Explainable AI for Breast Cancer Diagnosis: Comparative Analysis of ML Models Using Random Forest Feature Selection and SHAP Interpretability

Abstract

Breast cancer diagnosis is critical for improving patient outcomes, yet traditional methods face limitations such as invasiveness and human error. This study presents an explainable AI framework for breast cancer classification using six ML models: LR, NB, KNN, RF, SVC, and DT. SMOTE addresses class imbalance, while RF feature selection reduces dimensionality from 30 to 19 features. SHAP interpretability is integrated to provide clinical insights into feature contributions, enhancing trust in model predictions. The SVC model with RF-selected features achieves superior performance, with an accuracy of 0.9930 and recall of 1.0000, highlighting the importance of features such as smoothness mean. This framework balances accuracy, efficiency, and transparency, offering a foundation for clinical deployment and guiding future work on external validation and broader adoption of explainable ML in breast cancer care.

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

1 Introduction

Breast cancer is one of the most common cancers worldwide and remains a leading cause of cancer-related mortality among women Stewart et al. [2003]. According to global health statistics, it accounts for millions of new cases and hundreds of thousands of deaths annually, placing a significant burden on healthcare systems and societies. Its metastatic nature, often spreading to the bones, liver, lungs, and brain, makes treatment complex and reduces long-term survival rates. These challenges underscore the need for accurate and timely diagnosis to improve patient outcomes.

Clinically, breast cancer is categorized into benign and malignant types. Accurate classification plays a critical role in determining appropriate treatment pathways. Misclassification not only delays effective treatment but can also lead to unnecessary procedures, patient anxiety, and increased healthcare costs Chaurasia and Pal [2020], Gupta et al. [2020]. While early detection substantially improves survival, conventional diagnostic tools such as mammography, ultrasound, and biopsy face several drawbacks. They can be invasive, costly, time-consuming, and are often subject to human error and variability in interpretation Gupta et al. [2020], Montazeri et al. [2016]. These limitations motivate the development of alternative, data-driven diagnostic approaches that are accurate, reliable, and scalable.

Artificial intelligence (AI) and machine learning (ML) have emerged as powerful technologies capable of addressing these limitations Mutinda and Langat [2024], Mutinda and Geletu [2025], Chugh et al. [2021]. By learning patterns from large clinical datasets, ML algorithms can assist in identifying subtle differences between benign and malignant tumors that may not be easily detected by human experts. In breast cancer diagnosis, ML approaches have demonstrated strong potential in analyzing datasets such as the Wisconsin Breast Cancer Dataset (WBCD), providing high accuracy and reproducibility in tumor classification Gupta et al. [2020], Shravya et al. [2019], Wei et al. [2023]. Early efforts primarily relied on statistical models such as logistic regression Li and Chen [2018], Montazeri et al. [2016], which offered interpretability but were limited in handling complex, non-linear data. Subsequently, ensemble learning methods like Random Forest and kernel-based models such as Support Vector Classifier (SVC) advanced predictive performance by capturing more intricate patterns. More recently, deep learning models have further improved classification accuracy but often require large datasets, substantial computational resources, and are hindered by a lack of interpretability Gupta et al. [2020], Islam et al. [2020], Yarabarla et al. [2019]. This progression highlights an ongoing trade-off between predictive power and transparency.

Interpretability is particularly crucial in medical applications, where decisions must be justified and trusted by clinicians. Many high-performing ML models act as "black boxes," making it unclear how predictions are generated. This lack of transparency can reduce clinical adoption, as healthcare professionals must understand the reasoning behind diagnostic recommendations to integrate them into patient care. Explainable AI (XAI) techniques, such as SHAP (SHapley Additive exPlanations), address this challenge by attributing importance scores to features, thereby providing interpretable explanations that align model outputs with clinical knowledge Shaon et al. [2024], Maheswari et al. [2024]. In the context of breast cancer, SHAP can reveal how tumor-related features—such as smoothness, texture, or cell symmetry—influence classification, enabling clinicians to validate predictions.

Another critical aspect of building robust diagnostic models is feature selection. High-dimensional datasets often contain redundant or irrelevant variables that reduce model efficiency and increase the risk of overfitting. Feature selection techniques help isolate the most informative predictors, leading to simpler, faster, and more generalizable models. Random Forest feature importance has proven particularly effective for this task, as it identifies variables that contribute most to classification performance Iranzad and Liu [2024], Nguyen et al. [2013]. In medical settings, efficient models are especially valuable, as they allow integration into real-time diagnostic workflows and resource-constrained environments.

