 99.70% AUC). However, it is important to note that Linear Regression is not theoretically suitable for classification tasks. Its predictions are unbounded and do not represent probabilities, which makes these results less meaningful for real-world diagnostic use. The observed performance underscores why Logistic Regression is the more appropriate linear model for classification, even though Linear Regression can appear competitive in controlled experiments.

Another important aspect of this study was the analysis of model stability and consistency, as quantified by the standard deviations of cross-validation metrics. This analysis is vital in assessing the reliability of a model across different samples of data. A model that performs well on average but exhibits high variability may not be dependable in clinical applications where decisions based on a single patient's data must be consistently accurate. In our cross-validation results, the Linear Regression Classifier demonstrated low variability across key metrics, particularly in AUC (standard deviation of 0.0086) and F1-score (0.0530), suggesting that its predictions are stable across different training subsets. Logistic Regression followed closely in terms of consistency, exhibiting minimal deviation in accuracy (0.0296) and precision (0.0176), which further reinforces its robustness as a classification tool.

On the other hand, K-Nearest Neighbors (KNN) exhibited relatively high standard deviations across most metrics, particularly recall (0.0941) and F1-score (0.0680). This is indicative of KNN’s sensitivity to data partitioning, which arises due to its instance-based, non-parametric nature. When the number of neighbors (k) is small, small shifts in training data can significantly alter classification boundaries. This instability, combined with the lower performance metrics on the test set, suggests that KNN may not be ideal for sensitive medical diagnostics without further ensemble strategies or neighborhood weighting adjustments.

Random Forest, though slightly behind the top models in terms of average test performance, exhibited balanced results in both consistency and accuracy. Its ensemble nature—constructing multiple decision trees and averaging their outputs—provides a buffer against overfitting, and the model maintained respectable metrics such as 97.37% accuracy and 96.30% F1-score on the test set. It also offers inherent feature importance interpretability, which could be valuable in biomedical contexts where understanding the relevance of each input feature is clinically meaningful.

Notably, the SVM (RBF) model maintained low standard deviation across accuracy (0.0357) and AUC scores (0.0154) and was less variable than KNN and Random Forest, though slightly more variable than Logistic and Linear Regression in some metrics. Nonetheless, its overall performance profile strongly suggests that it is a highly suitable model for breast cancer prediction, especially when the cost of misclassification is high. In particular, the combination of perfect precision and the highest recall (97.62%) among all evaluated models on the test set makes it an excellent candidate for diagnostic systems aiming to minimize both false positives and false negatives.

In comparing these models, it is important to emphasize that model interpretability, computational complexity, and ease of deployment are also critical factors. Logistic Regression, despite being an older and simpler model, offers high transparency and explainability. The weights learned by the model directly correspond to the influence of each feature on the prediction, making it favorable in medical settings where interpretability is often mandated by ethical or regulatory guidelines. Similarly, the Linear

Regression Classifier, while unorthodox and unsuitable for classification tasks, maintains a similar level of simplicity and clarity in how predictions are made, though its outputs must be interpreted with caution.

4.1 Model Selection Guidelines:

Given the results of both cross-validation and final test set evaluation, model selection for practical applications—particularly in clinical settings like breast cancer diagnosis—must be informed by multiple considerations. These include not only performance metrics such as accuracy and recall, but also stability, interpretability, computational cost, and the clinical risk associated with false negatives or false positives. Below, we provide a comprehensive breakdown of the model-specific insights and practical recommendations derived from this study.

