JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 03, Issue No. 02, 2026

ARTICLE TYPE: Research Article

Article ID: JEPCS-2026-0302-0046

A Comparative Analysis of Machine Learning Algorithms for Diagnostic Breast Cancer Classification: Logistic Regression, Support Vector Machine and Random Forest

Ismayilov Togrul Atabay[1,*]

[1] Department of Digital Technologies and Exact Sciences, Azerbaijan State Agricultural University, Ganja, Azerbaijan.

*Corresponding author: togrul.ismayilov@adau.edu.az

Contributing authors: togrul.ismayilov@adau.edu.az

Received: September 15, 2026 **Accepted:** October 21, 2026

Abstract- The choice of a classification algorithm directly affects the practical value of data-driven decision systems. This study compares three algorithms logistic regression (LR), a radial-basis-function support vector machine (SVM) and random forest (RF) on the Wisconsin Diagnostic Breast Cancer dataset (569 samples, 30 features). The study follows the CRISP-DM methodology as evaluation uses complexed 5-fold cross-validation repeated 10 times (50 folds), and differences are tested with the Nadeau-Bengio corrected resampled t-test. Mean accuracy was 97.89% for LR, 97.56% for SVM and 95.92% for RF, yet the corrected test did not confirm a significant difference for any pair (smallest $p = 0.085$). RF was about 94 times slower than SVM in training and about 10 times slower in prediction. The results indicate that when the classes are close to linearly separable, simple models can match ensemble models, and that tests accounting for the dependence between cross-validation folds are essential when comparing classifiers.

Keywords- Machine learning; classification; logistic regression; support vector machine; cross-validation

1. Introduction

Supervised machine learning is used to support decision-making across a wide range of fields, from medical diagnosis to financial risk assessment [1]. In these applications, the main practical question is often framed as "Which algorithm performs best?" However, the answer depends on the size of the dataset, the structure of the features, the evaluation protocol, and the computational costs. Extracting quantitative measurements of cell nuclei from digitized images of fine-needle aspiration (FNA) samples of breast lesions and predicting whether a lesion is benign or malignant based on these measurements is a topic extensively studied in machine learning literature [2], [3]. The resulting Wisconsin Diagnostic Breast Cancer dataset is one of the most widely used test datasets for comparing algorithms, owing to its open availability and small size [4]. However, in many such comparisons, results are presented based on a single train/test split or as an average across folds. The statistical significance of differences is not assessed, and computational costs are often not reported at all. With small datasets, random variability between folds can exceed the differences between algorithms, rendering claims about the "best" algorithm unstable. The aim of this study is to determine whether there are statistically significant differences among three well-known classifiers-LR, SVM, and RF in terms of overall accuracy, sensitivity for the malignant class, and F1-score, under identical conditions. The study objectives are: (1) to characterize the dataset through exploratory analysis; (2) to evaluate the three models using the same repeated cross-validation protocol and test for differences using a method that accounts for correlation; and (3) to analyze error structure, computational costs, sensitivity to hyperparameters, and learning curves.

The remainder of the article is organized as follows: Section 2 reviews related work, Section 3 describes the materials and methods (dataset, data preparation, and the three classification algorithms), and Section 4 presents the experimental setup and evaluation protocol. Section 5 presents and discusses the results, while Section 6 outlines the conclusions and directions for future research.

2. Related Work

In this section, literature relevant to the research topic is reviewed across three areas.

2.1. FNA-Based Diagnosis of Breast Cancer

Wolberg and Mangasarian demonstrated the feasibility of automating FNA-based diagnosis by applying a linear programming-based multisurface separation method to breast cytology features [3]. Meanwhile, Street et al. proposed an approach to automatically extract cell nucleus features from digital images and this work forms the basis of a dataset widely used today [2]. These studies focus on feature extraction and diagnostic capability but do not entail a systematic comparison of modern algorithms.

2.2. General Comparisons of Classification Algorithms

Fernández-Delgado et al. compared 179 classifiers across 121 datasets and reported that, on average, the random forest family yielded the best results, with Gaussian kernel SVMs performing closely behind [8]. Caruana and Niculescu-Mizil compared supervised learning algorithms for binary classification tasks using multiple metrics and demonstrated that ensemble methods-particularly tree-based ensembles-were generally superior [9]. At the same time, Wolpert's "no free lunch" results indicate that an average ranking across all problems does not guarantee the same ranking for a specific dataset [10].

Recent breast-cancer machine-learning research also emphasizes that strong internal discrimination is not sufficient for clinical translation. Contemporary studies increasingly separate benchmark evaluation from clinical validation and emphasize external validation and calibration when models are intended for clinical use. These considerations motivate the deliberately limited clinical interpretation adopted in this study [20], [21].

2.3. Statistical Evaluation Methodology

Demšar recommended the Wilcoxon signed-rank test for comparing two classifiers across multiple datasets, and the Friedman test for comparing multiple classifiers [11]. However, within a single dataset, training sets overlap significantly across repeated cross-validation folds, meaning the folds are not independent. Nadeau and Bengio proposed a corrected resampled t-test to account for this dependency [12]. Bouckaert and Frank investigated the reproducibility of such tests and recommended combining repeated cross-validation with the corrected test [13]. The evaluation in this study is based on this body of literature.

Thus, existing studies have focused either on the general ranking of algorithms or on statistical methodology. This research combines these approaches using a specific, widely used medical dataset, evaluating three algorithms under the same protocol in terms of accuracy, error structure, computational costs, and hyperparameter sensitivity.

3. Materials and Methods

The research is structured according to the CRISP-DM methodology, which describes the lifecycle of data-driven projects [14]. Sections 3 and 4 are organized according to its phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

3.1. Understanding the Problem

The problem is a binary classification task: predicting whether a tumor is malignant or benign based on nuclear features. Since the cost of errors is asymmetric, sensitivity and the F1-score for the malignant class are monitored as key metrics alongside overall accuracy.

