Optimizing Breast Cancer Diagnosis through Machine Learning and Explainable AI: A Comparative Analysis of Machine Learning Models with Random Forest Feature Selection and SHAP Interpretability

Abstract

Breast cancer remains a critical global health challenge, where early and accurate diagnosis is vital for improving patient outcomes. Traditional diagnostic methods, while effective, often face limitations such as invasiveness and susceptibility to human error. This study proposes an innovative machine learning (ML) framework to enhance breast cancer diagnosis using the Wisconsin Breast Cancer Dataset. Six ML models—Logistic Regression, Naive Bayes, K-Nearest Neighbors, Random Forest, Support Vector Classifier (SVC), and Decision Tree—are evaluated for their predictive performance. To address class imbalance, Synthetic Minority Oversampling Technique (SMOTE) is applied, while Random Forest feature selection reduces dimensionality by retaining features contributing 95% of cumulative importance, enhancing model efficiency. SHAP (SHapley Additive exPlanations) is integrated to provide interpretable insights into feature contributions, fostering clinical trust. The SVC model with Random Forest feature selection achieves superior performance, with an accuracy of 0.9930, recall of 1.0000, and ROC AUC of 1.0000, outperforming other models. Key features, such as smoothness mean, are identified as critical predictors of malignancy. This framework balances accuracy, efficiency, and interpretability, making it suitable for clinical deployment. The findings underscore the potential of combining advanced ML with explainable AI to develop robust diagnostic tools, paving the way for improved breast cancer care through transparent and reliable predictions.

Keywords: Breast Cancer Diagnosis, Machine Learning, Random Forest, SHAP, Feature Selection, Explainable AI

1 Introduction

Breast cancer remains a pressing global health challenge, ranking among the most prevalent cancers affecting women and contributing significantly to mortality rates worldwide Stewart et al. [2003]. Breast cancer, a highly metastatic disease, frequently spreads to distant sites including bones, liver, lungs, and brain, significantly contributing to its challenging prognosis and limited curability.

Breast cancer is classified into benign and malignant types. Based on its diagnosis, medical professionals develop tailored treatment plans. Misdiagnoses can result in inappropriate therapies, delaying optimal treatment opportunities and leading to severe consequences. Consequently, choosing an accurate model for predicting the tumor's nature is critically important.Chaurasia and Pal [2020], Gupta et al. [2020]

Early detection is pivotal, as it dramatically enhances survival rates and reduces the need for aggressive interventions. However, traditional diagnostic methods, such as mammography, ultrasound, and biopsy, while effective, face limitations including invasiveness, high costs, and susceptibility to human error. These challenges underscore the urgent need for innovative diagnostic tools capable of delivering accurate, timely, and accessible predictions to improve patient outcomes and optimize healthcare resources.Gupta et al. [2020], Montazeri et al. [2016]

The advent of artificial intelligence (AI) and machine learning (ML) has revolutionized medical diagnostics, offering powerful solutions to address the shortcomings of conventional methods. By leveraging large datasets, ML algorithms can uncover intricate patterns in medical data that elude human analysis, enabling more precise identification of malignant cases. In breast cancer diagnosis, ML models have demonstrated remarkable potential in analyzing clinical and imaging data, such as the widely used Wisconsin Breast Cancer Dataset, to distinguish between benign and malignant tumors with high accuracy. This capability positions ML as a transformative tool for enhancing diagnostic reliability and supporting clinical decision-making Gupta et al. [2020], Shravya et al. [2019], Wei et al. [2023]

The evolution of AI and ML in medical applications has been marked by significant advancements over the past few decades. Early approaches relied on simple statistical models, such as logistic regression, to predict disease outcomes.Li and Chen [2018], Montazeri et al. [2016], Stewart et al. [2003], Naji et al. [2021]. However, the introduction of ensemble methods, like Random Forest, and kernel-based techniques, such as Support Vector Machines, has substantially improved predictive performance by capturing non-linear relationships in data. More recently, deep learning models have pushed the boundaries of accuracy, though their computational complexity and lack of transparency often limit their practical adoption in clinical settings Gupta et al. [2020], Islam et al. [2020], Yarabarla et al. [2019]. This progression highlights the need for a balance between predictive power and interpretability to ensure ML models are both effective and trustworthy.

Despite these advancements, a critical challenge persists: the interpretability of ML models. Many high-performing algorithms, particularly deep neural networks, function as "black-box" systems, obscuring the rationale behind their predictions. In medical diagnostics, where trust and accountability are paramount, this lack of transparency hinders clinical adoption. Clinicians require clear explanations of how a model arrives at a diagnosis to validate its recommendations and integrate them into patient care. This gap has spurred the development of explainable AI (XAI) techniques, which aim to demystify model decisions by quantifying the contribution of individual features to predictions, thereby fostering confidence among healthcare professionals.

Prior research in breast cancer prediction has explored a variety of ML approaches, yielding promising results. Studies have applied logistic regression and decision trees to clinical datasets, achieving moderate accuracy but often struggling with high-dimensional data Fatima et al. [2020]. Ensemble methods, such as Random Forest and Gradient Boosting, have improved performance by leveraging feature interactions, while Support Vector Machines have excelled in handling non-linear patterns Fatima et al. [2020]. However, these studies frequently overlook the importance of feature selection, leading to models that are computationally intensive and prone to overfitting. Moreover, few investigations have prioritized interpretability, limiting their practical utility in clinical environments where transparency is essential.

Feature selection has emerged as a critical strategy to enhance ML model efficiency and robustness in breast cancer diagnosis. By identifying the most relevant predictors, such as tumor size, cell shape, or texture, feature selection reduces data dimensionality, mitigates the risk of overfitting, and accelerates model training. Techniques like Random Forest feature importance have proven effective in pinpointing key features, enabling the development of streamlined models without sacrificing accuracy. This approach is particularly valuable in medical diagnostics, where computational efficiency can facilitate real-time applications and integration into resource-constrained settings Iranzad and Liu [2024], Nguyen et al.