Motivated by these considerations, this study proposes a breast cancer diagnostic framework that integrates Random Forest feature selection with SHAP-based interpretability. Using the WBCD, we evaluate six ML models—Logistic Regression (LR), Naive Bayes(NB), K-Nearest Neighbors(KNN), Random Forest(RF), Support Vector Classifier (SVC), and Decision Tree (DT)—to compare predictive performance. The framework addresses key challenges by balancing accuracy, computational efficiency, and transparency. Random Forest feature selection ensures dimensionality reduction, while SHAP provides interpretable explanations of model predictions, enabling trust and practical applicability in clinical settings.

The main contributions of this study are summarized as follows:

1. Development of a diagnostic pipeline integrating SMOTE for addressing class imbalance, Random Forest feature selection for dimensionality reduction, and SHAP for interpretability.

2. Comprehensive comparison of six ML models to identify the most accurate and reliable algorithm for classifying breast tumors.

3. Demonstration of how Random Forest feature selection improves predictive accuracy and computational efficiency while reducing the risk of overfitting.

4. Application of SHAP to elucidate the contribution of individual features, enhancing transparency and fostering clinical trust in ML-assisted diagnostics.

The remainder of this paper is organized as follows: Section 2 describes the dataset and methodology. Section 3 presents the experimental results and discussion. Section 4 concludes with key findings and outlines 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], Assegie [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 Kabiraj et al. [2020].

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], Safia [2023], Minnoor and Baths [2023].

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: $0 \leq \alpha_i \leq C$, 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], Sharma et al. [2025].

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], Hazra et al. [2020].

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.

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 Bifarin [2023], Xia et al. [2024].

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 ML models: LR, NB, KNN, RF, SVC, and 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. To address class imbalance, SMOTE was applied to generate synthetic samples of the minority class. RF-based feature selection was then performed, retaining features contributing 95% of cumulative importance, thereby reducing dimensionality from 30 to 19 features.

The dataset was split into an 80:20 ratio (train:test) with stratified sampling to preserve class proportions. Hyperparameter optimization was carried out using grid search for each model. The best-performing model, SVC, was further analyzed using SHAP to provide interpretability.

Figure 1 presents the flowchart of the proposed prediction framework.

