Leveraging Machine Learning Algorithms for Early Detection of Breast Cancer: A Comparative Study Using Diagnostic Features

Aulia Abikhair[1], Huang Guanghui[2,*]

[1]*Teacher Education Study Program, School of Postgraduate Studies, Universitas Pendidikan Indonesia, Bandung, Indonesia*
[2]*Department of Educational Administration, Faculty of Education, Universitas Pendidikan Indonesia, Bandung, Indonesia*

(Received June 7, 2025; Revised October 10, 2025; Accepted January 25, 2026; Available online March 28, 2026)

Abstract

Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide, making early and accurate diagnosis essential for improving survival rates and treatment outcomes. To address limitations associated with conventional diagnostic methods, Machine Learning (ML) techniques have been increasingly adopted to enhance classification accuracy and reduce diagnostic variability. This study presents a comparative evaluation of four widely used ML algorithms Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression applied to a structured breast cancer diagnostic dataset. The dataset comprises morphological and texture-based features extracted from digitized tumor samples, enabling binary classification of benign and malignant cases. The models were trained using an 80:20 train–test split and validated through k-fold cross-validation. Performance evaluation was conducted using accuracy, precision, recall, F1-score, and confusion matrix analysis to ensure comprehensive assessment of classification behavior. Experimental results indicate strong predictive performance across all models, with overall accuracy values ranging from 0.95 to 0.96. Among the evaluated approaches, Random Forest demonstrated the most balanced performance, particularly achieving the highest recall for malignant tumors and the lowest false-negative rate, which is critical in clinical diagnostics. Feature importance analysis further revealed that tumor area, concave points, radius, and perimeter were the most influential predictors in classification decisions, consistent with established clinical indicators of malignancy. These findings confirm that classical and interpretable machine learning algorithms, especially ensemble-based methods, remain highly effective for structured breast cancer classification tasks. The study contributes to the advancement of reliable and transparent ML-based decision-support systems, supporting improved early detection and diagnostic accuracy in breast cancer care.

Keywords: Breast Cancer Detection; Machine Learning; Random Forest; Tumor Classification; Feature Importance

1. Introduction

Breast cancer remains one of the most prevalent and deadly diseases among women worldwide, with approximately 2.3 million new cases and over 685,000 deaths reported in 2020 [1], [2]. Early detection is crucial for improving survival rates and treatment outcomes. However, many cases are still diagnosed at advanced stages due to limited awareness, inadequate access to screening, and socio-cultural barriers [3]. Although strategies such as Breast Self-Examination (BSE) have demonstrated effectiveness in promoting early detection behaviors, knowledge gaps regarding routine screening persist among many women [4]. Strengthening educational initiatives and improving diagnostic technologies are therefore essential to reduce mortality and improve patient prognosis.

Conventional breast cancer detection methods, including mammography and ultrasound imaging, rely heavily on subjective interpretation by radiologists. This subjectivity may result in diagnostic variability, false positives, or false negatives, potentially leading to delayed treatment or unnecessary psychological distress [5]. Moreover, manual diagnostic procedures are time-consuming and highly dependent on individual expertise, which may limit consistency across healthcare settings [6]. These limitations highlight the need for automated, data-driven approaches capable of improving diagnostic accuracy and efficiency.

***Corresponding author: Huang Guanghui (huangguanghui@upi.edu)**
DOI: https://doi.org/10.47738/ijiis.v9i2.298

 ISSN 2579-7069

In recent years, the application of Machine Learning (ML) has significantly transformed medical diagnostics, particularly in cancer detection. ML algorithms are capable of analyzing complex medical data, identifying subtle patterns, and producing reliable classifications with reduced human bias [7]. Various studies have demonstrated the effectiveness of ML techniques in distinguishing between benign and malignant tumors, showing promising results in improving predictive accuracy [8]. However, comparative evaluations of classical machine learning classifiers on structured diagnostic datasets remain important to determine the most effective and computationally efficient approach for clinical support systems.

This study focuses specifically on four widely used and interpretable machine learning algorithms: Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression. These algorithms were selected due to their strong theoretical foundations, proven effectiveness in classification tasks, and suitability for structured medical datasets. Random Forest offers robustness through ensemble learning and feature importance analysis; KNN provides instance-based classification using distance metrics; SVM constructs optimal decision boundaries in high-dimensional spaces; and Logistic Regression serves as a statistical baseline model with strong interpretability. By comparing these four algorithms, this research aims to identify the most effective method for classifying breast cancer tumors as benign or malignant using diagnostic feature data.