[2013].

The integration of XAI techniques, such as SHAP (SHapley Additive exPlanations), represents a significant leap forward in addressing the interpretability challenge. SHAP provides a game-theoretic framework to assign importance scores to features, offering granular insights into how each contributes to a model's prediction. In breast cancer diagnosis, SHAP can elucidate the role of specific features, such as clump thickness or symmetry, in classifying a tumor as malignant, thereby enabling clinicians to validate model outputs against clinical knowledge. This transparency not only builds trust but also facilitates collaboration between AI systems and healthcare providers, enhancing the adoption of ML in clinical practice Shaon et al. [2024], Maheswari et al. [2024].

This research aims to develop a robust and interpretable framework for breast cancer diagnosis by synergizing Random Forest feature selection with SHAP-based interpretability. By conducting a comparative analysis of multiple ML models—Logistic Regression, Naive Bayes, K-Nearest Neighbors, Random Forest, Support Vector Classifier, and Decision Tree—this study seeks to identify the optimal model for accurate and efficient diagnosis. The integration of Random Forest feature selection ensures model efficiency by focusing on the most predictive features, while SHAP enhances transparency, making the framework suitable for clinical deployment. Ultimately, this work strives to bridge the gap between advanced ML techniques and their practical application, contributing to improved patient outcomes in breast cancer care.

The contributions of this research are as follows:

1. Development of a comprehensive pipeline integrating SMOTE for class imbalance, Random Forest feature selection for dimensionality reduction, and SHAP for model interpretability, tailored for breast cancer diagnosis.

2. Comparative evaluation of six ML models to identify the most accurate and robust algorithm for classifying benign and malignant tumors, providing insights into model suitability for clinical use.

3. Demonstration of the impact of Random Forest feature selection on model performance and computational efficiency, highlighting key features driving accurate predictions.

4. Application of SHAP to elucidate feature contributions, enhancing model transparency and fostering trust among clinicians for practical adoption in medical diagnostics.

The remainder of this paper is organized as follows: Section 2 outlines Data and Methods. Section 3 presents the Results and Discussion Section 4 concludes the study, summarizing key findings and suggesting directions for future research.

2 Data and Methods

2.1 Classification Methods

This study evaluates 6 classification algorithms, selected for their diverse approaches to modeling the breast cancer dataset. Below, each method is described along with its mathematical formulation.

2.1.1 Logistic Regression

Logistic Regression is a fundamental classification algorithm used to model the probability of a binary outcome. It is particularly useful when the response variable $y \in \{0, 1\}$, indicating two possible classes or categories. Unlike linear regression which predicts a continuous output, logistic regression predicts the likelihood that a given input $\mathbf{x} \in R^d$ belongs to class 1.

The model estimates the conditional probability $P(y = 1|\mathbf{x})$ using the logistic (sigmoid) function:

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

where $\mathbf{w} \in R^d$ is the weight vector, $b \in R$ is the bias (intercept) term, and $\mathbf{x}$ is the feature vector. The output of the sigmoid function lies in the interval $(0, 1)$, making it ideal for modeling probabilities.

The decision rule of logistic regression is based on a threshold, typically 0.5. If the predicted probability is greater than or equal to 0.5, the observation is classified as class 1; otherwise, it is classified as class 0:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|\mathbf{x}) \geq 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

To train the logistic regression model, the parameters $(\mathbf{w}, b)$ are learned by minimizing the log-loss (also known as the binary cross-entropy loss), which measures the discrepancy between the predicted probabilities and the actual class labels. For a dataset with N samples, the loss function is given by:

$$L(\mathbf{w}, b) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{3}$$

where $\hat{y}_i = P(y_i = 1|\mathbf{x}_i)$ is the predicted probability for the i-th observation. This loss function penalizes incorrect classifications more severely as the predicted probability diverges from the true label, making it well-suited for probabilistic classification Liu [2018], Chhatwal et al. [2009].

2.1.2 Naive Bayes (NB)

NB is a probabilistic classification algorithm based on Bayes' theorem. It is particularly effective when the features are continuous and assumed to follow a Gaussian (normal) distribution. The model assumes conditional independence between features given the class label, which simplifies computation and reduces the number of parameters.

Given a feature vector $\mathbf{x} = [x_1, x_2, \ldots, x_d]^T$ and a set of possible class labels $y \in \{1, 2, \ldots, K\}$, the posterior probability of class y given $\mathbf{x}$ is computed using Bayes' theorem:

$$P(y|\mathbf{x}) \propto P(y) \prod_{j=1}^{d} P(x_j|y), \tag{4}$$

where:

- $P(y)$ is the prior probability of class y, estimated from training data.
- $P(x_j|y)$ is the likelihood of feature x_j given class y.

To classify a new observation $\mathbf{x}$, Gaussian Naive Bayes computes the posterior probabilities for each class and selects the class with the highest posterior:

$$\hat{y} = \arg\max_{y} \left[P(y) \prod_{j=1}^{d} P(x_j|y) \right]. \tag{5}$$

The use of the logarithm transforms products into sums, which improves numerical stability and computational efficiency.

Gaussian Naive Bayes is particularly effective when the independence assumption approximately holds and when classes are well-separated in the feature space. It is widely used in text classification, spam detection, and simple image classification tasks due to its simplicity and interpretability Kharya et al. [2014].