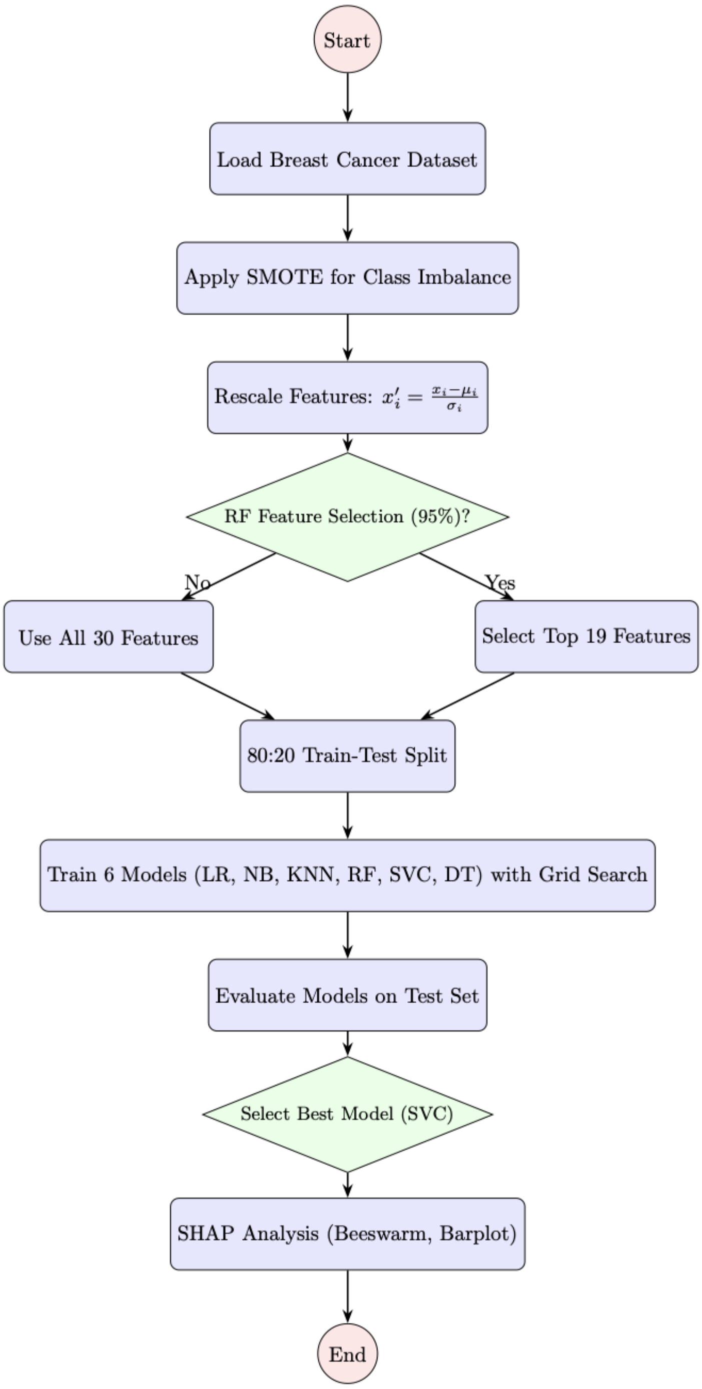

Start
Load Breast Cancer Dataset
Apply SMOTE for Class Imbalance
Rescale Features: $x_i' = \frac{x_i - \mu_i}{\sigma_i}$
RF Feature Selection (95%)?
No
Yes
Use All 30 Features
Select Top 19 Features
80:20 Train-Test Split
Train 6 Models (LR, NB, KNN, RF, SVC, DT) with Grid Search
Evaluate Models on Test Set
Select Best Model (SVC)
SHAP Analysis (Beeswarm, Barplot)
End

Figure 1: The flowchart illustrates the experimental design for breast cancer classification, detailing data preprocessing, feature selection, model training, and evaluation. The pipeline highlights the application of SMOTE, Random Forest (RF) feature selection, and Support Vector Classifier (SVC) with SHAP analysis.

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
fractal_dimension_worst	Largest (worst) "coastline approximation" - 1.	Continuous

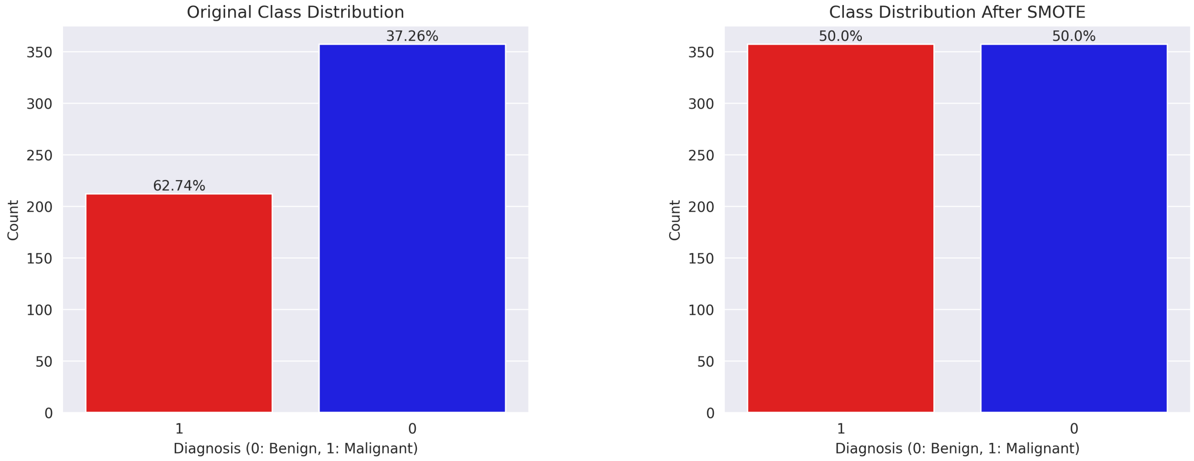
Feature	Mean	Std Dev	Min	Max	Skewness	Kurtosis	Median	25th Pctl	75th Pctl
radius mean	14.127	3.524	6.981	28.110	0.940	0.828	13.370	11.700	15.780
texture mean	19.290	4.301	9.710	39.280	0.649	0.741	18.840	16.170	21.800
perimeter mean	91.969	24.299	43.790	188.500	0.988	0.953	86.240	75.170	104.100
area mean	654.889	351.914	143.500	2501.000	1.641	3.610	551.100	420.300	782.700
smoothness mean	0.096	0.014	0.053	0.163	0.455	0.838	0.096	0.086	0.105
compactness mean	0.104	0.053	0.019	0.345	1.187	1.625	0.093	0.065	0.130
concavity mean	0.089	0.080	0.000	0.427	1.397	1.971	0.062	0.030	0.131
concave points mean	0.049	0.039	0.000	0.201	1.168	1.047	0.034	0.020	0.074