Diagnostic Safety (Minimizing False Negatives):

In medical diagnostics, particularly oncology, false negatives carry severe consequences. Missing a malignant tumor could result in delayed treatment and worsened patient outcomes. Therefore, recall becomes the most critical metric. Based on this criterion, the Support Vector Machine (RBF) demonstrated outstanding performance with a recall of 97.62% on the test set. This suggests that it is most effective in identifying malignant cases correctly, making it a strong candidate for diagnostic applications where patient safety is paramount. Logistic Regression also maintained high recall (95.24%) while offering added benefits in terms of simplicity, stability, and interpretability. Importantly, the Linear Regression Classifier, despite achieving the same recall (95.24%), is not inherently a classifier. Its predictions are unbounded and do not represent probabilities, which can lead to unpredictable behavior on unseen or imbalanced datasets. While it may provide a lightweight alternative in controlled settings or for exploratory analysis, caution is required before applying it in clinical decision-making.

Overall Balanced Performance (Accuracy, Precision, F1-Score):

When the objective includes maintaining a balance between correctly identifying both malignant and benign cases, the SVM (RBF) again stands out. Its accuracy (99.12%), F1-score (98.80%), and perfect precision (100%) reflect both sensitivity and specificity. This is particularly important in systems where minimizing both false negatives and false positives is critical—such as automated triage or pre-screening tools used prior to physician review.

Model Interpretability and Transparency:

In real-world deployment, especially in regulated domains like healthcare, models must often be explainable. Black-box models, while performant, may face resistance in clinical adoption due to their opacity. Logistic Regression is widely accepted in clinical environments for precisely this reason; its coefficients can be directly interpreted to understand the effect of each feature on the predicted class. The Linear Regression Classifier maintains similar coefficient transparency, but lacks the probabilistic interpretation inherent to logistic models, and its use in classification is methodologically questionable.

In contrast, SVM and Random Forest are more difficult to interpret. SVMs with RBF kernels operate in high-dimensional feature spaces, making it non-trivial to trace input features to decision outcomes. Random Forests offer some interpretability through feature importance metrics, but the ensemble's multiple decision paths complicate reasoning about overall behavior.

Stability and Generalization Consistency:

A model's variability across different training subsets is a key indicator of reliability. In our cross-validation analysis, the Linear Regression Classifier exhibited some of the lowest standard deviations in AUC (0.0086) and RMSE (0.0902), suggesting stable behavior across folds. However, its methodological limitations restrict its practical relevance for classification. Logistic Regression followed closely, with minimal deviations in accuracy (0.0296) and precision (0.0176), reinforcing its robustness. SVM (RBF) maintained relatively low variance across most metrics, while K-Nearest Neighbors showed the highest

variability, particularly in recall (0.0941) and F1-score (0.0680), highlighting potential instability in production environments.

Computational Efficiency and Scalability:

From a deployment perspective, training time and scalability influence model choice. KNN, although simple to train, is computationally expensive during inference due to distance calculations to all training points. Random Forest can become slow to train with large numbers of trees or high-dimensional features, though it benefits from parallelization. In contrast, Logistic Regression is computationally lightweight and fast to train and deploy, as is the Linear Regression Classifier, making them suitable for real-time or embedded diagnostic systems. SVM, depending on dataset size and kernel complexity, can be computationally intensive, though acceptable for datasets of WDBC's size.

Domain Suitability and Flexibility:

Finally, the clinical context determines the most appropriate model. If interpretability is prioritized, Logistic Regression is recommended. If maximizing diagnostic accuracy and minimizing misclassification risk is paramount and computational resources are sufficient, SVM (RBF) should be selected. For rapid prototyping, benchmarking, or computationally constrained scenarios, KNN or Linear Regression may provide useful reference points, but the latter should be applied with caution due to its unsuitability as a classifier in real-world applications.

4.2 Limitations and Future Work:

Despite the methodological rigor and promising results demonstrated throughout this study, several limitations must be acknowledged to contextualize our findings and guide future research directions. Machine learning experiments, particularly in healthcare-related domains, must be critically assessed in terms of dataset characteristics, modeling assumptions, and deployment feasibility. While our models performed exceptionally well on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, their generalizability, scalability, and interpretability may face challenges when extended to real-world clinical settings or more complex datasets.