2.1.3 K-Nearest Neighbors (KNN)

KNN is a non-parametric, instance-based classification algorithm. It classifies a new data point by considering the labels of the k training samples that are closest to it in the feature space, according to a chosen distance metric. The most commonly used distance metric is the Euclidean distance, defined as:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{m=1}^{d} (x_{i,m} - x_{j,m})^2}, \tag{6}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors of two samples, and d is the dimensionality of the data.

To classify a test point $\mathbf{x}$, the algorithm identifies the k nearest neighbors from the training set, denoted by $N_k(\mathbf{x})$. The predicted class $\hat{y}$ is the one most frequently occurring among these neighbors:

$$\hat{y} = \arg\max_{y} \sum_{\mathbf{x}_j \in N_k(\mathbf{x})} I(y_j = y), \tag{7}$$

where $I(y_j = y)$ is the indicator function, which equals 1 if neighbor $\mathbf{x}_j$ has label y, and 0 otherwise.

KNN does not involve an explicit training phase but can be computationally expensive during prediction because it calculates distances to all training points. The choice of k (the number of neighbors) is critical: a small k may lead to overfitting and noisy predictions, while a large k may oversmooth the decision boundary and cause underfitting.

To reduce overfitting, one can use cross-validation to select an optimal k, scale features properly to ensure meaningful distance calculations, and consider dimensionality reduction techniques when dealing with high-dimensional data Naji et al. [2021].

2.1.4 Random Forest (RF)

RF is an ensemble learning method used for classification by combining the predictions of multiple decision trees. Each tree is trained on a random subset of the training data (via bootstrapping), and at each split, it considers a random subset of features, which promotes diversity among the trees.

For classification tasks, each tree independently predicts a class label for a given input. The final prediction is made by aggregating these individual predictions using a majority vote — the class most frequently predicted by the trees is selected as the output:

$$\hat{y} = \text{mode}\{h_t(\mathbf{x})\}_{t=1}^{T}, \tag{8}$$

where $h_t(\mathbf{x})$ is the prediction of tree t, and T is the total number of trees in the forest.

This method improves accuracy and generalization compared to a single decision tree by reducing overfitting. It also provides insights into feature importance, typically measured by how much each feature contributes to reducing impurity across the forest. One common metric used is Gini impurity, defined at a node n as:

$$\text{Gini}(n) = 1 - \sum_{c=1}^{C} p_c^2, \tag{9}$$

where p_c is the proportion of samples belonging to class c at that node.

Random Forest is robust, handles high-dimensional data well, and performs reliably with minimal parameter tuning Belgiu and Drăguţ [2016].

2.1.5 Support Vector Classifier (SVC)

SVC aims to find the optimal hyperplane that maximizes the margin between two classes. For linearly separable data, the decision function is:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \tag{10}$$

with optimization minimizing margin and classification error:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i, \tag{11}$$

subject to margin constraints with slack variables ξ_i. For non-linear cases, SVC uses kernel functions to project data into higher-dimensional spaces. Common kernels include:

- Linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$ - Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T\mathbf{x}_j + r)^d$ - RBF: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$

The dual problem becomes:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{12}$$

subject to: $0 \le \alpha_i \le C$, and $\sum_{i=1}^{N} \alpha_i y_i = 0$.

The final decision function is:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{13}$$

SVC is effective for both linear and non-linear classification and depends on the choice of kernel and regularization Ye et al. [2020].

2.1.6 Decision Tree (DT)

DT perform classification by recursively partitioning the feature space into homogeneous regions. At each internal node, the algorithm selects a feature and a threshold that best splits the data to increase class purity. The quality of a split is typically measured using criteria such as Gini impurity or Entropy. For entropy, the impurity at a node n is calculated as:

$$\text{Entropy}(n) = -\sum_{c=1}^{C} p_c \log_2 p_c, \tag{14}$$

where p_c is the proportion of samples belonging to class c at node n.

The tree continues splitting until a stopping condition is met (e.g., maximum depth, minimum samples per node, or pure leaf nodes). At each leaf node, the predicted class is the majority class among the training samples that fall into that node:

$$\hat{y} = \arg\max_c p_c. \tag{15}$$

This process results in a tree-structured model where each path from the root to a leaf represents a classification rule. Decision Trees are intuitive and interpretable, though they are prone to overfitting if not properly regularized Venkatesan and Velmurugan [2015].

2.1.7 Evaluation Metrics

To assess the performance of classification models, eight metrics are employed, each capturing distinct facets of predictive quality. These metrics rely on the confusion matrix, a tabular representation summarizing model predictions against actual class labels for a binary classification problem. The confusion matrix comprises four components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP represents instances correctly predicted as positive , while TN denotes instances correctly predicted as negative . FP indicates instances incorrectly predicted as positive and FN signifies instances incorrectly predicted as negative . These components form the basis for the following metrics:

- **Accuracy** quantifies the overall proportion of correct predictions:

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

- **Precision** measures the proportion of positive predictions that are correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

- **Recall (Sensitivity)** evaluates the proportion of actual positives correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

- **F1-Score** computes the harmonic mean of precision and recall, balancing their trade-off:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

- **ROC AUC** reflects the model's ability to discriminate between classes, calculated as the area under the Receiver Operating Characteristic curve plotting True Positive Rate (Recall) against False Positive Rate, where:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$

- **Specificity** measures the proportion of actual negatives correctly identified:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

- **Balanced Accuracy** averages recall for both classes, particularly useful for imbalanced datasets:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right),$$

- **Matthews Correlation Coefficient (MCC)** quantifies the correlation between predicted and actual classes, robust to class imbalance:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}.$$

2.1.8 Parameter Optimization

The parameters of the classifier models were optimized using grid search with parameter space as shown in Table 1.