symmetry mean	0.181	0.027	0.106	0.304	0.724	1.266	0.179	0.162	0.196
fractal dimension mean	0.063	0.007	0.050	0.097	1.301	2.969	0.062	0.058	0.066
radius se	0.405	0.277	0.112	2.873	3.080	17.521	0.324	0.232	0.479
texture se	1.217	0.552	0.360	4.885	1.642	5.292	1.108	0.834	1.474
perimeter se	2.866	2.022	0.757	21.980	3.435	21.204	2.287	1.606	3.357
area se	40.337	45.491	6.802	542.200	5.433	48.767	24.530	17.850	45.190
smoothness se	0.007	0.003	0.002	0.031	2.308	10.368	0.006	0.005	0.008
compactness se	0.025	0.018	0.002	0.135	1.897	5.051	0.020	0.013	0.032
concavity se	0.032	0.030	0.000	0.396	5.097	48.423	0.026	0.015	0.042
concave points se	0.012	0.006	0.000	0.053	1.441	5.071	0.011	0.008	0.015
symmetry se	0.021	0.008	0.008	0.079	2.189	7.816	0.019	0.015	0.023
fractal dimension se	0.004	0.003	0.001	0.030	3.914	26.040	0.003	0.002	0.005
radius worst	16.269	4.833	7.930	36.040	1.100	0.925	14.970	13.010	18.790
texture worst	25.677	6.146	12.020	49.540	0.497	0.212	25.410	21.080	29.720
perimeter worst	107.261	33.603	50.410	251.200	1.125	1.050	97.660	84.110	125.400
area worst	880.583	569.357	185.200	4254.000	1.854	4.347	686.500	515.300	1084.000
smoothness worst	0.132	0.023	0.071	0.223	0.414	0.503	0.131	0.117	0.146
compactness worst	0.254	0.157	0.027	1.058	1.470	3.002	0.212	0.147	0.339
concavity worst	0.272	0.209	0.000	1.252	1.147	1.591	0.227	0.114	0.383
concave points worst	0.115	0.066	0.000	0.291	0.491	-0.541	0.100	0.065	0.161
symmetry worst	0.290	0.062	0.156	0.664	1.430	4.395	0.282	0.250	0.318
fractal dimension worst	0.084	0.018	0.055	0.208	1.658	5.188	0.080	0.071	0.092