The primary objective of this study is therefore to conduct a comparative performance evaluation of Random Forest, KNN, SVM, and Logistic Regression in breast cancer tumor classification. By systematically assessing accuracy, precision, recall, and F1-score, this research seeks to determine which algorithm provides the most reliable diagnostic performance. Ultimately, the findings are expected to contribute to improving early breast cancer detection through evidence-based integration of machine learning techniques, supporting more accurate, efficient, and objective diagnostic practices in oncology.

2. Literature Review

2.1. Breast Cancer Detection

Early detection of breast cancer is a critical determinant of survival rates and treatment success. As one of the most commonly diagnosed cancers among women worldwide, breast cancer continues to pose a significant public health challenge, with early-stage identification substantially improving prognosis and reducing treatment complexity [9]. Conventional diagnostic methods such as mammography, ultrasound, Magnetic Resonance Imaging (MRI), and Computed Tomography (CT) play essential roles in screening and tumor evaluation. Mammography remains the standard screening modality but shows reduced sensitivity in dense breast tissue and involves radiation exposure [10]. Ultrasound serves as a complementary tool, particularly for dense breasts, yet its performance is highly operator-dependent [11]. MRI provides high sensitivity but is costly and may produce false positives, while CT and PET imaging are primarily used for staging rather than early detection [12]. These limitations including variability in sensitivity and specificity, operator dependence, false-positive outcomes, and accessibility constraints underscore the need for more objective and efficient diagnostic approaches.

In response to these challenges, ML techniques have emerged as powerful tools for enhancing breast cancer detection and classification. ML algorithms enable the analysis of complex medical data, including imaging and structured diagnostic features, thereby reducing human bias and improving predictive accuracy [13]. Ensemble and classification-based models have demonstrated promising performance in distinguishing benign from malignant tumors, particularly when effective feature selection strategies are applied [14]. Computer-aided diagnostic systems powered by ML further support clinical decision-making by improving consistency and reducing diagnostic variability. Despite the growing use of deep learning in imaging-based detection, there remains a need for systematic comparative evaluation of classical and interpretable machine learning algorithms on structured diagnostic datasets. Models such as Random Forest, KNN, SVM, and Logistic Regression offer strong theoretical foundations, computational efficiency, and explain ability, making them suitable candidates for clinical decision-support systems. Therefore, evaluating and comparing these algorithms is essential to determine the most reliable and practical approach for breast cancer tumor classification and to strengthen the integration of ML into early detection strategies.

2.2. Machine Learning Algorithms for Cancer Classification

Several machine learning algorithms have been widely applied in breast cancer classification to distinguish benign from malignant tumors, including SVM, Random Forest, KNN, and Logistic Regression. SVM is particularly effective in high-dimensional datasets and constructs optimal hyperplanes to separate classes, demonstrating strong classification performance in complex medical data [15], [16]. However, its effectiveness depends on appropriate kernel selection and parameter tuning, which may complicate implementation. Random Forest, an ensemble-based learning method, aggregates multiple decision trees to enhance robustness and reduce overfitting, frequently outperforming other classifiers in noisy and high-dimensional datasets [6], [17]. Additionally, its ability to provide feature importance supports interpretability in clinical decision-making. KNN, a simple instance-based algorithm, classifies samples based on proximity in feature space and has shown competitive performance in tumor classification [17], [18]. Nevertheless, it becomes computationally expensive during prediction and is sensitive to high dimensionality. Logistic Regression is commonly employed as a benchmark model due to its simplicity, probabilistic output, and interpretability [15], [18], [19]. Although it assumes linear relationships between predictors and outcomes, it remains valuable in clinical contexts where transparency and computational efficiency are essential.

Comparative analyses in previous studies reveal trade-offs among these algorithms in terms of accuracy, interpretability, and computational cost. Random Forest often achieves superior predictive performance across multiple evaluation metrics, while SVM performs well in high-dimensional feature spaces [6], [15]. KNN provides simplicity but may struggle with scalability, and Logistic Regression serves as a reliable baseline for performance comparison [18], [19]. Therefore, selecting the most appropriate model depends on dataset characteristics, feature complexity, and the balance between predictive accuracy and interpretability required in breast cancer diagnostic systems.