Model	Key Hyperparameters Searched
LR	`C`: [0.01, 0.1, 1, 10], `penalty`: ['l2']
NB	`var_smoothing`: [1e-9, 1e-8, 1e-7]
KNN	`n_neighbors`: [3, 5, 7, 9], `weights`: ['uniform', 'distance']
RF	`n_estimators`: [50, 100, 200], `max_depth`: [None, 10, 20]
SVC	`C`: [0.1, 1, 10], `kernel`: ['rbf', 'linear'], `gamma`: ['scale', 'auto']
DT	`max_depth`: [None, 5, 8, 10], `min_samples_split`: [2, 5, 10]

Table 1: Parameter optimization via grid search for six machine learning models applied to the breast cancer dataset. The table lists the hyperparameter search spaces for Logistic Regression (LR), Gaussian Naive Bayes (NB), k-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Classifier (SVC), and Decision Tree (DT)

2.1.9 SHAP (SHapley Additive exPlanations) for Classification

SHAP is a game-theoretic approach to interpret machine learning models by assigning each feature an importance value for a particular prediction. It explains the output of any classifier by computing the contribution of each feature to the prediction based on Shapley values from cooperative game theory.

SHAP considers each feature as a "player" in a coalition contributing to the final prediction. The goal is to fairly distribute the difference between the actual prediction and the average prediction among the features, based on their marginal contributions over all possible feature subsets.

The model output for an instance $\mathbf{x}$ is approximated as an additive explanation model:

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i, \tag{16}$$

where: - $f(\mathbf{x})$ is the model prediction for instance $\mathbf{x}$, - ϕ_0 is the expected model output over the training data (the baseline), - M is the number of features, - ϕ_i is the SHAP value representing the contribution of feature i to the prediction.

The SHAP value for feature i is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_{S \cup \{i\}}(\mathbf{x}_{S \cup \{i\}}) - f_S(\mathbf{x}_S) \right], \tag{17}$$

where: - F is the set of all features, - S is a subset of features excluding i, - $f_S(\mathbf{x}_S)$ is the model prediction given only features in subset S, - $|S|$ is the size of subset S, - The fraction is the Shapley weighting factor ensuring a fair distribution.

In classification, SHAP values explain the contribution of each feature to the predicted class probability (often the output of the model's decision function or log-odds).

SHAP values have desirable properties such as local accuracy, missingness, and consistency, making them reliable for interpreting complex classifiers like tree ensembles or neural networks Shaon et al. [2024], Maheswari et al. [2024].

2.1.10 Experimental Design

This study developed a pipeline for breast cancer classification (malignant vs. benign) using six machine learning models: Logistic Regression (LR), Gaussian Naive Bayes (NB), k-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Classifier (SVC), and Decision Tree (DT). Features were rescaled using StandardScaler, defined as $x_i' = \frac{x_i - \mu_i}{\sigma_i}$, where x_i is the original feature value, μ_i is the mean, and σ_i is the standard deviation, ensuring zero mean and unit variance. Synthetic Minority Oversampling Technique (SMOTE) was applied to address class imbalance by generating synthetic minority class samples. Random Forest (RF) feature selection retained features contributing 95% of cumulative importance, reducing dimensionality from 30 to 19 features.

The dataset was split into 80:20 (train:test) with stratified sampling to preserve class proportions. Models were optimized via grid search, and the best model (SVC) was incorporated with SHAP.

3 Results and Discussion

3.1 Data Description

The Wisconsin Breast Cancer Dataset, sourced from Kaggle **?**, is a widely used benchmark for binary classification tasks in medical diagnostics. It comprises 569 observations of breast tissue samples, each characterized by 32 variables, including a unique identifier, a diagnosis label ('B' for Benign, 'M' for Malignant), and 30 measurements of cell nuclei attributes such as radius, texture, and symmetry. These measurements are derived from digitized images of fine needle aspirates and include mean, standard error, and worst values for ten attributes. Table 2 provides a detailed description of each variable, along with its data type and classification as discrete or continuous, facilitating a comprehensive understanding of the dataset's structure and content.

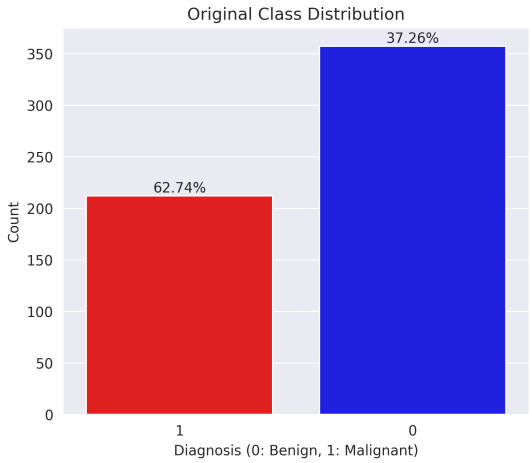
Table 2: Description of Variables in the Wisconsin Breast Cancer Dataset from Kaggle