Dataset-Specific Limitations:

The most immediate limitation stems from our reliance on the WDBC dataset. While it is a widely used and well-curated benchmark, it has several characteristics that limit the scope of generalization. The dataset is relatively small (569 samples), balanced between benign and malignant classes, and collected under controlled conditions. In real-world clinical scenarios, datasets are often much larger, imbalanced, noisy, and collected from heterogeneous sources. As a result, models that perform well on WDBC may not retain the same accuracy or robustness when deployed in diverse hospital environments or across different demographics.

Moreover, the WDBC dataset only includes 30 numerical features derived from fine-needle aspirate (FNA) digitized images. These features, while informative, represent a narrow slice of possible clinical data. In practice, decision-making may involve additional variables such as patient history, radiographic imaging, genetic markers, or physician assessments. The absence of such multimodal data in our study means that our models do not account for the complexity of real-world diagnosis pipelines.

Modeling Assumptions and Limitations:

Several modeling assumptions also constrain our study. For instance, we assumed that each model's performance could be optimized using hyperparameter tuning via RandomizedSearchCV, with a fixed number of iterations and predefined parameter distributions. While this approach is widely accepted and computationally efficient, it does not guarantee the global optimum. More exhaustive or adaptive tuning methods, such as Bayesian optimization or population-based training, may yield better configurations and should be explored in future work.

Additionally, although we evaluated a broad spectrum of models—including linear, nonlinear, ensemble, instance-based, and even a custom threshold-based regressor—more advanced architectures were not considered. For example, neural networks, gradient boosting machines (e.g., XGBoost or LightGBM), or hybrid stacking ensembles could outperform our selected models on complex data distributions. Their exclusion here was a conscious decision

to focus on interpretable and broadly accessible models, but it limits the comprehensiveness of our performance benchmarks.

Interpretability vs. Complexity Trade-off:

Another key limitation lies in the trade-off between model complexity and interpretability. While models such as SVM and Random Forest provided excellent predictive performance, they are comparatively opaque. Their internal decision-making processes are difficult to interpret without additional techniques like SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), or feature attribution maps. In high-stakes medical applications, the lack of interpretability may be a barrier to clinical adoption, as medical professionals must often justify the rationale behind decisions.

While we highlighted this trade-off in our model selection guidelines, we did not integrate interpretability methods into the study due to scope constraints. Future work could incorporate these tools not only to gain insight into model behavior but also to validate whether the models are learning medically meaningful patterns or overfitting to spurious correlations.

Limitations in Evaluation Strategy:

Our evaluation framework, though thorough, could be further enhanced. While 10-fold stratified cross-validation reduces the likelihood of sampling bias and provides a robust estimate of generalization performance, it does not simulate time-dependent or real-world deployment scenarios. Prospective validation—where models are tested on entirely future or independent datasets—would better emulate real clinical decision-making.

Additionally, although our study considered multiple metrics, including accuracy, precision, recall, F1, AUC, and RMSE, it did not evaluate calibration metrics (e.g., Brier score or calibration curves). These are particularly relevant for probabilistic classifiers in medical domains, where not only classification but also confidence in predictions is critical. Miscalibrated models can be dangerous if their predicted probabilities are systematically biased—leading to under- or over-treatment based on false confidence.

Opportunities for Future Work:

Future research can address these limitations by expanding both the data and methodological dimensions of this study. One of the most promising directions involves validating our current models on external datasets from different populations, imaging modalities, or clinical contexts. Datasets such as Breast Cancer Wisconsin (Original) Dataset, The Cancer Genome Atlas (TCGA), or even hospital-specific EMR databases could provide richer testbeds for evaluation.

Exploration of more complex and high-capacity models, including convolutional neural networks (CNNs) for image inputs, recurrent neural networks (RNNs) for sequential data, or transformers for multi-modal inputs, could further improve diagnostic performance. Ensemble methods such as stacking or boosting may also yield improvements by leveraging the complementary strengths of different models.