2.3. Feature Selection in Breast Cancer Classification

Feature selection is a fundamental step in machine learning that aims to identify the most relevant variables from a larger set of features to improve predictive performance and model efficiency [20]. In cancer detection, effective feature selection significantly enhances classification accuracy while reducing computational complexity and minimizing the risk of overfitting [21]. By reducing dimensionality, models become more generalized and perform better on unseen data, as irrelevant or noisy variables are eliminated [22]. Moreover, selecting a smaller subset of meaningful features improves interpretability, which is crucial in healthcare contexts where clinicians must understand and justify diagnostic decisions [23]. In addition, reduced feature space leads to shorter training times and faster model deployment, an important factor in time-sensitive clinical environments [24].

In breast cancer detection, several diagnostic features have been consistently identified as highly indicative of malignancy. Tumor radius and perimeter are strongly associated with tumor size and boundary irregularity, both of which correlate with malignant behavior [9], [25]. Texture features capture variations in pixel intensity and structural heterogeneity, providing essential information for differentiating benign and malignant growths [26]. Tumor area is similarly significant, as larger surface areas are often linked to aggressive cancer types [7], [9]. Additional morphological descriptors such as smoothness, compactness, and symmetry further contribute to classification performance by reflecting tumor shape and structural consistency [24], [27].

Comparative studies emphasize the substantial impact of feature selection on model effectiveness. Random Forest models incorporating selected features have demonstrated improved predictive accuracy by eliminating redundant variables [28]. Similarly, SVM classifiers trained on carefully selected features have shown superior performance compared to models using the full feature set [15]. Logistic Regression, often employed as a baseline model, is particularly sensitive to feature selection, with performance improvements observed when relevant predictors are properly chosen [21], [22]. Overall, the integration of effective feature selection techniques enhances the accuracy, interpretability, and efficiency of machine learning models in breast cancer classification, thereby supporting more reliable early diagnosis and improved patient care.

2.4. Challenges in Using Machine Learning Models

Although ML models have demonstrated strong potential in breast cancer diagnosis, several critical challenges limit their clinical reliability and applicability. One major issue concerns data quality, as medical datasets often contain

noise, missing values, and inconsistencies that can distort learning processes and reduce model generalizability [29], [30]. When training data fail to accurately represent the broader patient population, predictive outcomes may become biased, increasing the risk of misdiagnosis. Class imbalance is another prevalent challenge in healthcare datasets, where benign cases frequently outnumber malignant ones, leading algorithms to favor the majority class and reduce sensitivity in detecting critical minority outcomes [31]. Although resampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE) have been proposed to mitigate imbalance, they may introduce overfitting or generate ambiguous synthetic samples [29], [32].

Generalization capability also remains a concern, as complex models may overfit training data and fail to perform consistently on unseen datasets or diverse patient populations [33]. Beyond technical limitations, ethical considerations play a crucial role in ML-based healthcare applications. Algorithmic bias resulting from unrepresentative training data can produce inequitable outcomes across demographic groups [34]. Furthermore, the limited interpretability of certain ML models complicates clinical trust and decision justification [35], while concerns regarding patient data privacy must be carefully managed during model development and deployment. Collectively, these challenges underscore the need for robust data preprocessing, balanced training strategies, rigorous validation, and transparent modeling approaches to ensure safe and reliable integration of ML systems into breast cancer diagnostics.

3. Methodology

3.1. Data Collection

This study utilizes the Breast Cancer Diagnostic Dataset, a widely used benchmark dataset for tumor classification research. The dataset contains diagnostic measurements derived from digitized images of Fine Needle Aspirate (FNA) of breast masses. Each record represents a tumor sample labeled as either benign or malignant, enabling binary classification analysis. The dataset is structured and numerical, making it suitable for evaluating classical machine learning algorithms such as Random Forest, KNN, SVM, and Logistic Regression.

The dataset consists of multiple quantitative features computed from cell nuclei characteristics. These features include morphological and texture-based measurements such as Mean Radius, Mean Texture, Mean Perimeter, Mean Area, Mean Smoothness, Mean Compactness, Mean Concavity, Mean Concave Points, Mean Symmetry, and Mean Fractal Dimension. In addition to mean values, the dataset also provides corresponding standard error and worst (maximum) values for each characteristic, resulting in a comprehensive representation of tumor geometry and structural irregularity. Such features capture critical information related to tumor size, shape, boundary irregularity, and internal structure, all of which are clinically relevant indicators of malignancy. The combination of these statistical descriptors provides a robust foundation for training and evaluating classification models in breast cancer diagnosis research.

