*Salud, Ciencia y Tecnología*. 2025; 5:2241
doi: 10.56294/saludcyt20252241

**ORIGINAL**

# Threshold-Optimized and Calibrated Logistic Regression for Breast Cancer Classification

## Regresión logística calibrada y optimizada por umbral para la clasificación del cáncer de mama

Ronal Watrianthos[1], Rayendra[1], Ervan Asri[1], Yuhefizar[1], Humaira[1]

[1]Politeknik Negeri Padang, Department of Information Technology. Padang, Indonesia.

**Cite as:** Watrianthos R, Rayendra, Asri E, Yuhefizar, Humaira. Threshold-Optimized and Calibrated Logistic Regression for Breast Cancer Classification. Salud, Ciencia y Tecnología. 2025; 5:2241. https://doi.org/10.56294/saludcyt20252241

**Submitted:** 29-08-2025 **Revised:** 26-09-2025 **Accepted:** 15-10-2025 **Published:** 16-10-2025

**Editor:** Prof. Dr. William Castillo-González

**Corresponding Author:** Ronal Watrianthos

**ABSTRACT**

Breast cancer affects over 2,3 million individuals annually worldwide. Traditional diagnostic methods face limitations in consistency and objectivity, particularly in resource-constrained settings. This study developed a logistic regression-based clinical decision support system for breast cancer classification. We analyzed the Wisconsin Diagnostic Breast Cancer dataset containing 569 samples with 30 quantitative morphological features from fine needle aspirate cytology. The dataset comprised 357 benign and 212 malignant cases. Data underwent standardization via StandardScaler, followed by 75-25 train-test partitioning (426 training, 143 testing samples). We evaluated the logistic regression model through confusion matrix analysis, ROC curve assessment, threshold optimization via Youden's Index, and probability calibration using Expected Calibration Error (ECE). The model achieved 95,8 % accuracy, 96,2 % sensitivity, and 95,6 % specificity on independent testing data, with AUC-ROC of 0,993. Threshold optimization identified 0,560 as the optimal decision boundary, yielding 3,77 % false negative rate and 4,44 % false positive rate. Probability calibration demonstrated reliable predictions with ECE of 0,0390, improved to 0,0328 through isotonic regression. The model correctly classified 137 of 143 test samples (86 true negatives, 51 true positives, 4 false positives, 2 false negatives). The logistic regression model demonstrated strong discriminative performance for breast cancer classification. However, single train-test validation and dataset-specific characteristics require cautious interpretation. Cross-validation and external validation remain necessary for clinical translation.

**Keywords:** Breast Cancer Classification; Logistic Regression; Clinical Decision Support; Probability Calibration; Fine Needle Aspirate Cytology.

**RESUMEN**

El cáncer de mama afecta a más de 2,3 millones de personas al año en todo el mundo. Los métodos de diagnóstico tradicionales tienen limitaciones en cuanto a consistencia y objetividad, especialmente en entornos con recursos limitados. En este estudio se desarrolló un sistema de apoyo a la toma de decisiones clínicas basado en la regresión logística para la clasificación del cáncer de mama. Se analizó el conjunto de datos Wisconsin Diagnostic Breast Cancer, que contiene 569 muestras con 30 características morfológicas cuantitativas obtenidas mediante citología por aspiración con aguja fina. El conjunto de datos comprendía 357 casos benignos y 212 malignos. Los datos se sometieron a una estandarización mediante StandardScaler, seguida de una partición de entrenamiento-prueba 75-25 (426 muestras de entrenamiento y 143 de prueba). Evaluamos el modelo de regresión logística mediante el análisis de la matriz de confusión, la evaluación de la curva ROC, la optimización del umbral mediante el índice de Youden y la calibración de la probabilidad utilizando el error de calibración esperado (ECE). El modelo alcanzó una precisión del 95,8 %, una sensibilidad

del 96,2 % y una especificidad del 95,6 % en datos de prueba independientes, con un AUC-ROC de 0,993. La optimización del umbral identificó 0,560 como el límite de decisión óptimo, lo que dio lugar a una tasa de falsos negativos del 3,77 % y una tasa de falsos positivos del 4,44 %. La calibración de la probabilidad demostró predicciones fiables con un ECE de 0,0390, que mejoró a 0,0328 mediante regresión isotónica. El modelo clasificó correctamente 137 de las 143 muestras de prueba (86 negativos verdaderos, 51 positivos verdaderos, 4 falsos positivos y 2 falsos negativos). El modelo de regresión logística demostró un fuerte rendimiento discriminatorio para la clasificación del cáncer de mama. Sin embargo, la validación de un solo entrenamiento y las características específicas del conjunto de datos requieren una interpretación cautelosa. La validación cruzada y la validación externa siguen siendo necesarias para la traslación clínica.

**Palabras clave:** Clasificación del Cáncer de Mama; Regresión Logística; Apoyo a la Toma de Decisiones Clínicas; Calibración de Probabilidades; Citología por Aspiración con Aguja Fina.

## INTRODUCTION

Cancer represents one of the most formidable challenges in modern medicine, characterized by uncontrolled cellular proliferation and the potential for metastatic spread to distant organs. This group of diseases accounts for nearly 10 million deaths annually worldwide, with profound impacts on healthcare systems and patient quality of life. Different cancer types exhibit distinct biological behaviors, growth patterns, and clinical outcomes, ranging from indolent lesions requiring minimal intervention to aggressive malignancies demanding intensive multimodal therapy.