Integrating explainability frameworks, as mentioned earlier, would enhance clinical trust and facilitate collaboration with domain experts. Visualizing which features contribute most to the prediction, or which cases are borderline or uncertain, would be invaluable in aligning model behavior with physician expectations.

Another avenue is the incorporation of cost-sensitive learning or fairness-aware modeling. In medical applications, misclassifications have asymmetric costs, and performance across demographic subgroups must be equitable. Future models could be optimized to minimize false negatives while maintaining fairness across age, gender, or ethnicity—a step toward responsible and ethical AI deployment.

Finally, the deployment environment itself warrants attention. Future work could involve simulating deployment in clinical decision support systems (CDSS), analyzing user interaction, and measuring real-world impact. Collaborations with healthcare providers could yield feedback loops that refine model behavior post-deployment and ensure long-term reliability.

5. REFERENCES

[1] A. Maulidia, L. Lidyawati, L. Jambola, and L. Kristiana, "Analysis of logistic regression algorithm for predicting types of breast cancer based on machine learning," in *AIP Conf. Proc.*, vol. 2772, p. 040005, 2023.

[2] Breast Cancer Wisconsin (Diagnostic) Data Set, UCI Machine Learning Repository, 2016. [Online]. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[3] F. T. Johra and M. M. H. Shuvo, "Detection of breast cancer from histopathology image and classifying benign and malignant state using fuzzy logic," in *Proc. 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEEICT)*, 2016, pp. 1–6.

[4] J. Harlan, "Analisis Regresi Logistik," *J. Chem. Inf. Model.*, vol. 53, no. 9, 2013.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York: Springer, 2013.

[6] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[7] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[8] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[9] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[10] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[11] R. J. Hyndman and A. B. Koehler, "Another Look at Measures of Forecast Accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, 2006.

[12] World Health Organization, "Breast Cancer Fact Sheet," *World Health Organization*, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[13] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates," *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021.

[14] L. A. Torre *et al.*, "Global cancer incidence and mortality rates and trends," *Cancer Epidemiol. Biomarkers Prev.*, vol. 25, no. 1, pp. 16–27, 2016.

[15] American Cancer Society, "Breast Cancer Risk Factors." [Online]. Available: https://www.cancer.org/cancer/breast-cancer/risk-and-prevention.html

[16] R. H. Johnson *et al.*, “Breast cancer in adolescents and young adults,” *J. Clin. Oncol.*, vol. 28, no. 32, pp. 4785–4793, 2010.

[17] C. E. DeSantis *et al.*, “Breast Cancer Statistics, 2022,” *CA Cancer J. Clin.*, vol. 72, no. 6, pp. 524–541, 2022.

[18] H. D. Nelson *et al.*, “Screening for breast cancer: a systematic review,” *JAMA*, vol. 314, no. 15, pp. 1615–1634, 2016.

[19] W. A. Berg *et al.*, “Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography,” *JAMA*, vol. 307, no. 13, pp. 1394–1404, 2012.

[20] National Cancer Institute, “SEER Cancer Stat Facts: Female Breast Cancer.” [Online]. Available: https://seer.cancer.gov/statfacts/html/breast.html

[21] S. M. McKinney *et al.*, “International evaluation of an AI system for breast cancer screening,” *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[22] A. F. Agarap, “On breast cancer detection: An application of machine learning algorithms on the WBCD dataset,” *arXiv preprint* arXiv:1712.07061, 2017.

[23] M. Karabatak, “A new classifier for breast cancer detection based on Naive Bayes combined with feature selection,” *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5829–5834, 2015.

[24] R. Ghasemi *et al.*, “Explainable AI in breast cancer diagnosis: A systematic review,” *Diagnostics*, vol. 14, no. 2, p. 143, 2024.

[25] M. Kasongo Dahouda and I. Joe, “A Deep-Learned Embedding Technique for Categorical Features Encoding,” *IEEE Access*, vol. 9, pp. 114381–114392, Aug.