3.2. Data Preprocessing

Data preprocessing was conducted to ensure data quality, consistency, and suitability for machine learning modeling. The initial step involved removing non-informative attributes, specifically the ID column, as it serves only as a unique identifier and does not contribute to predictive learning. Retaining such attributes may introduce noise without improving model performance.

Next, the dataset was examined for missing values and inconsistencies. A completeness check confirmed that no missing values were present, eliminating the need for imputation techniques. However, data validation procedures were performed to ensure numerical consistency and proper formatting of all feature variables. Ensuring data integrity at this stage is critical, as poor data quality can negatively affect model generalization and predictive reliability.

The target variable, Diagnosis, originally labeled as categorical values (“Benign” and “Malignant”), was transformed into numerical format to facilitate algorithmic processing. Binary encoding was applied, where malignant tumors were assigned a value of 1 and benign tumors were assigned a value of 0. This transformation enables compatibility with classification algorithms such as Random Forest, KNN, SVM, and Logistic Regression.

Additionally, feature scaling was applied to standardize the input variables, particularly for distance-based and margin-based algorithms such as KNN and SVM, which are sensitive to differences in feature magnitude. Standardization

ensures that all features contribute proportionally during model training and prevents bias toward variables with larger numerical ranges.

Through these preprocessing steps removal of irrelevant attributes, validation of data integrity, label encoding, and feature scaling the dataset was prepared to ensure robust and reliable model training and evaluation.

3.3. Machine Learning Algorithms for Cancer Classification

In breast cancer diagnostics, several machine learning algorithms have demonstrated strong performance in classifying tumors as benign or malignant, including Random Forest, SVM, KNN, and Logistic Regression. These algorithms were selected due to their complementary strengths in predictive accuracy, robustness, interpretability, and suitability for structured medical datasets.

Random Forest is an ensemble learning method that aggregates multiple decision trees constructed from random subsets of training samples and features [36]. The final prediction is obtained through majority voting:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\} \quad (1)$$

where $T_i(x)$ represents the prediction from the i-th decision tree. Random Forest is widely recognized for reducing variance and overfitting while maintaining high predictive performance [37]. Additionally, it provides feature importance measures, supporting clinical interpretability [38].

SVM constructs an optimal separating hyperplane by maximizing the margin between classes [38]. The decision boundary is defined as:

$$f(x) = w^T x + b \quad (2)$$

where $w$is the weight vector and b is the bias term. The optimization objective is:

$$\min_{w,b} \frac{1}{2} || w ||^2 \quad (3)$$

subject to:

$$y_i(w^T x_i + b) \geq 1 \quad (4)$$

SVM performs particularly well in high-dimensional feature spaces and is robust to overfitting through margin maximization.

KNN is a distance-based, instance-based learning algorithm that classifies a new sample based on the majority class among its $k$nearest neighbors [37]. Using Euclidean distance, the metric is defined as:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (5)$$

KNN is simple and intuitive, making it a valuable benchmark model [38]. However, it is sensitive to high dimensionality and computationally expensive during prediction.

Logistic Regression is a probabilistic binary classifier that models the log-odds of class membership as a linear combination of features [38]. The logistic function is defined as:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \quad (6)$$

where β represents model coefficients. Logistic Regression is highly interpretable and frequently used as a baseline model in medical studies.

Comparative studies indicate that Random Forest and SVM typically achieve superior accuracy, particularly in high-dimensional datasets [37], [38]. Random Forest has reported accuracy up to 98.83%, followed by SVM at

 ISSN 2579-7069

approximately 97.5%. KNN often achieves competitive but slightly lower performance (around 90%), while Logistic Regression serves as a reliable and interpretable baseline, commonly reaching around 93% accuracy.

Overall, the selection of these four algorithms is justified by their demonstrated robustness, theoretical diversity, and complementary strengths in handling structured breast cancer diagnostic data. Their comparative evaluation provides a balanced assessment between predictive performance and clinical interpretability.