Among these malignancies, breast cancer continues to pose a significant challenge in modern oncology, ranking as the second leading cause of cancer-related mortality among women worldwide and affecting over 2,3 million individuals annually.[1] The heterogeneous nature of breast malignancies, along with the critical importance of early detection for treatment efficacy and patient survival, highlights the essential need for accurate, reliable, and accessible diagnostic methodologies.[2,3] While traditional diagnostic approaches form the foundation of current clinical practice, they encounter inherent limitations in consistency, objectivity, and accessibility across diverse healthcare settings, particularly in resource-constrained environments where specialized expertise may be limited.[4]

The advent of digital pathology and computational analysis has fundamentally transformed the field of medical diagnostics, presenting unprecedented opportunities to enhance diagnostic accuracy through quantitative morphological assessment.[5,6] Fine needle aspiration cytology, a minimally invasive diagnostic procedure, produces extensive morphological data from cellular specimens, which can be systematically analyzed using advanced computational techniques.[7] Nevertheless, the interpretation of these morphological features has traditionally depended heavily on the expertise of pathologists, introducing potential variability in diagnostic consistency and creating obstacles to standardized care delivery.[8]

Cancer classification represents a fundamental supervised learning problem wherein computational algorithms learn to distinguish between malignant and benign lesions based on quantifiable features extracted from clinical specimens. Machine learning technologies have emerged as formidable tools for addressing this diagnostic challenge.[9,10] Recent advancements in artificial intelligence have yielded promising results in various oncological applications, ranging from radiological imaging interpretation to histopathological analysis, thereby suggesting significant opportunities for the development of clinical decision support systems.[11,12]

Among classification algorithms, logistic regression occupies a distinctive position in clinical applications due to its interpretability, probabilistic output, and computational efficiency. Unlike black-box approaches, logistic regression provides transparent decision-making processes through interpretable coefficients, generates calibrated probability estimates rather than mere class labels, and requires minimal computational resources suitable for resource-constrained healthcare settings.

However, the clinical deployment of logistic regression for cancer classification necessitates careful attention to threshold optimization and probability calibration. Threshold optimization determines the decision boundary that balances sensitivity and specificity according to clinical priorities, while probability calibration ensures that predicted probabilities accurately reflect true malignancy risk. The integration of machine learning approaches with morphological feature analysis represents a convergent pathway toward more objective, consistent, and accessible diagnostic capabilities.[13]

Despite technological advancements, significant gaps remain in the clinical translation of machine learning diagnostic tools. Numerous existing studies predominantly focus on classification accuracy, often overlooking the probabilistic reliability, clinical interpretability, and threshold optimization requirements essential for practical clinical application.[14,15,16,17,18,19,20,21,22] The recognition that classification accuracy alone is insufficient for clinical applications emerged gradually through several decades of statistical and medical research. In the

1980s, researchers began documenting discrepancies between predicted probabilities and observed outcomes in logistic regression models applied to cardiovascular risk prediction.[23,24]

Contemporary machine learning research has witnessed renewed attention to calibration assessment, particularly as complex algorithms like deep neural networks often produce overconfident predictions despite high classification accuracy.[25,26] Modern frameworks now recognize calibration as essential for clinical decision support systems, where miscalibrated probabilities can lead to inappropriate treatment decisions regardless of discriminative performance.[27] The absence of comprehensive calibration assessment in diagnostic modeling can be a significant oversight, as clinical decision-making necessitates not only accurate classifications but also reliable confidence estimates that can be meaningfully interpreted as risk probabilities.[28,29,30]

Moreover, the predominant focus of machine learning methodologies in medical diagnosis is on algorithmic complexity rather than clinical interpretability. This results in the development of advanced models that, while achieving notable performance metrics, remain largely opaque to clinicians.[16,17] This lack of interpretability presents significant challenges to clinical adoption, as healthcare providers need to comprehend the underlying decision-making processes to effectively incorporate computational tools into existing diagnostic workflows.[18,31] In contrast to complex algorithmic approaches, logistic regression offers intrinsic interpretability through transparent coefficient estimation and probabilistic reasoning, making it particularly suitable for clinical decision support applications. The interpretability of logistic regression extends beyond model coefficients to encompass two critical components emphasized in this study: threshold optimization and probability calibration.[32,33]

The Wisconsin Diagnostic Breast Cancer dataset represents a landmark resource in computational diagnostic research.[34,35] This dataset provides a comprehensive quantitative characterization of cellular morphological features derived from fine needle aspirate samples, facilitating a detailed analysis of the relationship between morphological patterns and diagnostic outcomes. This publicly available dataset comprises 569 digitized fine needle aspirate (FNA) samples from breast masses, with each sample characterized by 30 quantitative morphological features derived from digital image analysis of cell nuclei present in the aspirate specimens.

The objective of this study was to evaluate the clinical applicability of logistic regression for breast cancer classification using the Wisconsin Diagnostic Breast Cancer dataset. This evaluation specifically addresses the identified gaps in current machine learning diagnostic research: the absence of systematic threshold optimization methodologies that align decision boundaries with clinical priorities, the limited assessment of probability calibration quality that ensures reliable risk estimation, and the need for interpretable models that facilitate clinical adoption. By focusing on these underexplored aspects rather than solely on classification accuracy, this study contributes a methodological framework for translating machine learning models into clinically actionable diagnostic support tools. This systematic approach provides a methodological contribution that enhances the understanding of how machine learning models can be effectively evaluated and implemented in clinical diagnostic applications.

## METHOD

This study represents a descriptive observational study with a retrospective cross-sectional design. No intervention was performed by the researchers; instead, we conducted secondary analysis of pre-existing data to develop and evaluate a diagnostic classification model. The study utilized previously collected morphological measurements to assess the discriminative performance, threshold optimization, and probability calibration of logistic regression for breast cancer classification.

This study employed the Wisconsin Diagnostic Breast Cancer (WDBC) dataset,[36] which consists of 569 digitized fine-needle aspirate samples from breast masses. Each sample is characterized by 30 quantitative morphological features derived from cell nuclei measurements. These features are systematically categorized into three statistical aggregations: mean values, standard errors, and worst (largest) values for ten core morphological characteristics, namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

The target variable pertains to the histopathological diagnosis, encoded as a binary classification wherein malignant cases are identified as positive instances (M=1) and benign cases as negative instances (B=0). The data preprocessing phase involved the removal of the patient identification variable to prevent data leakage and the application of label encoding transformation to convert categorical diagnoses into numerical format, utilizing scikit-learn's LabelEncoder class. An analysis of missing values was conducted across all features, confirming complete data integrity with zero null values, thereby obviating the need for imputation strategies. The final dataset retained all 30 morphological features for subsequent analysis.

Due to the considerable variation in scale among morphological features, which range from fractional smoothness measurements to calculations involving areas of thousands of units, standardization was employed to ensure that each feature contributes equally to the logistic regression objective function. The StandardScaler transformation was applied in accordance with the following equation:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Where $x_{ij}$ represents the original feature value, $\mu_j$ is the feature mean, $\sigma_j$ is the feature standard deviation, and $z_{ij}$ is the standardized value.