3.2. Wisconsin Diagnostic Breast Cancer Dataset

The dataset consists of 569 samples, and each sample contains 30 real-valued features calculated from an FNA image. The features are based on ten characteristics of cell nuclei (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension), with three statistics provided for each mean, standard error, and "worst" (the mean of the three largest values) [2], [4]. The target variable comprises two classes: malignant (212) and benign (357).

The key characteristics of the dataset are presented in Table 1. The data contains no missing values or duplicate rows. The class ratio is approximately 37:63 that is, the "naive" baseline accuracy based on the most frequent class is 62.7%.

Table 1: *Characteristics of the dataset.*

Parameter	Value
Source	UCI Machine Learning Repository, Breast Cancer Wisconsin (Diagnostic) [4]
Number of samples	569
Number of features	30 (10 measurements × 3 statistics: mean, standard error, "worst")
Classes	Malignant: 212 (37.3%); benign: 357 (62.7%)
Missing values / duplicates	0 / 0
Samples with $\|z\| > 3$ in at least one feature	74 (13.0%)
Baseline accuracy (most frequent class)	62.7%

3.3. Data Preparation

No imputation was performed, as there were no missing values. Features for LR and SVM were standardized using the z-score method to prevent data leakage, the standardization parameters were learned solely from the training set at each fold (within the pipeline). No additional preprocessing was applied for RF, as it is not sensitive to scaling. Samples with $|z| > 3$ were retained as manifestations of natural variability, as they did not indicate measurement errors. Highly correlated features were not removed, in order to ensure a fair comparison.

3.4. Logistic Regression

Logistic regression models the log-odds of class membership as a linear combination of features. The L2-regularized variant stabilizes weights in the presence of highly correlated features, allows for the interpretation of coefficients, and is computationally efficient [5]. Scaling the features to a common range is required for the regularization to function effectively.

3.5. Support Vector Machine

The Support Vector Machine is based on the principle of finding the boundary between classes using the widest separating strip (maximum margin). The kernel function enables the learning of non-linear boundaries [6]. The radial basis kernel is

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right).$$

where the two hyperparameters are the penalty coefficient C and γ, which relates to the kernel width. Since the kernel is distance-based, the SVM is sensitive to feature scaling.

3.6. Random Forest

Random Forest is an ensemble consisting of multiple decision trees trained on bootstrap samples and utilizing a random subset of features at each split. The prediction is determined by a majority vote of the trees [7]. The method is insensitive to scaling and allows for the assessment of relative feature importance. However, training and prediction costs increase as the number of trees grows.

3.7. Modeling

The configurations of the three models are shown in Table 2. In the initial comparison, hyperparameters were kept at fixed, generally accepted standard values, while the impact of hyperparameter tuning was investigated separately (Section 5.6). The algorithms were implemented using the scikit-learn library [15].

Table 2: *Model configurations.*

Model	Configuration	Preprocessing
LR	L2 penalty, $C = 1$, `lbfgs` solver, `max_iter = 1000`	StandardScaler
SVM	RBF kernel, $C = 1$, $\gamma =$ "scale"	StandardScaler
RF	300 trees, Gini criterion, `max_features = "sqrt"`, bootstrap	None

3.8. Evaluation Metrics

The malignant class was treated as the "positive" class because, from a clinical perspective, missing a malignant tumor (false negative) has more severe consequences than classifying a benign tumor as malignant (false positive). The metrics are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = 2\,\frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{4}$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Additionally, the Area Under the ROC Curve (ROC-AUC), which is independent of the classification threshold, was used.

4. Experimental Setup

4.1. Evaluation Protocol

The primary evaluation was conducted using 5-fold stratified cross-validation repeated 10 times (totaling 50 training/test pairs random_state = 42) [16]. For each fold, overall accuracy, precision, recall, and F1-score for the malignant class, as well as ROC-AUC, were calculated. Differences between models were assessed based on pairwise differences (d) across the same folds. The primary test used was the Nadeau-Bengio corrected repeated-measures t-test [12]:

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right) s_d^2}} \tag{5}$$

where k = 50, n_2/n_1 = 1/4, and the number of degrees of freedom is k − 1 = 49. Here, $\bar{d}$ and s_d^2 are the mean and sample variance of the fold differences, respectively, while n_1 and n_2 are the sizes of the training and test sets. The results of the Wilcoxon signed-rank test are also provided for comparison [17]. However, due to the dependence between the folds, the p-values from this test are considered anti-conservative (tending to indicate significance where there is none).

Additional analyses were performed for descriptive or robustness purposes and are kept distinct from the primary repeated 5-fold comparison: (i) confusion matrices and ROC curves based on a separate 10-fold cross-validation prediction set, used only to summarize error structure and discrimination, (ii) measurement of training and prediction times on a stratified split consisting of 426 training samples and 143 test samples (20 training and 200 prediction repetitions), (iii) sensitivity analysis regarding the (C, gamma) grid for SVM and the number of trees for the random forest, (iv) nested 5×5-fold cross-validation, with five outer folds for unbiased performance estimation and five inner folds within each outer training partition for hyperparameter selection, (v) learning curves and (vi) feature importance for the random forest.

4.2. Experimental Environment

All calculations were performed in the environment shown in Table 3. Since the duration measurements were obtained in a single-core virtual environment, the absolute values are indicative and the ratios between the models are what matter.

Table 3: *Experimental environment.*

Component	Specification
Operating environment	Linux (Ubuntu 24)
Processor	Intel® Xeon® Processor @ 2.10 GHz; 1 core
Software	Python 3.12.3, scikit-learn 1.8.0, NumPy 2.4.4
Timing tool	`time.perf_counter`

4.3. Deployment

The deployment phase of CRISP-DM was not carried out in this study: the work is retrospective in nature, the models have not undergone calibration or validation on external data, and they are not intended for clinical use.