3.5. Model Training and Evaluation

To ensure reliable performance assessment, the dataset was divided into training and testing subsets using an 80:20 ratio, where 80% of the data were used for model training and 20% for independent testing. This split enables the evaluation of model generalization on unseen data. The training set was used to learn model parameters for Random Forest, SVM, KNN, and Logistic Regression, while the testing set was reserved exclusively for performance validation. In addition to the hold-out method, k-fold cross-validation was applied during training to enhance robustness and reduce variance in performance estimation. In k-fold cross-validation, the dataset is partitioned into kequal subsets, and the model is trained and validated iteratively such that each subset serves as validation data once. The final performance is computed as the average across all folds:

$$CV\ Score = \frac{1}{k}\sum_{i=1}^{k} Scor\ e_i \tag{7}$$

Model performance was evaluated using standard classification metrics derived from the confusion matrix, which consists of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). From these values, several evaluation metrics were calculated:

Accuracy, representing overall correctness:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Precision, measuring the proportion of correctly predicted positive cases:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall (Sensitivity), indicating the model's ability to correctly identify malignant cases:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

F1-Score, the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

The confusion matrix was further analyzed to assess classification distribution and misclassification patterns, particularly focusing on false negatives, which are critical in medical diagnostics. By combining cross-validation with multiple evaluation metrics, the study ensures a comprehensive assessment of model robustness, predictive reliability, and clinical applicability.

3.6. Tools and Frameworks for Breast Cancer Detection Using Machine Learning

The implementation of machine learning models for breast cancer detection relies on several widely adopted programming tools and frameworks that support data processing, model development, and evaluation. Among these, Python has emerged as a dominant programming language in data science due to its readability, flexibility, and extensive ecosystem of libraries. Core libraries such as NumPy and Pandas facilitate efficient numerical computation and structured data manipulation, while Matplotlib and Seaborn enable comprehensive visualization of patterns and model outputs. Most importantly, Scikit-learn provides optimized implementations of classification algorithms including Random Forest, SVM, KNN, and Logistic Regression, along with built-in utilities for preprocessing, feature

selection, and cross-validation [36]. Python's strong community support and seamless integration with advanced frameworks further enhance its suitability for medical machine learning research [39].

Scikit-learn, in particular, plays a central role in structured breast cancer classification tasks. Its standardized API design simplifies model training, evaluation, and hyperparameter tuning, while its cross-validation and pipeline features enable reproducible and efficient workflows. The library's optimized implementations ensure reliable performance even with moderately large datasets, making it suitable for structured diagnostic datasets.

For deep learning applications, especially image-based classification, TensorFlow and Keras provide scalable and flexible frameworks for building neural networks. These frameworks support Convolutional Neural Networks (CNNs) and other advanced architectures increasingly used in medical imaging tasks [40]. TensorFlow's scalability allows deployment across various hardware configurations, from local machines to cloud environments, while Keras offers high-level abstractions that simplify neural network construction [41].

Additionally, R remains relevant in medical research due to its strong statistical capabilities. Packages such as caret and randomForest support predictive modeling, feature selection, and model validation, while visualization libraries like ggplot2 enhance exploratory data analysis [42]. R is particularly advantageous when statistical interpretability and rigorous hypothesis testing are required

4. Results

4.1. Model Performance

The evaluation metrics include accuracy, precision, recall, F1-score, as well as macro and weighted averages derived from the confusion matrix. The detailed results are summarized in Table 1.

Table 1. Performance Comparison of Machine Learning Models

Model	Accuracy	Precision (Malignant)	Recall (Malignant)	F1-Score (Malignant)	Macro Avg F1	Weighted Avg F1
Random Forest	0.96	0.98	0.93	0.95	0.96	0.96
Support Vector Machine	0.95	1.00	0.86	0.93	0.94	0.95
K-Nearest Neighbors	0.96	1.00	0.88	0.94	0.95	0.96
Logistic Regression	0.96	0.97	0.91	0.94	0.95	0.96

As shown in Table 1, all models achieved strong predictive performance, with accuracy values ranging from 0.95 to 0.96. Random Forest achieved an overall accuracy of 0.96, with a precision of 0.98 and recall of 0.93 for the malignant class, resulting in an F1-score of 0.95. The macro-average and weighted-average F1-scores were both 0.96, indicating stable classification performance across both classes.

The SVM model achieved an accuracy of 0.95. It obtained a perfect precision score of 1.00 for malignant tumors, indicating no false positive predictions. However, its recall was slightly lower at 0.86, suggesting that some malignant cases were misclassified as benign. The F1-score for malignant classification was 0.93, with a weighted-average F1-score of 0.95.