To mitigate the risk of data leakage, the scaler was exclusively fitted to the training data before being applied to both the training and testing datasets. The data was partitioned using a 75-25 train-test split, facilitated by scikit-learn's train_test_split function with a random_state set to 0 to ensure reproducibility. This process resulted in 426 training samples and 143 testing samples.

Figure 1 illustrates that this division ensures an adequate amount of training data for model development while retaining sufficient samples for independent performance evaluation. Following partitioning, feature standardization was conducted, with the scaler being fitted exclusively to the training data and the transformation applied to both sets. This approach prevents information leakage that could potentially compromise the validity of the evaluation.

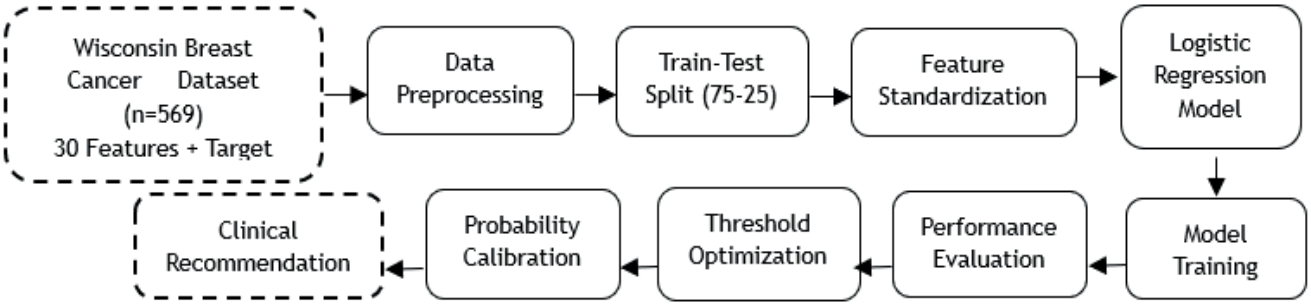


**Figure 1.** Research Flowchart

The model implementation employed logistic regression with the L-BFGS solver and L2 regularization, utilizing default parameters.[37,38] During the training phase, the model was fitted to standardized training data to determine optimal feature coefficients for binary classification, with the capability of predicting probabilities. The use of scikit-learn's LogisticRegression class with default parameters is indicative of standard practice in clinical machine learning applications.

The logistic regression model calculates the probability of malignancy by employing the sigmoid function, which transforms linear combinations of standardized features into probability values ranging from zero to one. Given that the WDBC dataset contains 30 morphological features (as described above), the model computes the following relationship:

$$\mathcal{P}(y = 1|X) = \frac{1}{1 + e^{\beta\emptyset + \sum_{i=1}^{\rho} \beta i x i}}$$

Where ßØ represents the intercept term, ßi denotes individual feature coefficients for each of the 30 morphological features, xi indicates the corresponding standardized feature values, and ρ equals the total number of features (30 in this study). This probabilistic framework enables not only binary classification decisions but also confidence quantification for clinical risk assessment.

**Performance Evaluation and ROC Analysis**

The assessment of model performance utilized a comprehensive evaluation framework tailored to meet the clinical demands of diagnostic support systems. The evaluation focused on confusion matrix analysis, addressing both sensitivity to malignant cases and specificity for benign classifications. The elements of the confusion matrix (True Negatives, False Positives, False Negatives, True Positives) facilitated the computation of standard binary classification metrics.

Receiver Operating Characteristic (ROC) analysis offers a threshold-independent evaluation of performance, which is crucial for comprehending model behavior across various clinical decision-making contexts.[39] The ROC curve, constructed by plotting the True Positive Rate against the False Positive Rate across different probability thresholds, facilitates the assessment of discriminative ability independent of specific cutoff values.[40]

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) serves as a quantitative measure of overall discriminative performance, with values ranging from 0,5, indicating random classification, to 1,0,

signifying perfect discrimination. This metric offers a singular summary statistic for the comparison of model performance and the assessment of clinical utility.

**Probability Calibration Assessment**

The assessment of model calibration is crucial for ensuring the provision of reliable probability estimates in clinical decision support systems. Accurately calibrated probabilities allow for the interpretation of model confidence as true risk estimates. The quality of calibration is quantified using the Expected Calibration Error (ECE),[41] which measures the average difference between predicted probabilities and actual outcomes across probability bins:

$$ECE = \sum_{m=1}^{M} \frac{Bm}{n} |acc(Bm) - conf(Bm)|$$

Where Bm represents samples in bin m, acc(Bm)denotes the accuracy within the bin, conf(Bm) indicates the average confidence, and n represents the total sample count. ECE values below 0,05 generally indicate well-calibrated models suitable for clinical applications.

Techniques for improving calibration, such as Platt scaling and isotonic regression,[42,43] were assessed to enhance the reliability of probability estimates for clinical decision support systems. Platt scaling utilizes a sigmoid transformation of model outputs, whereas isotonic regression employs a non-parametric monotonic transformation to improve calibration while maintaining the ranking order.

**Ethical Considerations**

This study utilized the Wisconsin Diagnostic Breast Cancer dataset, which is publicly available through the University of California Irvine Machine Learning Repository for research and educational purposes. The original data collection at the University of Wisconsin Clinical Sciences Center (1989-1991) was conducted under appropriate institutional review and informed consent procedures as documented in the original publications.[44,45]

The dataset contains fully de-identified, anonymized patient information with no personal identifiers, demographic data, or any information that could be used to re-identify individuals. Patient identification numbers were removed from the dataset prior to public release. The secondary analysis of this de-identified, publicly available dataset does not constitute human subjects research as defined by international ethical guidelines (Declaration of Helsinki) and institutional review board standards.

As this study involved only computational analysis of pre-existing, de-identified, publicly available data without any patient contact, recruitment, or intervention, formal institutional review board approval was not required according to the regulations governing secondary data analysis at our institution. Nevertheless, we adhered to principles of responsible research conduct, including transparent reporting of methodological procedures, acknowledgment of data sources, and commitment to reproducible science.

**Limitations of Secondary Data Analysis**

This study utilized the Wisconsin Diagnostic Breast Cancer dataset, which is publicl As a secondary analysis of pre-existing data, this study inherits limitations from the original data collection procedures, including potential selection bias in the patient population, temporal constraints (data collected 1989-1991 may not reflect current clinical populations), and lack of control over feature extraction protocols. The geographic restriction to a single institution (University of Wisconsin) may limit generalizability to diverse clinical settings and patient populations.