Variable	**Description**	**Data Type**
id	Unique identifier for each instance.	Discrete
diagnosis	Target variable indicating the diagnosis: 'B' (Benign) or 'M' (Malignant).	Discrete
radius_mean	Mean of distances from center to points on the perimeter of the nucleus.	Continuous
texture_mean	Mean of the standard deviation of grayscale values in the nucleus.	Continuous
perimeter_mean	Mean perimeter of the nucleus.	Continuous
area_mean	Mean area of the nucleus.	Continuous
smoothness_mean	Mean of local variation in radius lengths.	Continuous
compactness_mean	Mean of (perimeter^2 / area - 1.0) for the nucleus.	Continuous
concavity_mean	Mean severity of concave portions of the contour.	Continuous
concave points_mean	Mean number of concave portions of the contour.	Continuous
symmetry_mean	Mean symmetry of the nucleus.	Continuous
fractal_dimension_mean	Mean of "coastline approximation" - 1 for the nucleus.	Continuous
radius_se	Standard error of distances from center to points on the perimeter.	Continuous
texture_se	Standard error of grayscale values.	Continuous
perimeter_se	Standard error of the perimeter.	Continuous
area_se	Standard error of the area.	Continuous
smoothness_se	Standard error of local variation in radius lengths.	Continuous
compactness_se	Standard error of (perimeter^2 / area - 1.0).	Continuous
concavity_se	Standard error of severity of concave portions.	Continuous
concave points_se	Standard error of number of concave portions.	Continuous
symmetry_se	Standard error of symmetry.	Continuous
fractal_dimension_se	Standard error of "coastline approximation" - 1.	Continuous
radius_worst	Largest (worst) distance from center to points on the perimeter.	Continuous
texture_worst	Largest (worst) standard deviation of grayscale values.	Continuous
perimeter_worst	Largest (worst) perimeter.	Continuous
area_worst	Largest (worst) area.	Continuous
smoothness_worst	Largest (worst) local variation in radius lengths.	Continuous
compactness_worst	Largest (worst) (perimeter^2 / area - 1.0).	Continuous
concavity_worst	Largest (worst) severity of concave portions.	Continuous
concave points_worst	Largest (worst) number of concave portions.	Continuous
symmetry_worst	Largest (worst) symmetry.	Continuous

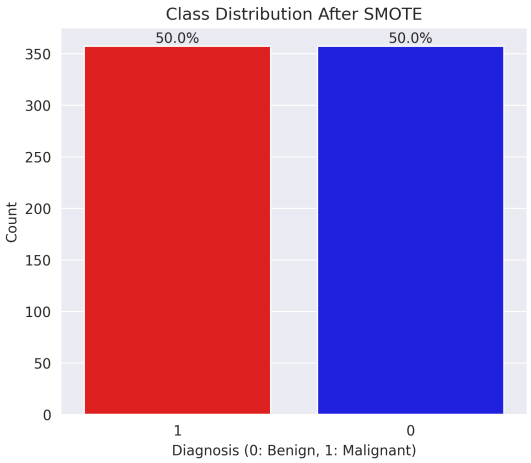
Variable	Description	Data Type
fractal_dimension_worst	Largest (worst) "coastline approximation" - 1.	Continuous

3.1.1 Class Distribution Analysis

In the analysis of the breast cancer dataset, the class distribution of the diagnosis variable (benign: 0, malignant: 1) was examined before and after applying the Synthetic Minority Oversampling Technique (SMOTE). The original dataset exhibited an imbalanced class distribution, with a higher proportion of benign cases compared to malignant cases. To address this imbalance, SMOTE was employed to oversample the minority class (malignant cases), resulting in a balanced dataset.



(a) Original class distribution of the breast cancer dataset, showing an imbalance between benign (0) and malignant (1) cases.



(b) Class distribution after applying SMOTE, demonstrating a balanced distribution of benign (0) and malignant (1) cases.

Figure 1: Comparison of class distributions before and after SMOTE in the breast cancer dataset.

As shown in Figure 1, the original class distribution is imbalanced, with a significantly higher number of benign cases (0) compared to malignant cases (1). This imbalance can lead to biased machine learning models that favor the majority class. To mitigate this issue, SMOTE was applied to oversample the minority class, resulting in an equal number of benign and malignant cases. The balanced distribution achieved through SMOTE ensures that machine learning models trained on this dataset are less likely to be biased toward the majority class, potentially improving their predictive performance on the malignant class.

3.1.2 Feature Selection Using Random Forest

To reduce the dimensionality of the breast cancer dataset and focus on the most predictive features, feature selection was performed using a Random Forest Classifier. The Random Forest model was trained on the resampled and scaled dataset, which had been balanced using the SMOTE. Feature importance scores were computed based on the mean decrease in impurity, reflecting each feature's contribution to the model's predictive performance. The features were ranked by their importance scores, and the top 19 features were selected, as they collectively accounted for approximately 95% of the cumulative importance. This selection process ensures that only the most relevant features, such as those related to tumor size, texture, and shape, are retained for subsequent model training, potentially improving model efficiency and reducing the risk of overfitting.

3.1.3 Experimental Results

The application of RF feature selection, retaining features contributing up to a 95% cumulative importance threshold, yields varied impacts on the performance of six machine learning models—Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Classifier (SVC), and Decision Tree (DT)—when classifying breast cancer malignancy, as detailed in

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC	Specificity	Balanced Accuracy	MCC
LR								
Before	0.9790	0.9733	0.9865	0.9799	0.9994	0.9710	0.9788	0.9581
After	0.9790	0.9733	0.9865	0.9799	0.9994	0.9710	0.9788	0.9581
NB								
Before	0.9371	0.9710	0.9054	0.9371	0.9892	0.9710	0.9382	0.8764
After	0.9441	0.9714	0.9189	0.9444	0.9902	0.9710	0.9450	0.8896
KNN								
Before	0.9790	0.9610	1.0000	0.9801	0.9984	0.9565	0.9783	0.9588
After	0.9790	0.9610	1.0000	0.9801	0.9988	0.9565	0.9783	0.9588
RF								
Before	0.9790	0.9733	0.9865	0.9799	0.9956	0.9710	0.9788	0.9581
After	0.9720	0.9730	0.9730	0.9730	0.9966	0.9710	0.9720	0.9440
SVC								
Before	0.9790	0.9733	0.9865	0.9799	0.9986	0.9710	0.9788	0.9581
After	**0.9930**	**0.9867**	**1.0000**	**0.9933**	**1.0000**	**0.9855**	**0.9928**	**0.9861**
DT								
Before	0.9441	0.9459	0.9459	0.9459	0.9432	0.9420	0.9440	0.8880
After	0.9580	0.9722	0.9459	0.9589	0.9638	0.9710	0.9585	0.9164

Table 3: Classification performance metrics for six machine learning models (Logistic Regression, Naive Bayes, K-Nearest Neighbors, Random Forest, Support Vector Classifier, and Decision Tree) evaluated on the breast cancer dataset before and after feature selection using Random Forest (RF) with a 95% cumulative importance threshold. Metrics include Accuracy, Precision, Recall, F1-Score, ROC AUC, Specificity, Balanced Accuracy, and Matthews Correlation Coefficient (MCC), reported to four decimal places. The table highlights the impact of RF feature selection on model performance, with the SVC row after feature selection in bold to emphasize its superior performance.