KNN demonstrated comparable performance, achieving an accuracy of 0.96. Similar to SVM, KNN achieved a precision of 1.00 for malignant cases, while recall reached 0.88. The F1-score for malignant classification was 0.94, and the weighted-average F1-score was 0.96. Although competitive, the slightly lower recall compared to Random Forest indicates a marginally higher false-negative rate.

Logistic Regression also achieved an overall accuracy of 0.96. The model produced a precision of 0.97 and recall of 0.91 for malignant tumors, resulting in an F1-score of 0.94. Both macro-average and weighted-average F1-scores were 0.95 and 0.96, respectively, demonstrating consistent performance despite the model's simpler linear structure.

 ISSN 2579-7069

From a clinical standpoint, recall (sensitivity) for the malignant class is particularly critical, as false negatives may delay treatment. Among the evaluated models, Random Forest achieved the highest recall (0.93) for malignant detection, indicating a stronger ability to correctly identify cancerous cases. While SVM and KNN achieved perfect precision, their lower recall values suggest a more conservative classification behavior.

Overall, the results indicate that although all models performed well, Random Forest demonstrated the most balanced performance across evaluation metrics, making it a robust candidate for structured breast cancer diagnostic classification.

4.2. Confusion Matrix

To provide a detailed evaluation of classification behavior, the confusion matrices of each model were analyzed to examine the distribution of correct and incorrect predictions for both benign and malignant classes. The confusion matrix offers granular insight into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which are critically important in medical diagnostic applications.

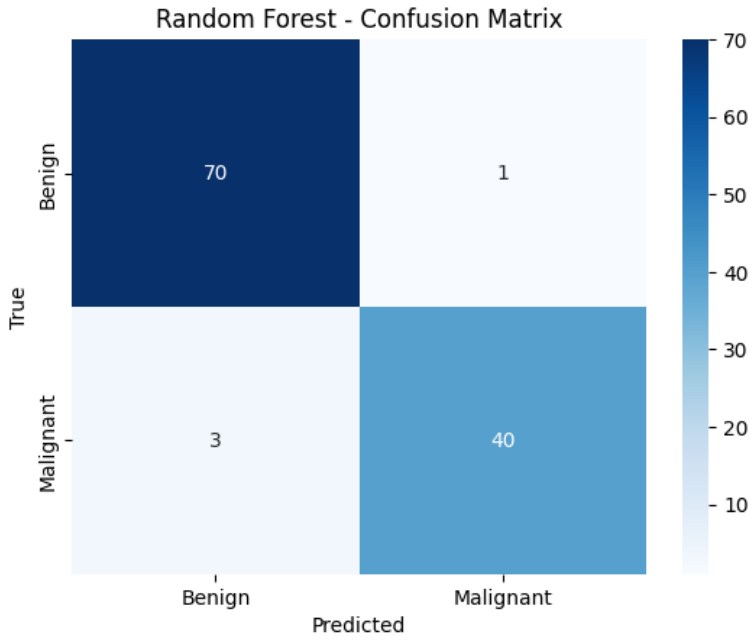

Figure 1. Confusion Matrix of Random Forest Model

As shown in Figure 1, the Random Forest model correctly classified 70 benign cases (TN = 70) and 40 malignant cases (TP = 40). Only 1 benign case was incorrectly predicted as malignant (FP = 1), while 3 malignant cases were misclassified as benign (FN = 3). The relatively low number of false negatives demonstrates strong sensitivity, which is particularly crucial in breast cancer diagnosis where missed malignant cases may delay treatment.