## RESULTS

The Wisconsin Diagnostic Breast Cancer dataset consists of 569 samples, all of which exhibit complete data integrity, with no missing values across the 30 morphological features. The distribution of the target variable indicates 357 benign cases (62,7 %) and 212 malignant cases (37,3 %), reflecting a moderate class imbalance with a benign-to-malignant ratio of 1,68:1. Following a 75-25 train-test split, the training set comprises 426 samples, while the test set includes 143 samples, maintaining similar class distributions.

The logistic regression model demonstrated an overall accuracy of 95,8 % on the independent test set, correctly predicting 137 out of 143 total samples. The analysis of the confusion matrix revealed 86 true negatives, 4 false positives, 2 false negatives, and 51 true positives (table 1).

**Table 1.** Confusion Matrix and Primary Performance Metrics

| Metric | Value | Clinical Interpretation |
|---|---|---|
| True Negatives (TN) | 86 | Benign cases correctly identified |
| False Positives (FP) | 4 | Benign cases misclassified as malignant |
| False Negatives (FN) | 2 | Malignant cases misclassified as benign |
| True Positives (TP) | 51 | Malignant cases correctly identified |
| Sensitivity (Recall) | 96,2 % | Proportion of malignant cases detected |
| Specificity | 95,6 % | Proportion of benign cases correctly identified |
| Precision (PPV) | 92,7 % | Accuracy of malignant predictions |
| F1-Score | 0,944 | Harmonic mean of precision and recall |
| Accuracy | 95,8 % | Overall classification correctness |

Figure 2 presents a heatmap of the confusion matrix, with the majority of predictions concentrated along the diagonal, indicating accurate classifications. The visualization displays the distribution of true negatives (86), false positives (4), false negatives (2), and true positives (51), with darker blue shades representing higher counts. The matrix evidences strong performance, as most predictions are concentrated along the diagonal.

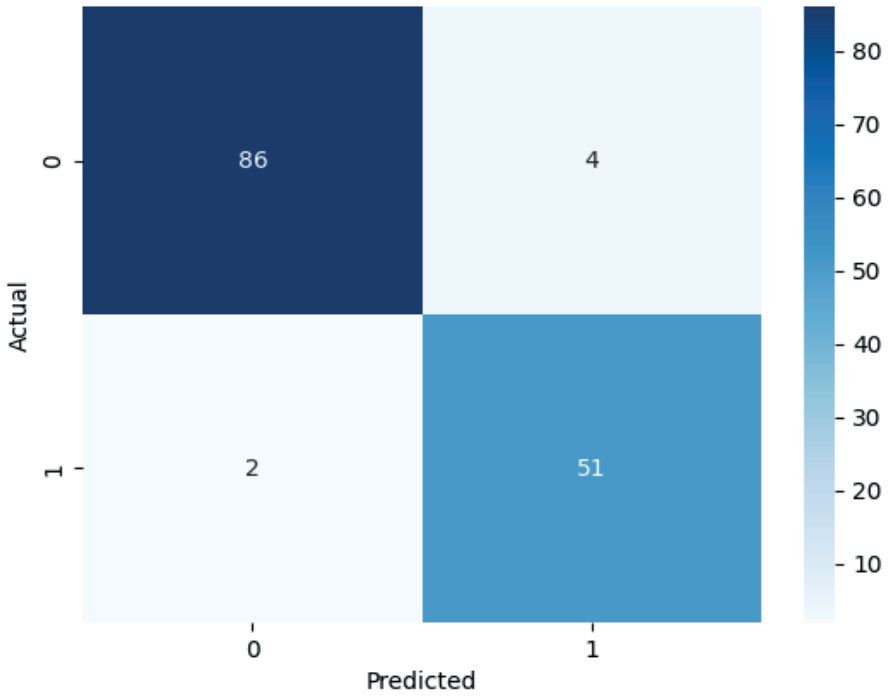


**Figure 2.** Confusion Matrix

**Note:** the matrix displays classification results on the independent test set (n=143 samples) using logistic regression with default threshold (0,5)

**ROC Analysis and Threshold Optimization**

The Receiver Operating Characteristic (ROC) analysis yielded an Area Under the Curve (AUC-ROC) of 0,993, indicating an almost perfect capacity to differentiate between malignant and benign cases. The ROC curve demonstrated a rapid ascent towards the upper-left corner, with minimal false positive rates across most sensitivity levels. Figure 3 presents the ROC curve alongside the sensitivity-specificity trade-off analysis, illustrating the model's exceptional discriminative performance and the relationship between sensitivity and specificity across various probability thresholds.

The left panel illustrates the Receiver Operating Characteristic (ROC) curve with an Area Under the Curve (AUC) of 0,993, indicating a near-perfect ability to discriminate between malignant and benign cases. The curve's rapid ascent towards the upper-left corner signifies excellent performance across all thresholds. The right panel presents the trade-off between sensitivity and specificity across various probability thresholds, demonstrating how these metrics fluctuate as the decision boundary is adjusted.