5. Results and Discussion

5.1. Exploratory Analysis

Figure 1(a) shows the class distribution. In the PCA projection of the standardized features (Figure 1(b)), the first two components explain 44.3% and 19.0% of the variance 63.2% combined and the two classes are largely separated even in just two dimensions, although overlap is observed in the boundary region.

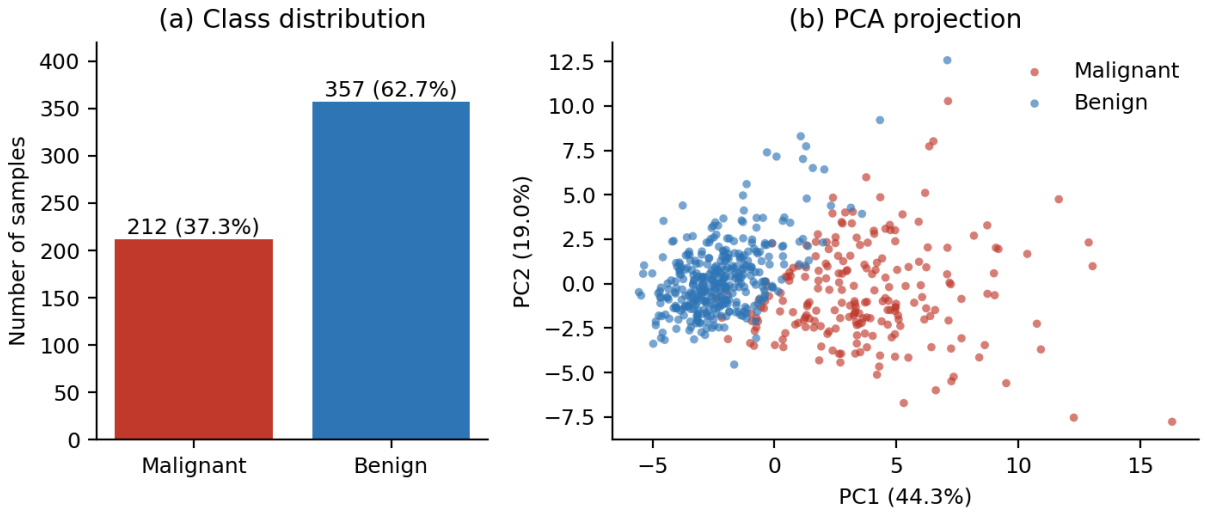


Figure 1: *Exploratory data analysis: (a) class distribution; (b) projection of standardized features onto the first two principal components.*

Strong correlations exist among the features (Figure 2): among the 435 pairs of the 30 features, 21 have $|r| > 0.9$, for example, the correlation between mean radius and mean perimeter is r = 0.9979, while the correlation between mean radius and mean area is r = 0.9874. This is important for interpreting feature importance (Section 5.8).

The class-wise distributions (Figure 3) show that size and shape-related measures, such as radius, perimeter, and concave points, are systematically higher in malignant tumors. According to the Mann-Whitney U test, the differences between the classes are statistically significant at the $p < 0.001$ level for 26 of the 30 features. In terms of single-feature ROC-AUC, the strongest feature is "worst perimeter" (AUC = 0.9755), the weakest features are "texture error" (AUC = 0.5116, $p = 0.644$) and "mean fractal dimension" (AUC = 0.5155, $p = 0.537$) (Table 4).

Table 4: *Discriminative power of single features (AUC based on Mann-Whitney U): the six strongest and two weakest features.*

Feature	AUC	Malignant (mean ± SD)	Benign (mean ± SD)	p
worst perimeter	0.9755	141.370 ± 29.457	87.006 ± 13.527	2.6×10^{-80}
worst radius	0.9704	21.135 ± 4.284	13.380 ± 1.981	1.1×10^{-78}
worst area	0.9698	1422.286 ± 597.968	558.899 ± 163.601	1.8×10^{-78}
worst concave points	0.9667	0.182 ± 0.046	0.074 ± 0.036	1.9×10^{-77}
mean concave points	0.9644	0.088 ± 0.034	0.026 ± 0.016	1.0×10^{-76}
mean perimeter	0.9469	115.365 ± 21.855	78.075 ± 11.807	3.6×10^{-71}
mean fractal dimension	0.5155	0.063 ± 0.008	0.063 ± 0.007	0.537
texture error	0.5116	1.211 ± 0.483	1.220 ± 0.589	0.644

5.2. Cross-Validation Results

Table 5 presents the mean values and standard deviations across 50 folds. LR achieves the highest overall accuracy (97.89%), precision (98.56%), and F1-score (97.11%), while SVM achieves the highest sensitivity (95.85%) and ROC-AUC (0.9952). RF has the lowest mean values across all five metrics and also exhibits greater variability across folds (SD = 2.02 percentage points for overall accuracy, compared with 1.43 for LR and 1.37 for SVM). Figure 4 illustrates these distributions using box plots.