 ISSN 2579-7069

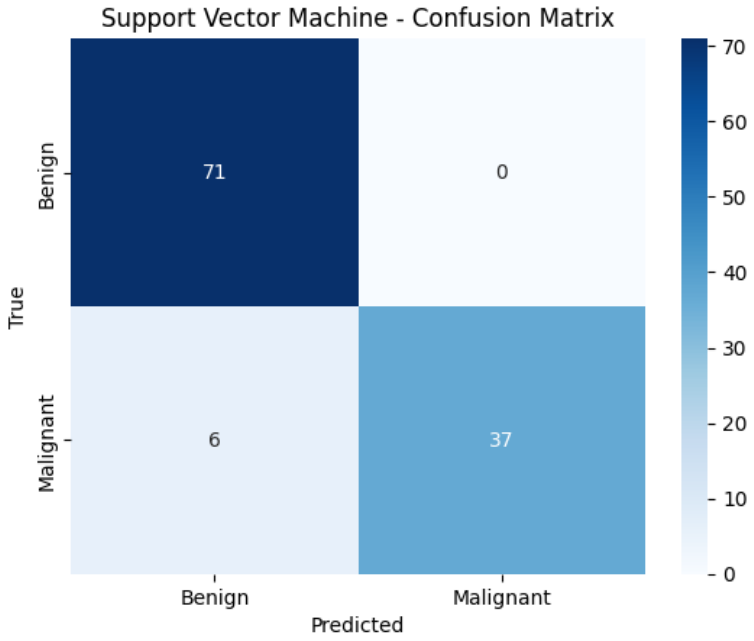

Figure 2. Confusion Matrix of SVM Model

In Figure 2, the SVM model correctly classified all benign cases (TN = 71) with zero false positives (FP = 0), indicating perfect specificity for the benign class. However, it misclassified 6 malignant cases as benign (FN = 6), correctly identifying 37 malignant cases (TP = 37). Although SVM achieved excellent precision for malignant predictions, the higher number of false negatives suggests slightly lower sensitivity compared to Random Forest.

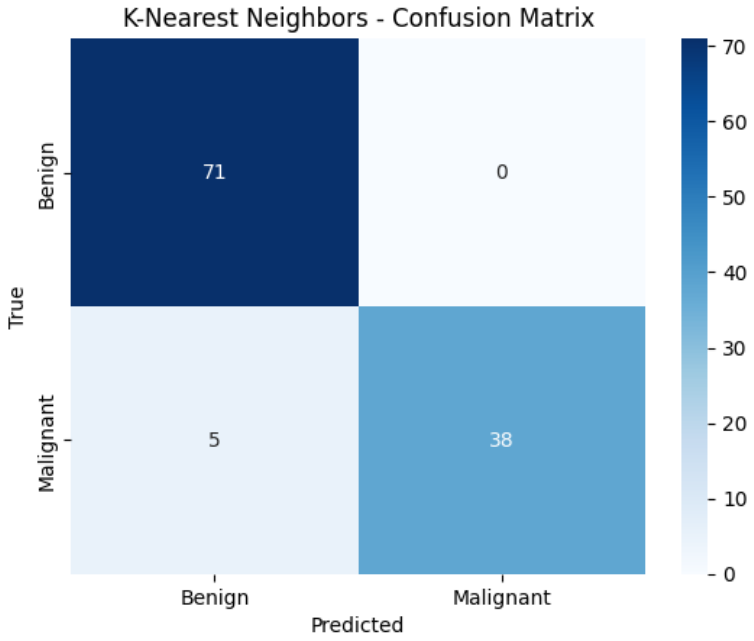

Figure 3. Confusion Matrix of KNN Model

Similarly, the KNN confusion matrix shown in Figure 3 indicates that all benign cases were correctly identified (TN = 71, FP = 0). The model correctly detected 38 malignant cases (TP = 38) but misclassified 5 malignant cases as benign (FN = 5). While KNN demonstrated strong specificity, its false-negative count remained higher than that of Random Forest, reflecting a moderate sensitivity gap.

 ISSN 2579-7069

The Logistic Regression confusion matrix presented in Figure 4 shows 70 correctly classified benign cases (TN = 70) and 39 correctly identified malignant cases (TP = 39). The model produced 1 false positive (FP = 1) and 4 false negatives (FN = 4), reflecting balanced classification behavior between sensitivity and specificity.