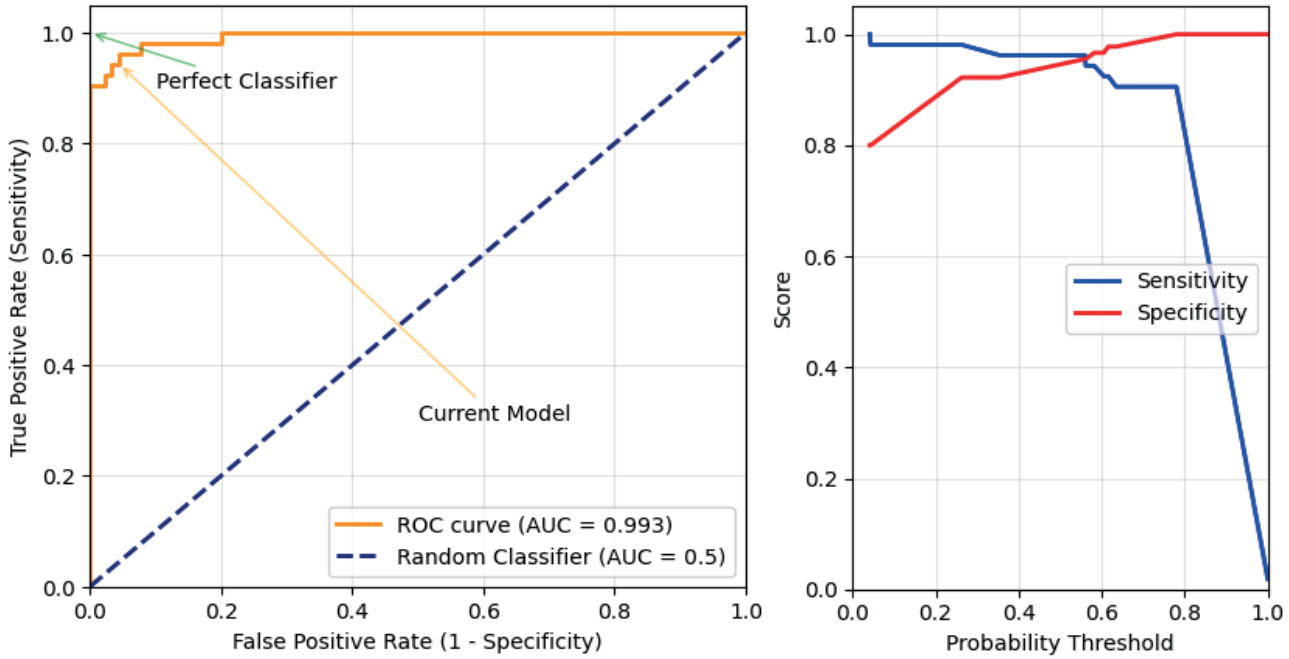


**Figure 3.** Receiver Operating Characteristic (ROC) Curve
**Note:** ROC curve generated from predicted probabilities on the test set (n=143 samples). The diagonal dashed line represents random classification (AUC=0,5)

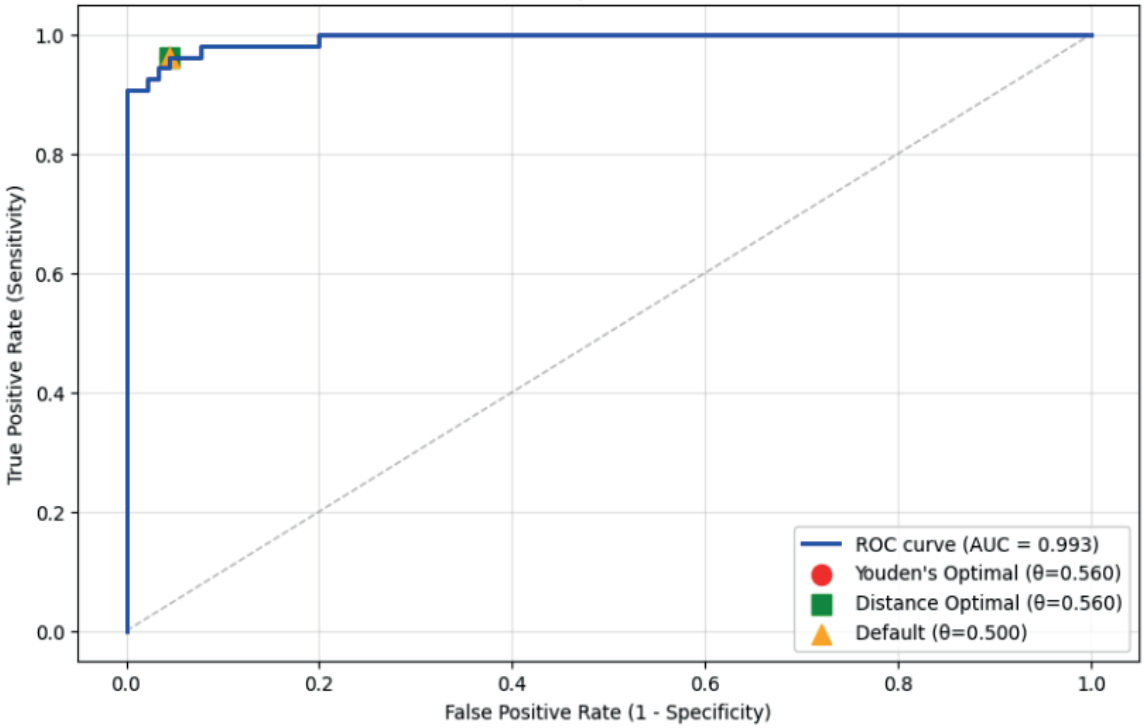


**Figure 4.** ROC Curve with Optimal Threshold Points
**Note:** optimal thresholds identified using Youden's Index (maximizing J = Sensitivity + Specificity - 1), minimum distance to perfect classification point (0,1), and default threshold (0,5). Test set, n=143 samples.

Threshold optimization utilizing Youden's Index determined the optimal decision boundary at 0,560, effectively balancing sensitivity and specificity for clinical decision-making. At this threshold, the model achieved a sensitivity of 96,2 % and a specificity of 95,6 %. Figure 4 illustrates the ROC curve, highlighting the optimal threshold points, including Youden's optimal (0,560), distance optimal (0,560), and default (0,500) thresholds, thereby demonstrating the convergence of multiple optimization criteria at similar threshold values.

Table 2 presents the comprehensive evaluative analysis of the developed logistic regression model, elucidating the diagnostic dynamics that arise with systematic modifications to the decision threshold. This table examines

the model's performance across various threshold values, ranging from 0,3 to 0,7, including the optimal point determined through the Youden index and the minimum distance approach to the ideal ROC point.