Table 5: *Results of* 5×10*-fold cross-validation.*

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 (%)	ROC-AUC
LR	97.89 ± 1.43	98.56 ± 1.53	95.76 ± 3.39	97.11 ± 2.00	0.9948 ± 0.0054
SVM	97.56 ± 1.37	97.62 ± 2.32	95.85 ± 3.10	96.68 ± 1.88	0.9952 ± 0.0047
RF	95.92 ± 2.02	95.70 ± 3.30	93.39 ± 4.80	94.43 ± 2.81	0.9899 ± 0.0084

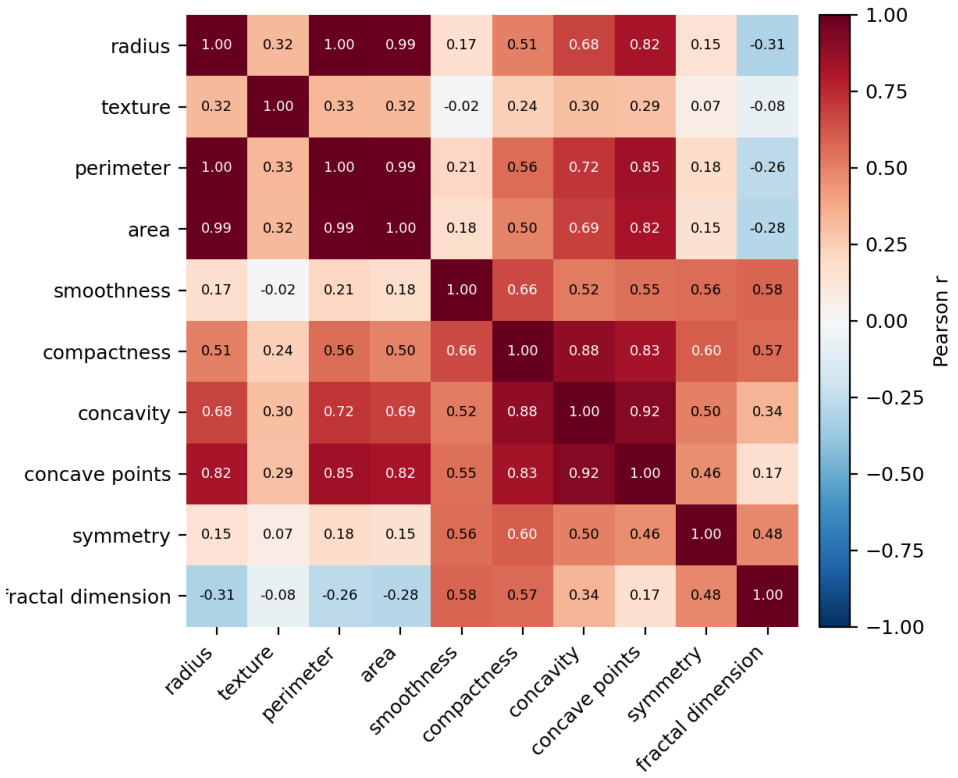


Figure 2: *Pearson correlation matrix among the "mean" features.*

5.3. Statistical Comparison

Table 6 presents the Nadeau-Bengio corrected pairwise comparisons. To control family-wise error across the three model-pair comparisons within each metric, Holm adjustment was applied to the corrected t-test p-values. The corrected pairwise tests are reported for Accuracy, Sensitivity, and F1, matching the stated study objective. The previously reported unadjusted corrected-test p-values for Accuracy and F1 all exceed 0.05. Therefore their Holm-adjusted values also exceed 0.05. Exact adjusted values are reported where they can be derived from the available results.

However, a non-significant difference does not imply equivalence: LR achieved higher accuracy than RF in 40 folds, while SVM did so in 32 folds (40 / 5 / 5 for LR-RF and 32 / 9 / 9 for SVM-RF, representing wins / ties / losses). Thus, the direction of the difference is consistent, but its magnitude is small relative to the variability across folds, and the statistical power of the test is limited given the dataset of 569 samples.

It is noteworthy that the Wilcoxon test yields very small p-values for the same fold differences (8.6×10^{-6} for SVM-RF, 7.3×10^{-7} for LR-RF). These values rely on the assumption of fold independence and do not account for the overlap between training sets. An uncorrected test would create a false sense of certainty here [12], [13].

Table 6: *Pairwise comparisons.*

Pair	Metric	Mean difference (pp)	95% CI (p.p.)	t (adj.)	p (N-B)	p (Wilcoxon)	Win / tie / loss
SVM − RF	Accuracy	+1.64	[−0.70; +3.97]	1.41	0.165	8.6×10^{-6}	32 / 9 / 9
SVM − RF	F1	+2.25	[−0.91; +5.41]	1.43	0.159	1.9×10^{-5}	34 / 3 / 13
SVM − LR	Accuracy	−0.33	[−1.46; +0.79]	−0.60	0.553	0.079	11 / 16 / 23
SVM − LR	F1	−0.42	[−1.95; +1.10]	−0.56	0.578	0.085	16 / 10 / 24
LR − RF	Accuracy	+1.97	[−0.28; +4.22]	1.76	0.085	7.3×10^{-7}	40 / 5 / 5
LR − RF	F1	+2.67	[−0.45; +5.80]	1.72	0.092	8.8×10^{-7}	40 / 3 / 7