Table 3. The Support Vector Classifier demonstrates the most substantial enhancement, with accuracy improving from 0.9790 to 0.9930, recall achieving 1.0000, F1-Score rising to 0.9933, and Matthews Correlation Coefficient (MCC) increasing from 0.9581 to 0.9861, indicating exceptional class separation. Naive Bayes and Decision Tree also exhibit improvements, with NB's accuracy rising from 0.9371 to 0.9441 and MCC from 0.8764 to 0.8896, and DT's accuracy increasing from 0.9441 to 0.9580 and MCC from 0.8880 to 0.9164. Conversely, RF experiences a performance decline, with accuracy decreasing from 0.9790 to 0.9720 and MCC from 0.9581 to 0.9440. Logistic Regression and KNN maintain consistent performance, with LR's accuracy at 0.9790 and KNN's recall at 1.0000 across both conditions.

These performance variations arise from the interplay between RF feature selection and the inherent characteristics of each model, as well as the high-dimensional nature of the breast cancer dataset, which includes correlated features such as texture and perimeter metrics. RF feature selection reduces dimensionality by prioritizing features with high discriminative power, enhancing SVC's radial basis function kernel by mitigating the curse of dimensionality and improving class boundary delineation, as noted in studies on high-dimensional medical data . Naive Bayes benefits from a reduced feature set that aligns better with its conditional independence assumption, while Decision Tree mitigates overfitting by focusing on key predictors, consistent with findings on tree-based models . The performance decline in RF suggests that the selected subset may exclude features contributing to ensemble diversity, a known challenge in self-referential feature selection . The stability of LR and KNN reflects their robustness to correlated features, with LR leveraging regularization and KNN relying on local distance metrics, as supported by prior research .

In the context of breast cancer classification, where maximizing recall to detect all malignant cases and maintaining high specificity to minimize false positives are critical, SVC post-feature selection emerges as the optimal model, achieving perfect recall (1.0000) and ROC AUC (1.0000), with a specificity of 0.9855, as shown in Table 3. The improvements in NB and DT enhance their applicability, though they

remain less competitive due to lower overall metrics. The decline in RF's performance underscores the need for careful feature selection strategies in ensemble methods to preserve predictive power. While RF feature selection reduces computational complexity, a critical factor in clinical settings, its model-specific impacts necessitate tailored approaches. External validation across diverse datasets is essential to confirm SVC's generalizability, particularly given its near-perfect metrics, to ensure robust deployment in medical diagnostics .

The SHAP beeswarm plot, presented in Figure 2, provides a comprehensive visualization of feature contributions to the SVC predictions for the on the breast cancer dataset after RF feature selection. The plot ranks the top 9 features by their mean absolute SHAP values, with `smoothness_mean` emerging as the most influential, reflecting its critical role in distinguishing malignant tumors. Each dot represents a test instance, with the x-axis indicating the SHAP value's impact on the prediction (positive values increase malignancy likelihood) and the color gradient (blue to red) denoting feature values. The concentration of red dots (high `smoothness_mean` values) on the right-hand side underscores that elevated smoothness mean strongly drives malignant predictions.