**Table 2.** Performance Across Multiple Decision Thresholds

| Threshold | Sensitivity | Specificity | Precision | F1-Score | Accuracy | FN | FP |
|---|---|---|---|---|---|---|---|
| 0,300 | 96,2 % | 92,2 % | 87,9 % | 91,9 % | 93,7 % | 2 | 7 |
| 0,400 | 96,2 % | 93,3 % | 89,5 % | 92,7 % | 94,4 % | 2 | 6 |
| 0,500 | 96,2 % | 95,6 % | 92,7 % | 94,4 % | 95,8 % | 2 | 4 |
| 0,560 | 96,2 % | 95,6 % | 92,7 % | 94,4 % | 95,8 % | 2 | 4 |
| 0,600 | 92,5 % | 96,7 % | 94,2 % | 93,3 % | 95,1 % | 4 | 3 |
| 0,700 | 90,6 % | 100,0 % | 100,0 % | 95,0 % | 96,5 % | 5 | 0 |

**Note:** performance metrics calculated across six probability thresholds using the test set (n=143 samples: 53 malignant, 90 benign). FN = False Negatives; FP = False Positives

At lower threshold values, such as 0,3, the sensitivity of the test reaches a peak of 96,2 %, although this comes at the cost of a significant reduction in specificity, leading to an increased rate of false positives. This configuration is advantageous in preliminary screening settings where the priority is on early detection rather than precise accuracy. In contrast, employing a higher threshold, such as 0,7, results in perfect specificity at 100 %, but this is accompanied by a marked decline in sensitivity, thereby increasing the likelihood of false negatives. Such an approach poses a risk in clinical environments, as it may result in the failure to identify actual cancer cases.

The optimal threshold, approximately 0,56, achieves a balance between sensitivity and specificity, consistently yielding a high F1-score and accuracy. This threshold mathematically maximizes the Youden index (J = Sensitivity + Specificity – 1), a metric frequently employed in epidemiology to ascertain the ideal diagnostic cut-off. Clinically, this table demonstrates that the selection of a threshold is not solely a technical decision but also an ethical and contextual one. For instance, hospitals with limited resources might opt for a more conservative threshold to alleviate the burden of follow-up examinations, whereas national screening centers might adopt a more aggressive threshold to minimize the risk of false negatives.

**Probability Calibration Analysis**

The evaluation of model calibration demonstrated a Brier Score of 0,0298 and an Expected Calibration Error (ECE) of 0,0390. The calibration plot displayed the relationship between predicted probabilities and observed outcomes across probability bins. The distribution of predicted probabilities showed that benign cases had predicted probabilities predominantly in the 0,0-0,1 range, while malignant cases had predicted probabilities concentrated in the 0,9-1,0 range. Few samples exhibited predicted probabilities in the intermediate range of 0,3-0,7.

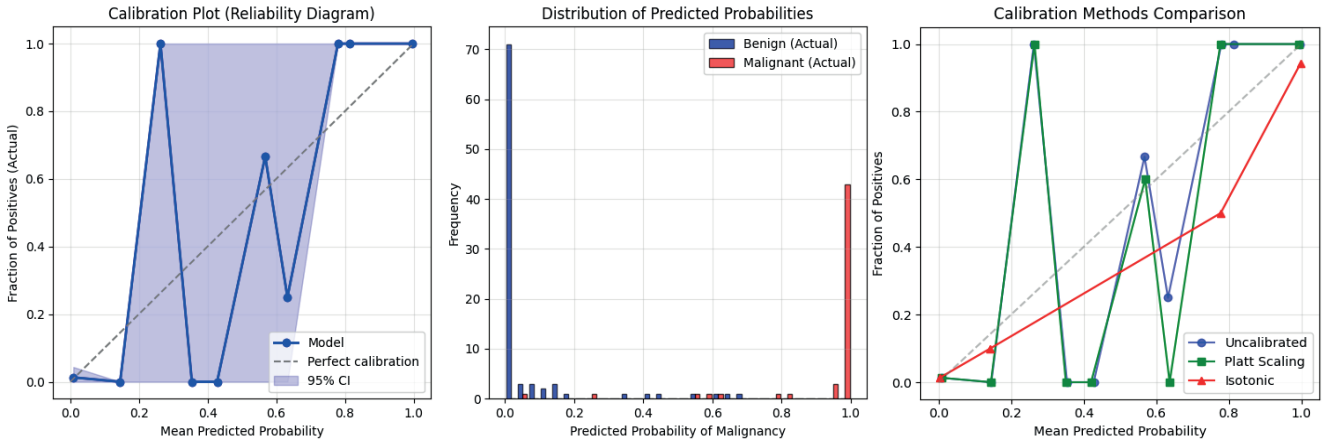


**Figure 5.** Calibration Characteristics

**Note:** Left panel: reliability diagram comparing predicted probabilities (x-axis) with observed frequencies (y-axis) on test set. Middle panel: distribution of predicted probabilities by true class. Right panel: comparison of calibration methods. Test set, n=143 samples

Figure 5 presents the principal calibration characteristics, emphasizing the reliability diagram that contrasts predicted with actual outcomes, the probability distribution patterns across classes, and the comparative

efficacy of various calibration methods. Techniques aimed at enhancing calibration demonstrated modest yet significant improvements in probability reliability.

**Table 3.** Calibration Metrics and Improvement Methods

| Calibration Method | Brier Score | Expected Calibration Error | Assessment |
|---|---|---|---|
| Original Model | 0,0298 | 0,0390 | Well calibrated |
| Platt Scaling | 0,0296 | 0,0378 | Slight improvement |
| Isotonic Regression | 0,0287 | 0,0328 | Best calibration |

**Note:** calibration metrics computed on test set (n=143 samples) for the original logistic regression model and two post-hoc calibration methods

As shown in table 3, isotonic regression emerged as the most effective calibration method, reducing the Expected Calibration Error from 0,0390 to 0,0328 and enhancing the Brier Score from 0,0298 to 0,0287. This non-parametric approach provided superior calibration enhancement compared to Platt scaling, which achieved a more limited improvement, reducing the ECE to 0,0378.

## DISCUSSION

The logistic regression model exhibited performance metrics that surpass conventional benchmarks for medical diagnostic systems. Optimal threshold analysis underscores the significance of balancing sensitivity and specificity in accordance with clinical priorities. The optimal Youden Index threshold of 0,560 offers an ideal mathematical equilibrium; however, clinical implementation necessitates consideration of the relative costs associated with false negatives and false positives. In the context of cancer screening, the clinical repercussions of failing to detect cancer generally surpass the costs of unnecessary follow-up procedures. Threshold analysis indicates that reducing the decision threshold to 0,400 preserves the same sensitivity while offering a more suitable clinical compromise. Conversely, a threshold of 0,700 achieves perfect specificity but results in five undetected cancer cases, which may be deemed clinically unacceptable.