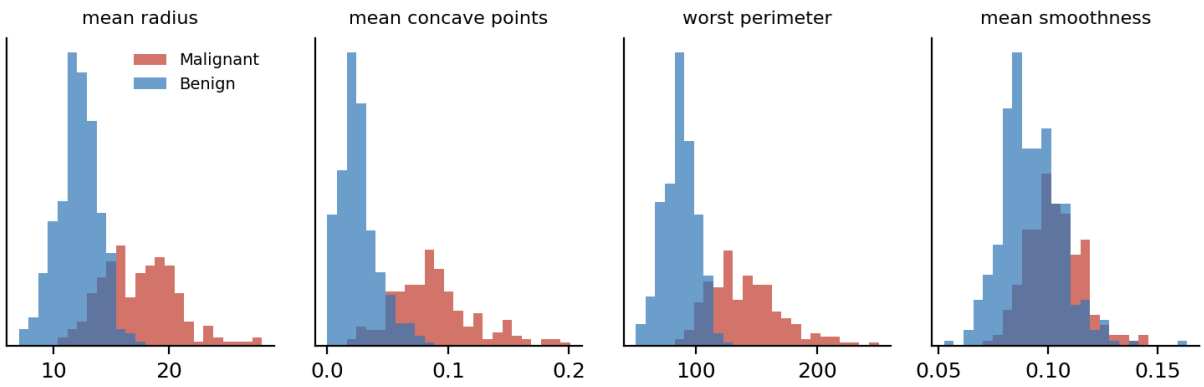


Figure 3: *Distribution of four features across classes (histogram).*

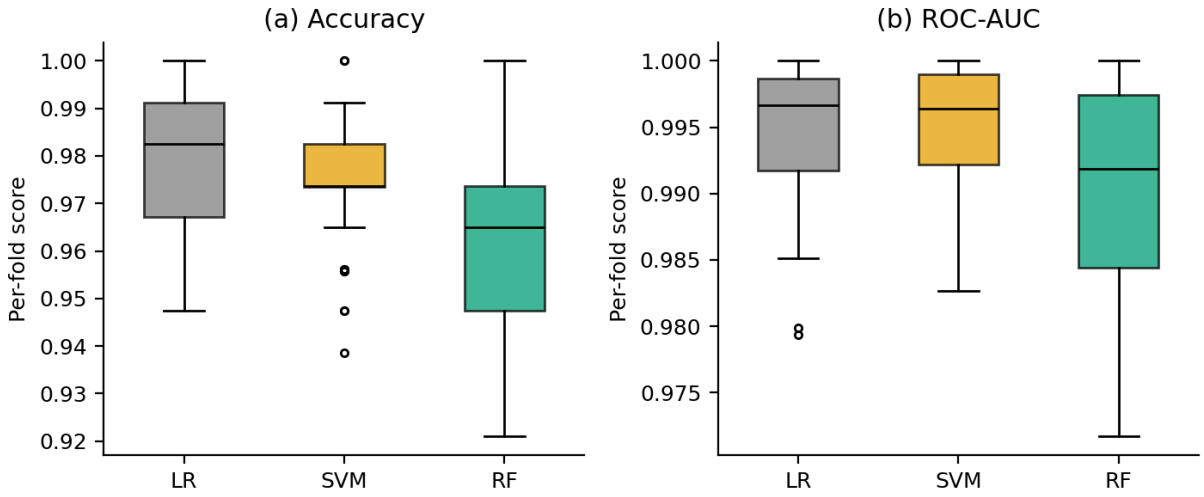


Figure 4: *Distribution of (a) overall accuracy and (b) ROC-AUC values across 50 cross-validation folds (box plot).*

5.4. Error Structure and ROC Analysis

Figure 5 presents the confusion matrices based on the fold-wise predictions from 10-fold cross-validation. Out of 212 malignant samples, LR correctly detected 202, SVM 204, and RF 198 (sensitivities of 95.28%, 96.23%, and 93.40%, respectively). Missed malignant cases (false negatives) numbered 10 for LR, 8 for SVM, and 14 for the random forest while false positive cases were 4, 6, and 11, respectively. The total number of errors was 14 for both LR and SVM (97.54% accuracy for both) and 25 for RF (95.61%). Missed malignant cases are the clinical priority, and SVM achieved the best result in this regard.

The ROC curves shown in Figure 6 confirm the high discriminative power of the three models: the AUC is 0.9947 for LR, 0.9951 for SVM, and 0.9892 for RF. Given such high AUC values, the differences between the models manifest primarily in the choice of the decision boundary and in the classification of samples located in the boundary region.

5.5. Computational Costs

Table 7 reports implementation-specific training and prediction measurements obtained on the stated single-core virtual environment and should not be interpreted as intrinsic algorithmic efficiency. LR and SVM had similar measured training times (4.22 and 4.25 ms), whereas RF required 397.3 ms in this particular implementation. Prediction timing was measured on the 143-sample test split and linearly normalized to a 1,000-sample equivalent for comparability: normalized time = measured time for 143 samples × (1000/143). These values characterize this dataset, software stack, hardware, and implementation only.

Note: Prediction values in the third column are 1,000-sample equivalents obtained by linear normalization of the measured 143-sample prediction time (×1000/143), not timings from a separate 1,000-sample test set.

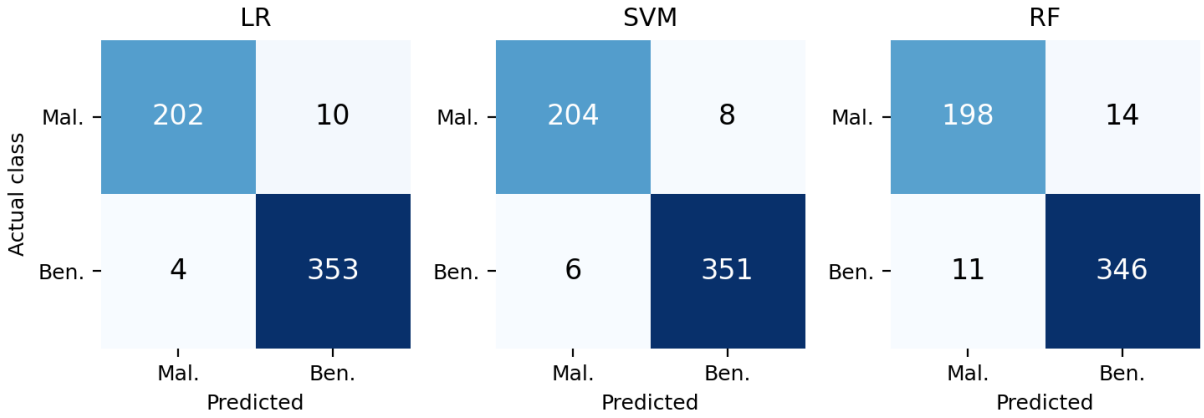


Figure 5: *Confusion matrices based on 10-fold cross-validation predictions.*

Table 7: *Training and prediction times.*

Model	Training (ms)	Prediction (normalized to 1,000 samples) (ms)	Training (SVM = 1)	Prediction (SVM = 1)
LR	4.22 ± 0.28	6.06 ± 0.88	1.0	0.5
SVM	4.25 ± 0.43	11.49 ± 1.09	1.0	1.0
RF	397.27 ± 14.02	116.44 ± 12.51	93.6	10.1