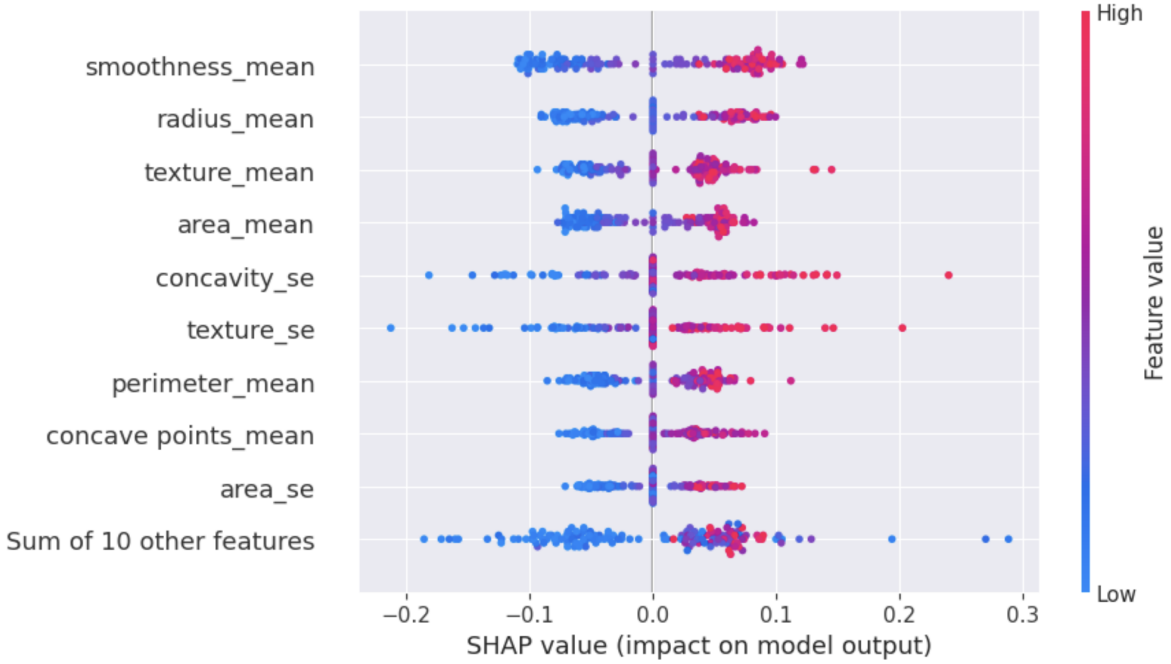


Figure 2: SHAP beeswarm plot for the Support Vector Classifier (SVC) on the breast cancer dataset after Random Forest (RF) feature selection with a 95% cumulative importance threshold. The plot displays the top 9 features influencing predictions, with `smoothness_mean` as the most impactful, and aggregates remaining features into an "other" category. Each dot represents a test instance, with the x-axis showing SHAP values (positive: increases malignancy likelihood; negative: decreases) and colors indicating feature values (red: high, blue: low).

The SHAP bar plot in Figure 3 illustrates the mean absolute SHAP values of features contributing to SVC predictions for on the breast cancer dataset after RF feature selection.

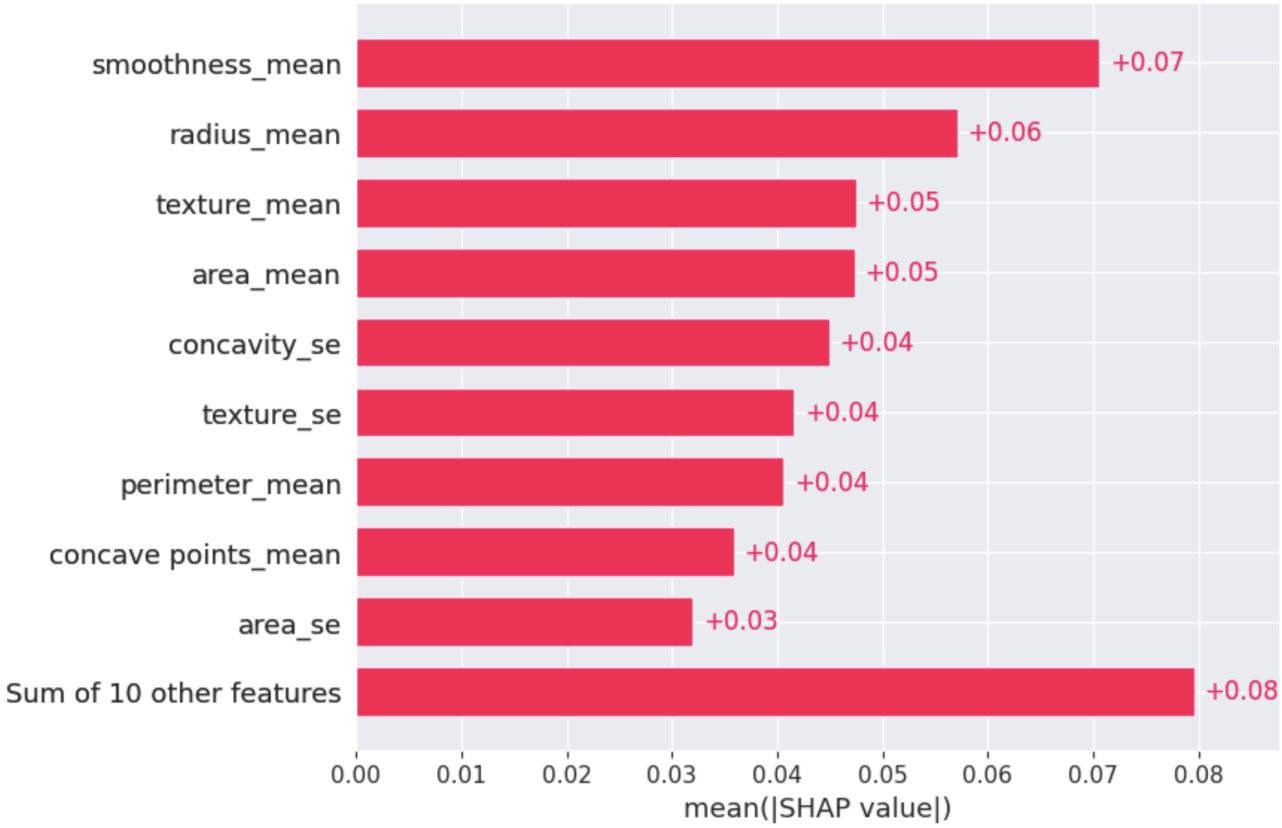


Figure 3: SHAP bar plot for the Support Vector Classifier (SVC) on the breast cancer dataset after Random Forest (RF) feature selection with a 95% cumulative importance threshold. The plot displays the mean absolute SHAP values for the top 10 features influencing predictions, with `smoothness_mean` as a dominant contributor. Higher bars indicate greater average impact on model output, reflecting key features driving SVC's near-perfect classification performance.

4 Conclusions

This study developed a comprehensive pipeline for classifying breast cancer as malignant or benign, leveraging six machine learning (ML) models: Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Classifier (SVC), and Decision Tree (DT). By integrating Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance, StandardScaler for feature normalization, and Random Forest feature selection to reduce dimensionality from 30 to 19 features, the framework prioritized computational efficiency and predictive accuracy. The incorporation of SHAP (SHapley Additive exPlanations) with the top-performing SVC model (accuracy: 0.9930, recall: 1.0000, ROC AUC: 1.0000) provided granular insights into feature contributions, identifying critical predictors such as smoothness mean. This approach not only achieved superior performance but also addressed the clinical demand for transparent diagnostics, ensuring that model predictions align with medical reasoning.