The robust calibration performance significantly enhances the clinical applicability of the model by delivering reliable probability estimates for risk stratification. An Expected Calibration Error below 0,05 suggests that the predicted probabilities can be interpreted as actual risk estimates, thereby facilitating more nuanced clinical decision-making beyond binary classification. This calibration quality compares favorably with existing literature on breast cancer classification systems. A systematic review by Dhanya et al.[46] examining machine learning approaches for breast cancer diagnosis using the WDBC dataset reported that most studies achieved classification accuracies between 93-97 %, similar to our 95,8 %, but fewer than 15 % of reviewed studies assessed probability calibration. Among those that did evaluate calibration, ECE values ranged from 0,042 to 0,089 for logistic regression models, indicating that our ECE of 0,0390 represents above-average calibration quality.[47,48,49]

The calibration improvement achieved through isotonic regression, although modest, underscores the value of post-processing techniques in enhancing probability reliability. Well-calibrated probabilities allow for integration with existing clinical risk assessment frameworks and support personalized patient counseling regarding malignancy risk. The calibration improvement achieved through isotonic regression in our study (ECE reduced from 0,0390 to 0,0328) aligns with findings from Bella et al.[50], who demonstrated that isotonic regression consistently outperforms Platt scaling for logistic regression calibration in medical datasets. However, our improvement magnitude (15,9 % ECE reduction) is more modest than the 23-35 % reductions reported by a study for imbalanced datasets, possibly reflecting that our original model was already reasonably well-calibrated.

Our findings contribute to the growing evidence base supporting interpretable, calibrated models for clinical deployment, as advocated by Christodoulou et al.[51], who found no performance advantage for complex machine learning methods over logistic regression in clinical prediction tasks when considering both discrimination and calibration. The systematic threshold optimization at 0,560 identified in our study provides actionable guidance for clinical implementation, contrasting with studies that report only default threshold (0,5) performance without exploring sensitivity-specificity trade-offs relevant to screening versus diagnostic contexts.[49,52]

### Methodological Considerations, Limitation, and Implications for Clinical Decision

Such exceptional performance metrics may suggest potential overfitting to the specific characteristics of the Wisconsin dataset. Although the single train-test validation approach is computationally efficient, it offers limited evidence regarding the model's robustness across diverse patient cohorts. The lack of cross-validation methodology hinders the assessment of performance consistency and the statistical significance of the reported

metrics. The Wisconsin Diagnostic Breast Cancer dataset, although extensively utilized in machine learning research, constitutes a curated academic dataset that may not entirely reflect the complexity and variability encountered in routine clinical practice. The systematic categorization of features into mean, standard error, and worst value categories, while methodologically robust, may introduce systematic patterns that enhance classification performance but potentially limit generalizability to real-world clinical scenarios.

The performance characteristics indicate potential utility as a diagnostic support tool, particularly in screening applications where high sensitivity is prioritized. However, several factors constrain its immediate readiness for clinical implementation. The exceptional performance metrics necessitate validation on independent clinical datasets from diverse institutions and patient populations to establish generalizability. Furthermore, the absence of feature importance analysis limits clinical interpretability and hinders the identification of the most diagnostically relevant morphological characteristics.

**Study Limitations and Future Research Directions**

The principal limitation of this study is the reliance on a single train-test validation approach, which fails to provide adequate evidence for statistical significance and performance stability. Although the Wisconsin dataset is well-established for research purposes, it may not fully represent the diverse range of cases encountered in various clinical settings. Furthermore, the absence of hyperparameter optimization suggests that the performance may not reflect the optimal configuration for this specific dataset and task. Additionally, the lack of comparison with alternative algorithms restricts the understanding of relative performance advantages and does not offer a baseline for performance assessment.

While the exceptional performance metrics are promising, they necessitate a critical evaluation and substantial further validation prior to consideration for clinical deployment. Despite these limitations, the systematic approach to threshold optimization and probability calibration exhibits methodological sophistication suitable for clinical applications and establishes a foundation for future validation studies with more rigorous experimental designs.

To validate these findings, it is essential to employ rigorous cross-validation methodologies, conduct external validation using independent clinical datasets, and compare the results with established diagnostic tools and alternative machine learning approaches. Conducting a feature importance analysis would enhance clinical interpretability by identifying the most diagnostically relevant morphological characteristics among the 30 available features. The integration of ensemble methods or more sophisticated algorithms could facilitate performance benchmarking and potentially enhance robustness. Developing clinical decision support interfaces that incorporate probability estimates into existing diagnostic workflows represents a logical progression for translating these research findings into practical clinical applications.

## CONCLUSIONS

This study addressed its objective of evaluating the clinical applicability of logistic regression for breast cancer classification by demonstrating that interpretable machine learning models can achieve clinically relevant performance when appropriately optimized and calibrated. The systematic evaluation framework encompassing threshold optimization and probability calibration assessment provides evidence that logistic regression represents a viable approach for clinical decision support in breast cancer diagnosis, particularly in settings where model interpretability and computational efficiency are priorities.

The principal conclusion of this study is that optimizing thresholds using clinically meaningful metrics enables classification decisions to align with institutional priorities concerning the balance between sensitivity and specificity. Additionally, the assessment and refinement of probability calibration techniques ensure that model outputs can be interpreted reliably as estimates of actual risk, rather than arbitrary confidence scores. Furthermore, a comprehensive model evaluation that encompasses discrimination, calibration, and threshold selection offers a more robust foundation for clinical application than traditional assessments that primarily focus on accuracy.

However, important limitations constrain immediate clinical implementation. The reliance on single train-test validation without rigorous cross-validation or external validation limits confidence in performance generalizability across diverse patient populations and clinical settings. The Wisconsin dataset, while methodologically valuable as a research benchmark, represents a curated collection from a single institution during a specific time period and may not fully reflect the morphological variability encountered in contemporary clinical practice. Dataset-specific optimization may contribute to the observed performance levels, necessitating cautious interpretation and mandatory validation on independent clinical cohorts before deployment.

Future research priorities include external validation using multi-institutional datasets representing diverse patient demographics and imaging protocols, implementation of rigorous cross-validation methodologies to establish statistical confidence in performance estimates, prospective clinical evaluation comparing model-assisted diagnosis against standard clinical workflows, and investigation of model performance across patient

subgroups to assess potential disparities. Additionally, development of clinical decision support interfaces that effectively communicate probability estimates and threshold-dependent trade-offs to clinicians represents an essential step toward translating research findings into practical diagnostic tools.