5.6. Hyperparameter Sensitivity and Nested Cross-Validation

Figure 7(a) shows the average accuracy of the SVM across the (C, gamma) grid. The best result was 97.65% at the point C = 10, gamma = 0.01. With a very small C (0.01) or a very large gamma (gamma = 1), the model effectively predicts the most frequent class, and the accuracy remains close to the baseline (62.7%)-specifically, between 62.7% and 63.3%. This demonstrates the SVM's sensitivity to hyperparameters. In Figure 7(b), the RF accuracy rises from 94.67% for 10 trees to approximately 96.0% for 50 or more trees, remaining virtually unchanged with further increases. Thus, the RF's lower performance cannot be attributed to an insufficient number of trees.

Nested 5×5-fold evaluation used five outer folds for performance estimation and five inner folds within each outer training partition for hyperparameter selection (Table 8). For each model, performance was aggregated as the mean and standard deviation of the five held-out outer-fold scores. The resulting mean accuracies were SVM 97.72%, LR 97.54%, and RF 96.49%. Because this nested analysis uses a different resampling design from the repeated 5-fold primary analysis in Table 5, the two sets of variability estimates should not be compared directly.

Table 8: *Overall accuracy (%) with standard and nested cross-validation.*

Model	Standard 5×10 (mean ± SD)	Nested 5×5 (mean ± SD)	Search space
LR	97.89 ± 1.43	97.54 ± 1.44	$C \in \{0.01, 0.1, 1, 10, 100\}$
SVM	97.56 ± 1.37	97.72 ± 1.47	$C \in \{0.1, 1, 10, 100\}$, $\gamma \in \{0.001, 0.01, 0.1, \text{“scale”}\}$
RF	95.92 ± 2.02	96.49 ± 2.91	`max_features` $\in$ {“sqrt”, “log2”, 0.5}, `min_samples_leaf` $\in \{1, 2, 4\}$; 200 trees

5.7. Learning Curves

Figure 8 illustrates the dependence of validation accuracy on the size of the training set. With the smallest training set (45 samples), LR yields 85.9%, compared with 79.3% for SVM and 82.1% for RF. When the training set reaches 455 samples, LR achieves 97.4% and SVM 97.7%. For RF, training accuracy is 100% while validation accuracy is 95.3%, giving a larger train-validation gap than LR (1.4 p.p.) and SVM (0.9 p.p.). This pattern is consistent with greater overfitting in this learning-curve experiment, but the gap is not a formal estimate of model variance, accordingly, no claim of intrinsically higher RF variance is made.

5.8. Importance of the Features

According to RF Gini importance, the most significant features are "worst perimeter" (14.6%), "worst area" (14.4%), and "worst concave points" (11.5%). These three features collectively account for 40.4% of the total importance, while the top ten features account for 80.2% (Figure 9(a)). However, in the hold-out test set (n = 143, RF accuracy 95.80%), the permutation importance for all these features is close to zero (Figure 9(b)), and the standard deviations are large relative

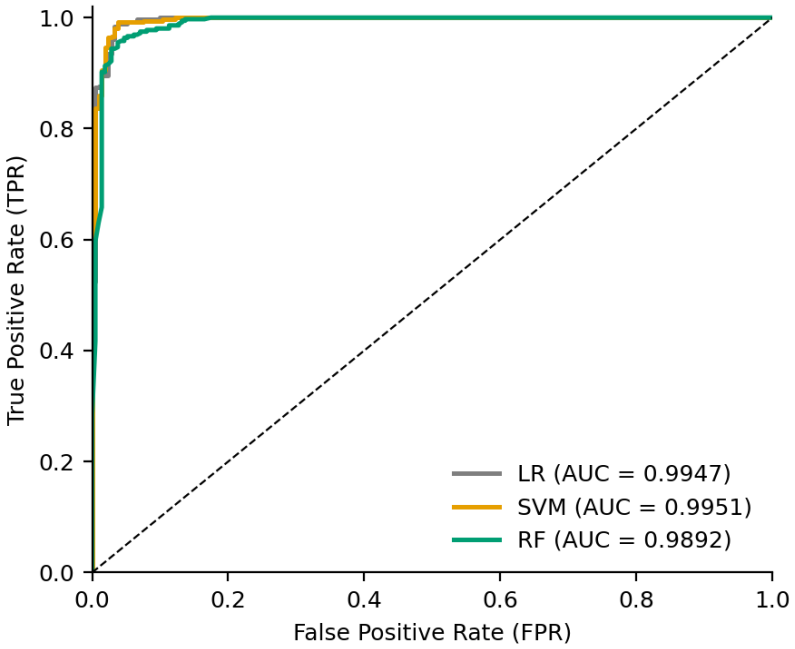


Figure 6: *ROC curves based on 10-fold cross-validation fold-wise predictions.*

to the mean values. This is due to strong correlation: when one feature is permuted, other features that are virtually equivalent to it (e.g., radius, perimeter, and area) retain the information, so accuracy does not drop. Consequently, neither Gini importance nor permutation importance allows for definitive conclusions regarding the "importance" of individual features in such datasets. This issue is also noted in the literature [18].

5.9. Discussion

The most significant aspect of the results is that, in this dataset of 569 samples, the complex ensemble model (RF) does not outperform the simpler models (LR, SVM) while it nominally yields weaker results, this difference is not statistically significant in the adjusted test. This does not contradict the general rankings from the extensive comparisons noted in Section 2.2, as those rankings reflect average performance across a large number of datasets [8], [9], whereas a different ranking is possible for a specific dataset [10].