Table 3: Descriptive Statistics of Selected Features

Table 3 show the summary statistics of all variables. These features, derived from digitized images of fine needle aspirates, include measurements such as radius, texture, and symmetry, each characterized by mean, standard error, and worst values. The table reports the mean, standard deviation, minimum, maximum, skewness, kurtosis, median, and the 25th and 75th percentiles for each feature, thereby providing insights into their distribution and variability.

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

As shown in Figure 2, 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.

Figure 3 show Random Forest Feature Importance plot.

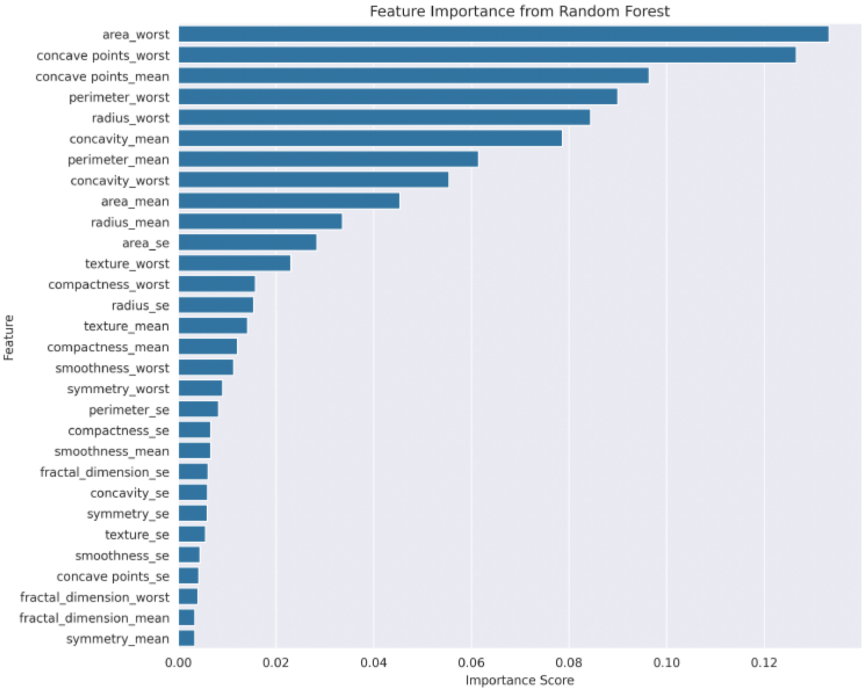


Figure 3: Feature importance scores derived from a Random Forest Classifier trained on the resampled and scaled breast cancer dataset. The bar plot displays the importance of each feature, with higher scores indicating greater contributions to the model's ability to distinguish between benign and malignant cases. The top 19 features, selected based on their cumulative importance, are used for subsequent analysis and model training.

3.1.3 Experimental Results

TThe application of RF feature selection, retaining features contributing up to a 95% cumulative importance threshold, yields varied impacts on the performance of six ML models—LR, NB, KNN, RF,

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 4: Classification performance metrics for six machine learning models. 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.

SVC, and DT—when classifying breast cancer malignancy, as detailed in Table 4. SVC demonstrates the most substantial improvement, with accuracy increasing from 0.9790 to 0.9930, recall reaching 1.0000, F1-Score rising to 0.9933, and MCC improving from 0.9581 to 0.9861, indicating exceptional class separation. NB and DT also show gains, with NB's accuracy increasing from 0.9371 to 0.9441 and MCC from 0.8764 to 0.8896, and DT's accuracy rising from 0.9441 to 0.9580 and MCC from 0.8880 to 0.9164. Conversely, RF experiences a slight decline, with accuracy decreasing from 0.9790 to 0.9720 and MCC from 0.9581 to 0.9440. LR and KNN maintain consistent performance, with LR's accuracy at 0.9790 and KNN's recall at 1.0000 across both conditions.

Figure 4 presents a barplot comparing classification performance of the six ML models before and after RF feature selection. Notably, SVC achieves near-perfect performance post-selection, demonstrating the effectiveness of RF-based feature reduction in enhancing model accuracy and reliability.

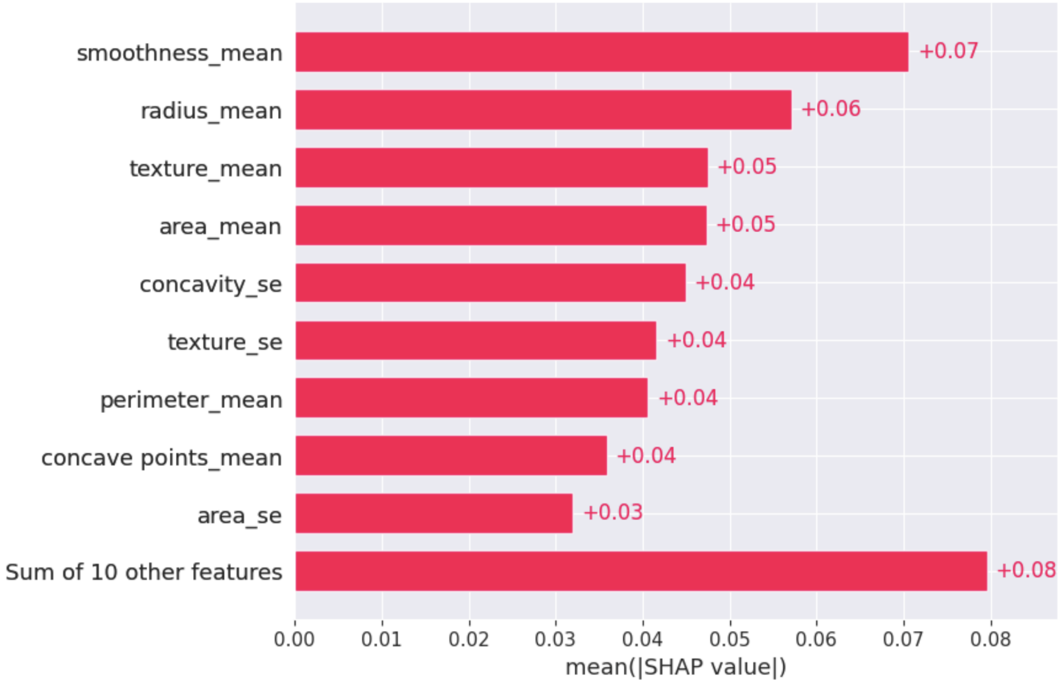
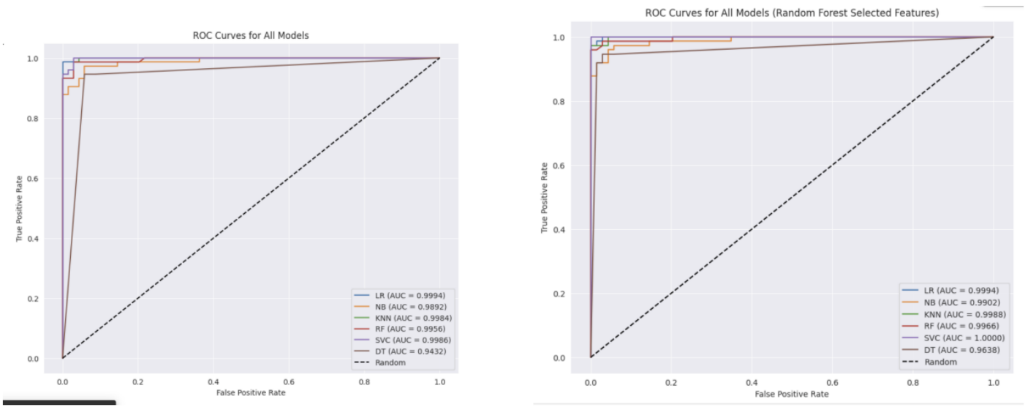