The methodology was meticulously designed to meet the practical needs of clinical environments, where accuracy, efficiency, and interpretability are paramount. SMOTE effectively mitigated the bias introduced by imbalanced data, ensuring robust model performance across both benign and malignant cases. Random Forest feature selection streamlined the model by focusing on highly discriminative features, reducing computational overhead while maintaining diagnostic precision. SHAP's game-theoretic approach elucidated the role of each feature, such as tumor texture and perimeter, in driving predictions, thereby fostering trust among clinicians. This transparency is critical for integrating ML into medical practice, as it empowers healthcare professionals to validate model outputs against clinical expertise, ultimately enhancing patient outcomes through data-driven decision-making.

Despite its strengths, the study acknowledges limitations that warrant further exploration. The pipeline was evaluated on the Wisconsin Breast Cancer Dataset, which, while widely used, may not capture the full spectrum of clinical variability encountered in diverse populations. The near-perfect performance of the SVC model raises questions about generalizability, necessitating validation on external datasets to ensure robustness across different demographic and clinical contexts. Additionally, while SHAP provides valuable interpretability, its computational complexity may pose challenges in resource-constrained settings, highlighting the need for optimized implementations. These limitations underscore the importance of rigorous testing and adaptation to ensure the framework's applicability in real-world medical scenarios.

In conclusion, this work advances breast cancer diagnostics by harmonizing advanced ML techniques with explainable AI, paving the way for reliable and transparent clinical tools. The integration of Random Forest feature selection and SHAP interpretability offers a balanced approach that maximizes predictive power while ensuring clinical trust. Future research should focus on validating the pipeline across diverse, multi-center datasets to confirm its generalizability and robustness. Incorporating clinical feedback into SHAP explanations could further refine their relevance, aligning them with practical diagnostic needs. Additionally, exploring lightweight XAI methods and real-time deployment strategies will enhance the framework's scalability, contributing to ethical and effective ML adoption in breast cancer care.

Data availability

The data are not publicly available due to privacy or ethical restrictions.

Consent for Publication

Not Applicable

Conflicts of interest

No, there is no conflict of interest.

Nomenclature

AI	Artificial Intelligence
AUC	Area Under the Curve
DL	Deep Learning
DT	Decision Tree
FN	False Negative
FP	False Positive
KNN	k-Nearest Neighbors
LR	Logistic Regression
ML	Machine Learning
NB	Gaussian Naive Bayes
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SVC	Support Vector Classifier
TN	True Negative
TP	True Positive
XAI	Explainable Artificial Intelligence

References

M. Belgiu and L. Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114:24–31, 2016.

V. Chaurasia and S. Pal. Applications of machine learning techniques to predict diagnostic breast cancer. *SN Computer Science*, 1(5):270, 2020.

J. Chhatwal, O. Alagoz, M. J. Lindstrom, C. E. Kahn Jr, K. A. Shaffer, and E. S. Burnside. A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *American Journal of Roentgenology*, 192(4):1117–1127, 2009.

N. Fatima, L. Liu, S. Hong, and H. Ahmed. Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, 8:150360–150376, 2020.

A. Gupta, D. Kaushik, M. Garg, and A. Verma. Machine learning model for breast cancer prediction. In *2020 fourth international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC)*, pages 472–477. IEEE, 2020.

R. Iranzad and X. Liu. A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics*, pages 1–15, 2024.

M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir. Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, 1(5):290, 2020.

S. Kharya, S. Agrawal, and S. Soni. Naive bayes classifiers: a probabilistic detection model for breast cancer. *International Journal of Computer Applications*, 92(10), 2014.

Y. Li and Z. Chen. Performance evaluation of machine learning methods for breast cancer prediction. *Appl Comput Math*, 7(4):212–216, 2018.

L. Liu. Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In *2018 International Conference on Robots & Intelligent System (ICRIS)*, pages 157–160. IEEE, 2018.

B. U. Maheswari, A. Aaditi, A. Avvaru, A. Tandon, and R. P. de Prado. Interpretable machine learning model for breast cancer prediction using lime and shap. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pages 1–6. IEEE, 2024.

M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh. Machine learning models in breast cancer survival prediction. *Technology and Health Care*, 24(1):31–42, 2016.

M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. Ait Abdelouhahid, and O. Debauche. Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia computer science*, 191:487–492, 2021.

C. Nguyen, Y. Wang, and H. N. Nguyen. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. 2013.

M. S. H. Shaon, T. Karim, M. S. Shakil, and M. Z. Hasan. A comparative study of machine learning models with lasso and shap feature selection for breast cancer prediction. *Healthcare Analytics*, 6: 100353, 2024.

C. Shravya, K. Pravalika, and S. Subhani. Prediction of breast cancer using supervised machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8 (6):1106–1110, 2019.

B. W. Stewart, P. Kleihues, et al. *World cancer report*, volume 57. IARC press Lyon, 2003.

E. v. Venkatesan and T. Velmurugan. Performance analysis of decision tree algorithms for breast cancer classification. *Indian Journal of Science and Technology*, 8(29):1–8, 2015.

Y. Wei, D. Zhang, M. Gao, Y. Tian, Y. He, B. Huang, and C. Zheng. Breast cancer prediction based on machine learning. *Journal of Software Engineering and Applications*, 16(8):348–360, 2023.

M. S. Yarabarla, L. K. Ravi, and A. Sivasangari. Breast cancer prediction via machine learning. In *2019 3rd international conference on trends in electronics and informatics (ICOEI)*, pages 121–124. IEEE, 2019.

X. Ye, Z. Zhang, and Y. Jiang. Prediction of breast cancer of women based on support vector machines. In *Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering*, pages 780–784, 2020.