## BIBLIOGRAPHIC REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6).

2. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistics, 2019. CA Cancer J Clin. 2019;69(6).

3. Al Muhaisen S, Safi O, Ulayan A, Aljawamis S, Fakhoury M, Baydoun H, et al. Artificial Intelligence-Powered Mammography: Navigating the Landscape of Deep Learning for Breast Cancer Detection. Cureus. 2024 Mar 26;

4. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3).

5. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. Vol. 33, Medical Image Analysis. 2016.

6. Evans AJ, Brown RW, Bui MM, Chlipala EA, Lacchetti C, Milner DA, et al. Validating Whole Slide Imaging Systems for Diagnostic Purposes in Pathology. Arch Pathol Lab Med. 2022;146(4).

7. Krane JF. Koss' Diagnostic Cytology and Its Histopathologic Bases, Fifth Edition. International Journal of Gynecological Pathology. 2007;26(3).

8. Ramos-Vara JA, Miller MA. When Tissue Antigens and Antibodies Get Along: Revisiting the Technical Aspects of Immunohistochemistry-The Red, Brown, and Blue Technique. Vol. 51, Veterinary Pathology. 2014.

9. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Vol. 28, Nature Medicine. 2022.

10. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Vol. 25, Nature Medicine. 2019.

11. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788).

12. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020;26(6).

13. Samsir S, Sitorus JHP, Zulkifli, Ritonga Z, Nasution FA, Watrianthos R. Comparison of machine learning algorithms for chest X-ray image COVID-19 classification. J Phys Conf Ser. 2021;1933(1):012040.

14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Vol. 25, Nature Medicine. 2019.

15. Jimma BL. Artificial intelligence in healthcare: A bibliometric analysis. Vol. 9, Telematics and Informatics Reports. 2023.

16. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: Past, present and future. Vol. 2, Stroke and Vascular Neurology. 2017.

17. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Vol. 1, Nature Machine Intelligence. 2019.

18. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Vol. 3, The Lancet Digital Health. 2021.

19. Beam AL, Kohane IS. Big data and machine learning in health care. Vol. 319, JAMA - Journal of the American Medical Association. 2018.

20. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. New England Journal of Medicine. 2018;378(11).

21. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. Vol. 17, BMC Medicine. 2019.

22. Naylor CD. On the prospects for a (Deep) learning health care system. Vol. 320, JAMA - Journal of the American Medical Association. 2018.

23. Scott AJ, Hosmer DW, Lemeshow S. Applied Logistic Regression. Biometrics. 1991;47(4).

24. Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. Am Heart J. 1991;121(1 PART 2).

25. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: 34th International Conference on Machine Learning, ICML 2017. 2017.

26. Minderer M, Djolonga J, Romijnders R, Hubis F, Zhai X, Houlsby N, et al. Revisiting the Calibration of Modern Neural Networks. In: Advances in Neural Information Processing Systems. 2021.

27. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making. 2006;26(6).

28. Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A, Subbarayalu AV, et al. A Review of the Role of Artificial Intelligence in Healthcare. Vol. 13, Journal of Personalized Medicine. 2023.

29. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Vol. 42, Japanese Journal of Radiology. 2024.

30. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ (Online). 2016;352.

31. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. Vol. 3, npj Digital Medicine. 2020.

32. López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C, Gude-Sampedro F. Optimalcutpoints: An R package for selecting optimal cutpoints in diagnostic tests. J Stat Softw. 2014;61(8).

33. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol. 2006;163(7).

34. Wu N, Phang J, Park J, Shen Y, Kim SG, Heacock L, et al. Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle. Kaggle. 2019;4(November).

35. W.N. S, W.H. W, O.L. M. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology. 1993;1905(870).

36. Alshayeji MH, Ellethy H, Abed S, Gupta R. Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach. Biomed Signal Process Control. 2022;71.

37. Panda NR, Pati JK, Mohanty JN, Bhuyan R. A Review on Logistic Regression in Medical Research. Vol. 13, National Journal of Community Medicine. 2022.

38. Prasad R, Anjali P, Adil S, Deepa N. Heart disease prediction using logistic regression algorithm using machine learning. Int J Eng Adv Technol. 2019;8(3 Special Issue).

39. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. Vol. 38, Journal of Biomedical Informatics. 2005.

40. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem. 1993;39(4).

41. Posocco N, Bonnefoy A. Estimating Expected Calibration Errors. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021.

42. Böken B. On the appropriateness of Platt scaling in classifier calibration. Inf Syst. 2021;95.

43. Huang L, Zhao J, Zhu B, Chen H, Broucke S Vanden. An Experimental Investigation of Calibration Techniques for Imbalanced Data. IEEE Access. 2020;8.

44. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med. 1997;16(9).

45. Cox DR. Two Further Applications of a Model for Binary Regression. Biometrika. 1958;45(3/4).

46. Dhanya R, Paul IR, Sindhu Akula S, Sivakumar M, Nair JJ. A comparative study for breast cancer prediction using machine learning and feature selection. In: 2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019. 2019.

47. LG A, AT E. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. J Health Med Inform. 2013;04(02).

48. Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Comput Appl. 2017;28(4).

49. Salama Gouda I, .Abdelhalim M.B., Zeid Magdy Abd-elghany. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. International Journal of Computer and Information Technology. 2012;01(01).

50. Bella A, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ. Quantification via probability estimators. In: Proceedings - IEEE International Conference on Data Mining, ICDM. 2010.

51. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Vol. 110, Journal of Clinical Epidemiology. 2019.

52. Asri H, Mousannif H, Al Moatassime H, Noel T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. In: Procedia Computer Science. 2016.

## FINANCING

This research was funded by the Directorate General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the 2025 Fundamental Research Grant Program. The authors express their sincere gratitude for the support provided.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

## AUTHORSHIP CONTRIBUTION

*Conceptualization:* Ronal Watranthos.
*Data curation:* Rayendra.
*Formal analysis:* Rayendra.
*Research:* Ronal Watranthos.
*Methodology:* Yuhefizar.
*Project management:* Yuhefizar.
*Resources:* Ervan Asri.
*Software:* Ervan Asri.

*Supervision:* Ronal Watranthos.
*Validation:* Yuhefizar.
*Display:* Humaira.
*Drafting - original draft:* Rayendra.
*Writing - proofreading and editing:* Humaira.