Figure 4: Barplot comparing classification metrics (Accuracy, Precision, Recall, F1-Score, ROC AUC, Specificity, Balanced Accuracy, and MCC) for six machine learning models

Figure 5 presents the ROC curves comparing the discriminative performance of six machine learning models on the breast cancer dataset.



(a) Before RF Feature Selection

(b) After RF Feature Selection

Figure 5: ROC curves comparing the discriminative performance of six machine learning models on the breast cancer dataset before and after feature selection.

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 4. 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 6, 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.

The SHAP bar plot in Figure 7 illustrates the mean absolute SHAP values of features contributing to SVC predictions for on the breast cancer dataset after RF feature selection.

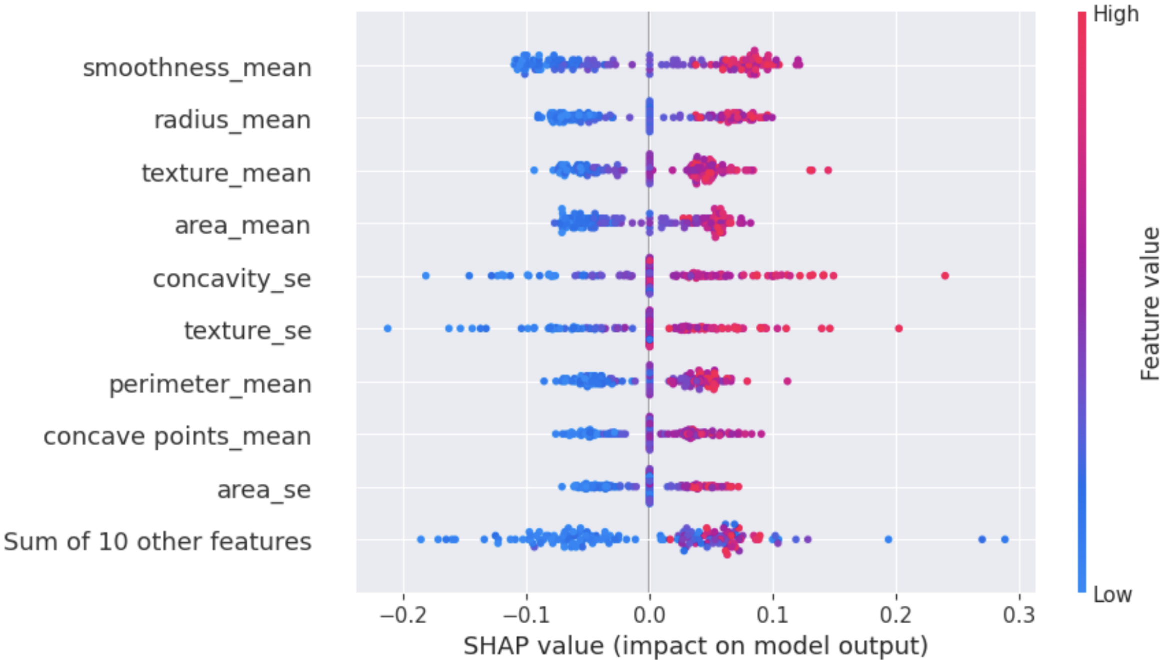


Figure 6: SHAP beeswarm plot for the SVC on the breast cancer dataset after RF feature selection with a 95% cumulative importance threshold. The plot displays the top 9 features influencing predictions, with `smoothness_mean` as the most impactful, and aggregates remaining features into an “other” category. 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).