The fact that LR performs on par with RBF-SVM indicates that the classes are largely linearly separable in this feature space. Consequently, the added flexibility of a non-linear model offers no significant advantage. This is likely due to the nature of the data: the features represent nuclear measurements, and indicators of size and shape are systematically higher in malignant lesions (Figure 3). The underperformance of RF can be attributed to the fact that tree-based splits are less efficient at approximating linear boundaries when dealing with strongly correlated, continuous features and small sample sizes; however, this explanation was not directly tested in this study and remains hypothetical.

From a benchmarking standpoint, the results can be summarized as follows: (i) on this dataset, standardized LR provides a simple and computationally inexpensive baseline (ii) SVM achieves comparable aggregate performance and fewer missed malignant cases in the separate 10-fold descriptive analysis, while remaining sensitive to hyperparameters and (iii) RF showed higher implementation-specific computational cost without higher accuracy in this experiment. These findings concern comparative benchmarking on the Wisconsin dataset only. The models were not externally validated, clinically calibrated, or evaluated prospectively, so the results should not be interpreted as evidence for clinical diagnosis or deployment. Furthermore, statistical comparisons should account for dependence introduced by overlapping cross-validation training sets.

5.10. Study Limitations

First, the study utilized only a single dataset (569 samples) and three algorithms. The results should not be generalized to other domains or larger datasets. Second, the features used were pre-extracted from the images rather than raw image data. Third, standard hyperparameters were employed for the primary comparison while hyperparameter tuning was applied only within a nested cross-validation setup involving smaller search spaces and model-specific configurations. Fourth, execution-time measurements were obtained in one single-core virtual environment and reflect this specific implementation rather than intrinsic algorithmic efficiency. Fifth, the models did not undergo probability calibration, decision-threshold

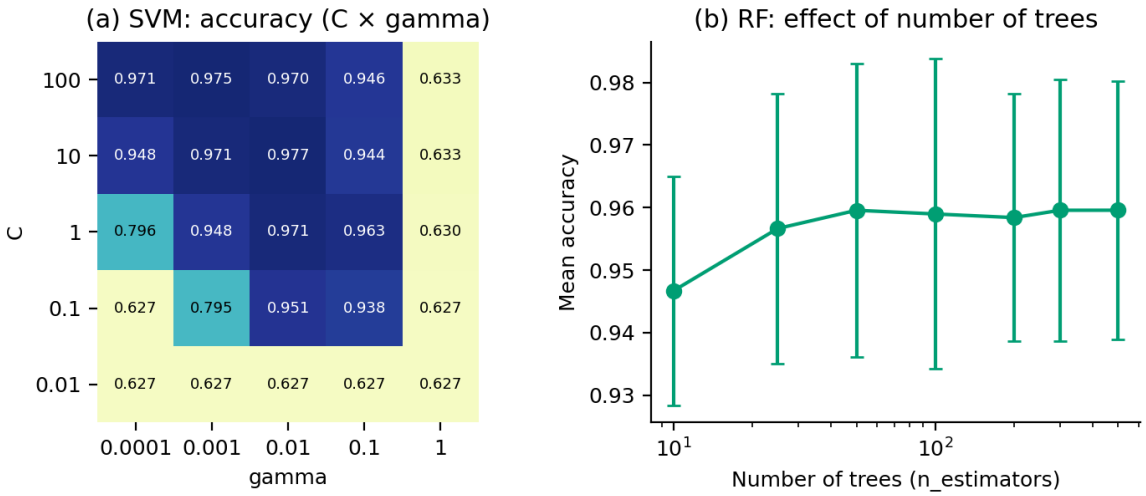


Figure 7: *Hyperparameter sensitivity: (a) SVM accuracy across the (C, γ) grid (b) RF accuracy as a function of the number of trees.*