[26] S. Roy and C. Chakrabarti, “Stratified Random Sampling for Power Estimation,” in *Proc. Int. Conf. Computer-Aided Design (ICCAD)*, Nov. 1996, pp. 577–582, doi: 10.1109/ICCAD.1996.569913.

[27] G. Patro and K. K. Sahu, “Normalization: A Preprocessing Stage,” *arXiv preprint arXiv:1503.06462*, Mar. 2015. Available: https://arxiv.org/pdf/1503.06462

[28] D. Dey *et al.*, “The proper application of logistic regression model in complex survey data: a systematic review,” *BMC Med. Res. Methodol.*, vol. 25, art. 15, 2025.

[29] Ramyasjoshi, “Breast Cancer Diagnosis Prediction using Logistic Regression,” *Medium*, 2024.

[30] “Breast Cancer Prediction Based on Multiple Machine Learning …,” *SAGE Journals*, 2024.

[31] J. Miah *et al.*, “Improving Cardiovascular Disease Prediction,” *arXiv preprint*, Nov. 2023.

[32] H. Pang *et al.*, “Electronic Health Records-Based Data-Driven Diabetes Knowledge,” *arXiv preprint*, Dec. 2024.

[33] C. Chakraborty and N. Mukherjee, “Bayesian Hybrid Machine Learning of Gallstone Risk,” *arXiv preprint*, Jun. 2025.

[34] M. Khan and R. S. Khan, “Linear regression and two-class classification with gene expression data,” *Journal of Biomedical Informatics*, vol. 37, no. 4, pp. 279–287, 2004.

[35] Z. Qin, M. Ding, and Y. Yang, “A classification algorithm based on linear regression and linear programming for predicting breast cancer,” *Soft Computing*, vol. 25, pp. 12745–12756, 2021.

[36] M. L. Wallace *et al.*, “Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction,” *BMC Med. Res. Methodol.*, vol. 23, art. 144, 2023.
[37] “Using random forest and biomarkers for differentiating COVID-19,” *Sci. Rep.*, 2024.
[38] D. Hartmann *et al.*, “Assessing the Role of Random Forests in Medical Image Segmentation,” *arXiv preprint*, 2021.

[39] K. B. Balsters *et al.*, “Random forest and SHAP classification of ASD based on rs-fMRI functional network features,” *Neuropsychopharmacology*, vol. 50, 2025.

[40] “Machine learning and SHAP value interpretation for predicting cardiovascular disease,” *Med. Data Anal.*, vol. 6, no. 1, 2024.

[41] M. Fascia, “Predictive Health Monitoring Using Random Forest and SHAP,” *SSRN Preprint*, 2025.

[42] M. A. S. Jilani, et al., “Breast Cancer Subtype Classification Using Feature-Optimized KNN,” *Computational Biology and Chemistry*, vol. 103, 2023.

[43] S. B. Khan, et al., “Skin Lesion Classification Using Feature Engineering and KNN,” *Biomedical Signal Processing and Control*, vol. 83, 2023.

[44] L. Deng, et al., “Early Alzheimer’s Detection Using Multi-Model KNN-Based Ensemble,” *Journal of Biomedical Informatics*, vol. 147, 2024.

[45] Y. Lin and H. Zhang, “K-Nearest Neighbors for Adaptive Patient Monitoring,” *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 1, pp. 92–100, 2024.

[46] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

[47] Recent breast cancer prediction studies using SVM-RBF (2023-2025), *Journal of Clinical Oncology Informatics*.

[48] Cardiovascular disease risk prediction with SVM: Smith et al., *Cardiology AI Journal*, 2024.

[49] Neurological disease classification using SVM: Chen et al., *Neuroinformatics*, 2025.

[50] Multi-modal heart failure prediction with SVM-RBF, *IEEE Transactions on Biomedical Engineering*, 2025.

[51] Early Alzheimer’s diagnosis via MRI and SVM: Zhao et al., *NeuroImage*, 2024.

[52] SVM in precision medicine applications: Review article, *Nature Medicine*, 2025.