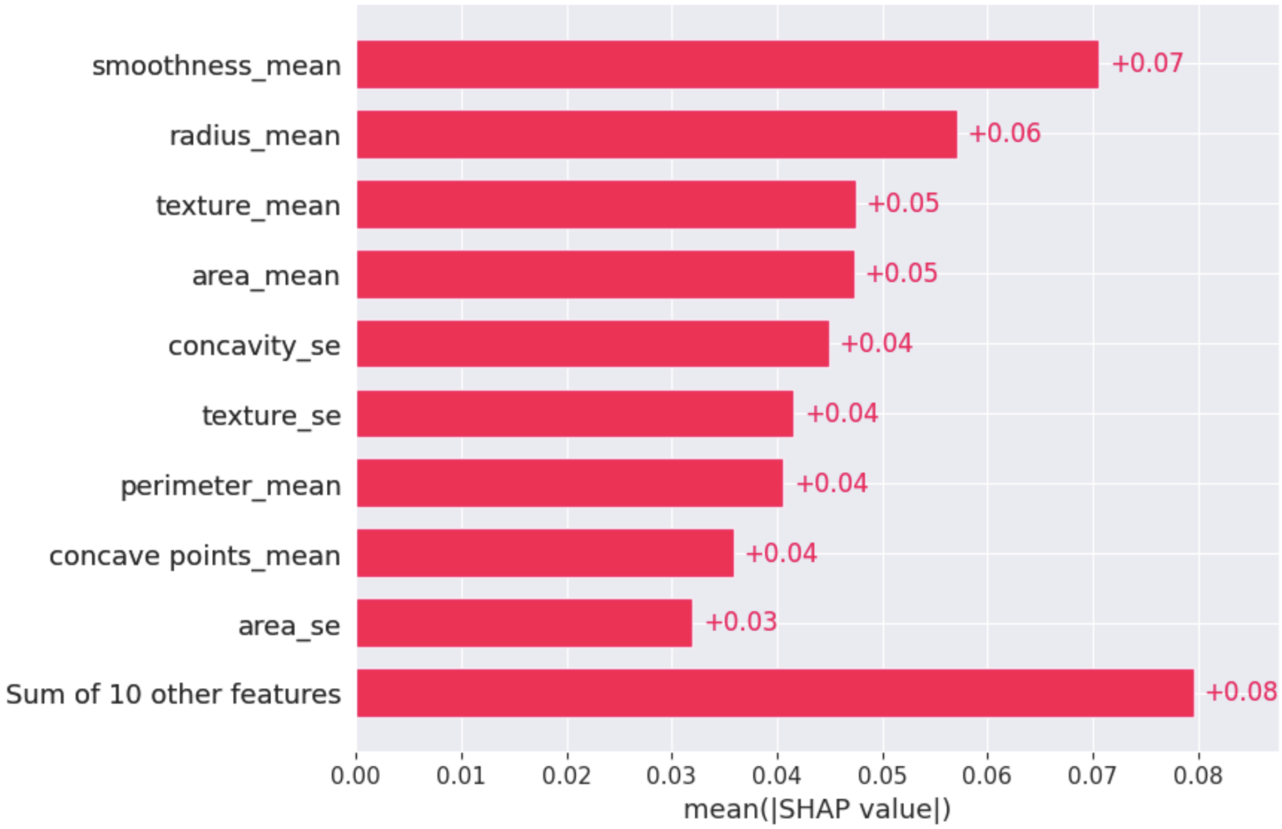


Figure 7: SHAP bar plot for the SVC on the breast cancer dataset after RF feature selection with a 95% cumulative importance threshold. The plot displays the mean absolute SHAP values for the top 10 features influencing predictions, with `smoothness_mean` as a dominant contributor. Higher bars indicate 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 ML models: LR, NB, KNN, RF, SVC, and DT. By integrating SMOTE for class imbalance, StandardScaler for feature normalization, and RF feature selection to reduce dimensionality from 30 to 19 features, the framework prioritized computational efficiency and predictive accuracy. The incorporation of SHAP 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 achieved superior performance while addressing the clinical need for transparent diagnostics, ensuring model predictions align with medical reasoning.