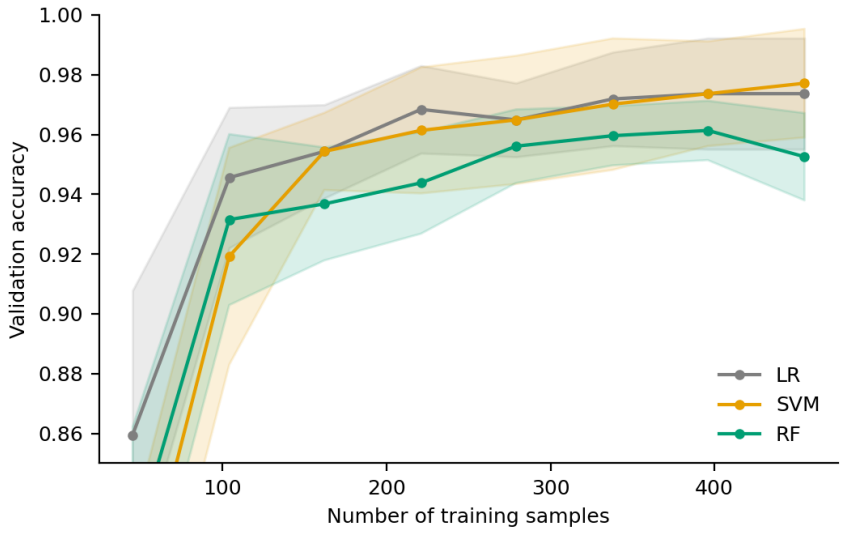


Figure 8: *Learning curves (5-fold validation, shaded area = ±1 SD).*

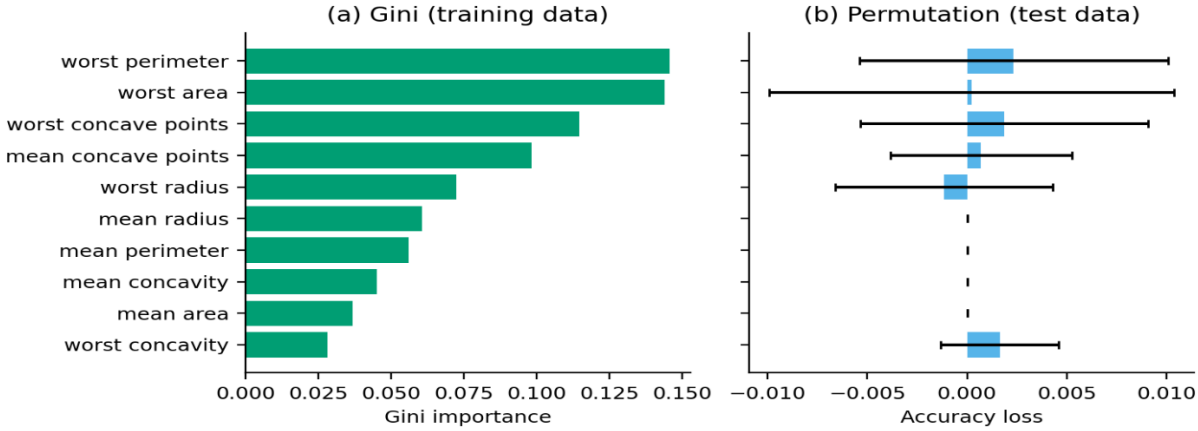


Figure 9: *Top 10 most important features for the random forest model: (a) Gini importance; (b) permutation importance on the test set (30 repetitions).*

optimization, prospective evaluation, or external validation. Consequently, this work is a benchmarking study and cannot establish clinical diagnostic utility or deployment readiness. Finally, the relative class imbalance (37:63) is moderate. In cases of greater imbalance, an analysis of precision-recall curves might yield more informative insights [19].

6. Conclusion and Future Work

This study benchmarked LR, SVM, and RF on the Wisconsin Diagnostic Breast Cancer dataset using repeated cross-validation as the primary comparison protocol. LR (97.89%) and SVM (97.56%) nominally outperformed RF (95.92%) in mean accuracy, but the Nadeau-Bengio corrected tests reported in the study did not establish statistically significant pairwise accuracy differences. The separate 10-fold prediction analysis summarized error structure and ROC behavior and was not used to support the primary inferential comparison. RF had higher measured training and prediction cost in the specific single-core implementation, but these timing ratios should not be generalized as intrinsic algorithmic efficiency. Because the models were not externally validated or clinically calibrated, the findings support benchmarking conclusions only and not clinical diagnosis or deployment.

Future work plans include: (i) incorporating gradient boosting and neural networks (ii) performing multiple comparisons across several datasets using the Friedman test (iii) calibrating probabilities and selecting a decision threshold focused on sensitivity (iv) validation on external data and (v) interpretation of features using grouped importance methods.

Author Declarations

Author Contributions (CRediT): Ismayilov Togrul Atabay: Conceptualization; Data Curation; Methodology; Software; Formal Analysis; Validation; Visualization; Writing – Original Draft; Writing – Review & Editing. The author has read and approved the submitted version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in the study is open-source (UCI Machine Learning Repository [4]).

Acknowledgments: Not applicable.

Conflicts of Interest: The author declares no financial, professional, or personal conflicts of interest regarding this study.

Ethics Statement: Not applicable.

Declaration of Generative AI: During the preparation of this manuscript, the author used Claude (Anthropic) for formatting the manuscript according to the journal template. The author reviewed and verified the resulting content and takes full responsibility for the accuracy, integrity, and content of the manuscript.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second edition, New York, New York, United States: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[2] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Biomedical Image Processing and Biomedical Visualization*, Proc. SPIE, vol. 1905, pp. 861-870, 1993, doi: 10.1117/12.148698.

[3] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 23, pp. 9193-9196, 1990, doi: 10.1073/pnas.87.23.9193.

[4] W. H. Wolberg, O. L. Mangasarian, and W. N. Street, "Breast Cancer Wisconsin (Diagnostic)," *UCI Machine Learning Repository*, 1995, doi: 10.24432/C5DW2B.

[5] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, Third edition, Hoboken, New Jersey, United States: Wiley, 2013, doi: 10.1002/9781118548387.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995, doi: 10.1007/BF00994018.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[8] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research*, vol. 15, pp. 3133-3181, 2014.

[9] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proc. 23rd Int. Conf. Machine Learning (ICML)*, 2006, pp. 161-168, doi: 10.1145/1143844.1143865.

[10] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, no. 7, pp. 1341-1390, 1996, doi: 10.1162/neco.1996.8.7.1341.

[11] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006.

[12] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239-281, 2003, doi: 10.1023/A:1024068626366.

[13] R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Advances in Knowledge Discovery and Data Mining: PAKDD 2004*, Lecture Notes in Computer Science, vol. 3056. Berlin, Germany: Springer, 2004, pp. 3-12, doi: 10.1007/978-3-540-24775-3_3.

[14] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. SPSS Inc., 2000.

[15] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, vol. 2, 1995, pp. 1137-1143.

[17] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80-83, 1945, doi: 10.2307/3001968.

[18] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, Art. no. 307, 2008, doi: 10.1186/1471-2105-9-307.

[19] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[20] S. S. Das, A. Mahaprasad, N. Padhy, S. Misra, R. Panigrahi, P. K. Mahapatro, and D. Arangi, "Comparative evaluation of machine learning classifiers for breast cancer diagnosis: A comprehensive statistical analysis," *Engineering Proceedings*, vol. 124, no. 1, Art. no. 35, 2026, doi: 10.3390/engproc2026124035.

[21] J.-J. Jung et al., "Development and external validation of a machine learning model to predict pathological complete response after neoadjuvant chemotherapy in breast cancer," *Journal of Breast Cancer*, vol. 26, no. 4, pp. 353-362, 2023, doi: 10.4048/jbc.2023.26.e14.