The methodology was designed to meet practical clinical needs, where accuracy, efficiency, and interpretability are essential. SMOTE mitigated bias from imbalanced data, ensuring robust performance across both benign and malignant cases. RF feature selection streamlined the model by focusing on highly discriminative features, reducing computational overhead without compromising diagnostic precision. SHAP's game-theoretic approach elucidated the influence of features such as tumor texture and perimeter, fostering trust among clinicians and supporting data-driven decision-making.

Despite its strengths, the study has several limitations. The pipeline was evaluated on the Wisconsin Breast Cancer Dataset, which may not capture the full spectrum of clinical variability. The near-perfect performance of SVC suggests the potential for overfitting, highlighting the need for cross-validation and testing on larger, multi-center datasets to ensure generalizability. While SHAP provides interpretability, its computational cost may be challenging in resource-constrained environments. Future work should also explore deep learning architectures such as CNN, TCN, and Transformers, as well as ensemble models, to further enhance predictive performance and robustness.

In conclusion, this work advances breast cancer diagnostics by combining ML techniques with explainable AI, offering a balanced approach that maximizes accuracy while ensuring clinical trust. Future research should focus on validating the pipeline across diverse datasets, integrating clinical feedback into SHAP explanations, and investigating lightweight XAI and real-time deployment strategies to improve scalability and practical adoption in breast cancer care.

Author Contributions

John Kamwele Mutinda: Conceptualization, Methodology, Software, Data Curation, Writing – Original Draft, Writing – Review
Tecla Mutave Kyalo: Methodology, Software, Data Curation, Writing – Original Draft, Writing – Review
Joyce Akhalakwa Mukolwe: Methodology, Software, Data Curation, Writing – Original Draft, Writing – Review
Jackson Ndoto Munyao: Methodology, Software, Data Curation, Writing – Original Draft, Writing – Review
Millicent Auma Omondi: Methodology, Software, Data Curation, Writing – Original Draft, Writing – Review
Wycliffe Nzoli Nzomo: Methodology, Software, Data Curation, Writing – Original Draft, Writing – Review
Titus Mutua Kioko: Methodology, Data Curation, Writing – Review
David Chepkonga: Methodology, Data Curation, Writing – Review
Samuel Kipsang Kaptum: Methodology, Data Curation, Writing – Review
Erick Munala Sifuna: Methodology, Data Curation, Writing – Review
Amos Kipkorir Langat: Supervision, Writing – Review

Disclaimer (Artificial intelligence)
Option 1:
Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.
Option 2:
Author(s) hereby declare that generative AI technologies such as Large Language Models, etc. have been used during the writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology
Details of the AI usage are given below:
1.
2.
3.

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

T. A. Assegie. An optimized k-nearest neighbor based breast cancer detection. *Journal of Robotics and Control (JRC)*, 2(3):115–118, 2021.

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.

O. O. Bifarin. Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. *Plos one*, 18(5):e0284315, 2023.

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.

G. Chugh, S. Kumar, and N. Singh. Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cognitive Computation*, 13(6):1451–1470, 2021.

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. Hazra, M. Banerjee, and L. Badia. Machine learning for breast cancer classification with ann and decision tree. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0522–0527. 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. Kabiraj, M. Raihan, N. Alvi, M. Afrin, L. Akter, S. A. Sohagi, and E. Podder. Breast cancer risk prediction using xgboost and random forest algorithm. In *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–4. IEEE, 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. Minnoor and V. Baths. Diagnosis of breast cancer using random forests. *Procedia Computer Science*, 218:429–437, 2023.

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.

J. K. Mutinda and A. Geletu. Stock market index prediction using ceemdan-lstm-bpnn-decomposition ensemble model. *Journal of Applied Mathematics*, 2025(1):7706431, 2025.

J. K. Mutinda and A. K. Langat. Exploring the role of dimensionality reduction in enhancing machine learning algorithm performance. *Asian Journal of Research in Computer Science*, 17(5):157–166, 2024.

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.

N. Safia. Prediction of breast cancer through random forest. *Current Medical Imaging*, 19(10):1144–1155, 2023.

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.

S. Sharma, S. Rani, and V. Sumalatha. Robust classification of breast cancer with support vector machine. In *2025 International Conference on Intelligent Control, Computing and Communications (IC3)*, pages 337–341. IEEE, 2025.

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.

H. Xia, X. Li, J. Pang, J. Liu, K. Ren, and L. Xiong. P-shapley: Shapley values on probabilistic classifiers. *Proceedings of the VLDB Endowment*, 17(7):1737–1750, 2024.

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.