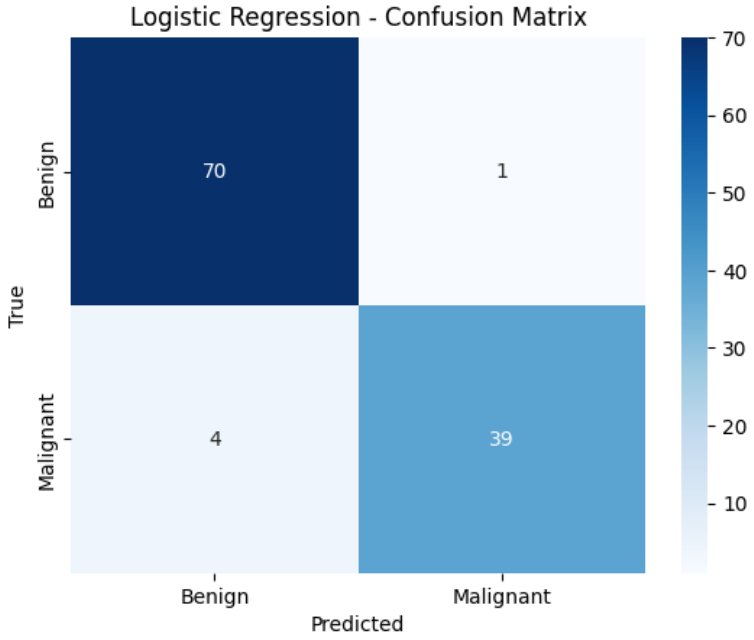

Figure 4. Confusion Matrix of Logistic Regression Model

From a clinical perspective, minimizing false negatives is paramount because classifying malignant tumors as benign poses a direct risk to patient safety. Among the evaluated models, Random Forest produced the lowest number of false negatives (FN = 3), followed by Logistic Regression (FN = 4), KNN (FN = 5), and SVM (FN = 6). Although SVM and KNN achieved perfect specificity (no false positives), their higher false-negative rates indicate a more conservative classification tendency.

Overall, the confusion matrix analysis confirms that while all models demonstrate high diagnostic capability, Random Forest provides the most balanced performance by maintaining both high sensitivity and high specificity. This balance enhances its suitability as a reliable decision-support model for structured breast cancer diagnostic data.

4.3. Comparative Analysis

This section presents a comparative evaluation of the four machine learning algorithms Random Forest, SVM, KNN, and Logistic Regression based on accuracy, precision, recall, F1-score, and confusion matrix analysis. The objective is to determine which model performs most effectively in identifying malignant and benign tumors.

In terms of overall accuracy, Random Forest, KNN, and Logistic Regression achieved the highest performance (0.96), while SVM followed closely with 0.95. Although the difference in accuracy is marginal, further analysis using class-specific metrics reveals meaningful distinctions in clinical relevance.

When focusing on the malignant class, which is the most critical category in breast cancer diagnosis, Random Forest achieved the highest recall (0.93), followed by Logistic Regression (0.91), KNN (0.88), and SVM (0.86). Recall (sensitivity) measures the model's ability to correctly identify malignant tumors, and higher recall reduces the likelihood of false negatives. Since false negatives represent missed cancer cases, Random Forest demonstrated superior performance in identifying malignant tumors.

Regarding precision for malignant cases, SVM and KNN achieved perfect precision (1.00), meaning they did not incorrectly classify benign tumors as malignant. Random Forest and Logistic Regression also showed high precision values (0.98 and 0.97, respectively). While high precision is desirable to minimize false alarms, in medical diagnostics,

recall is often considered more critical than precision because missing a malignant case poses greater risk than a false positive.

The F1-score, which balances precision and recall, further supports Random Forest as the most stable model, achieving the highest malignant F1-score (0.95). Logistic Regression and KNN both achieved 0.94, while SVM achieved 0.93. Additionally, macro and weighted average F1-scores indicate that Random Forest maintains consistent performance across both benign and malignant classes.

In terms of benign tumor classification, SVM and KNN achieved perfect specificity (no false positives), correctly identifying all benign cases. Random Forest and Logistic Regression each misclassified one benign case as malignant, resulting in slightly lower specificity but maintaining better balance overall.

From a clinical standpoint, minimizing false negatives (malignant tumors classified as benign) is paramount. Random Forest produced the lowest number of false negatives (3 cases), followed by Logistic Regression (4), KNN (5), and SVM (6). Therefore, despite similar overall accuracy across models, Random Forest demonstrated the most favorable balance between sensitivity and specificity.

Overall, while all four algorithms performed strongly, Random Forest provided the most balanced and clinically reliable classification performance, particularly in identifying malignant tumors. This suggests that ensemble-based approaches may offer superior robustness and generalization capability for structured breast cancer diagnostic data.

4.4. Feature Importance

To further investigate the contribution of individual diagnostic variables to the classification decision, feature importance analysis was conducted. For the Random Forest model, feature importance was computed based on the average decrease in impurity across all decision trees. The resulting importance ranking is illustrated in Figure 5.

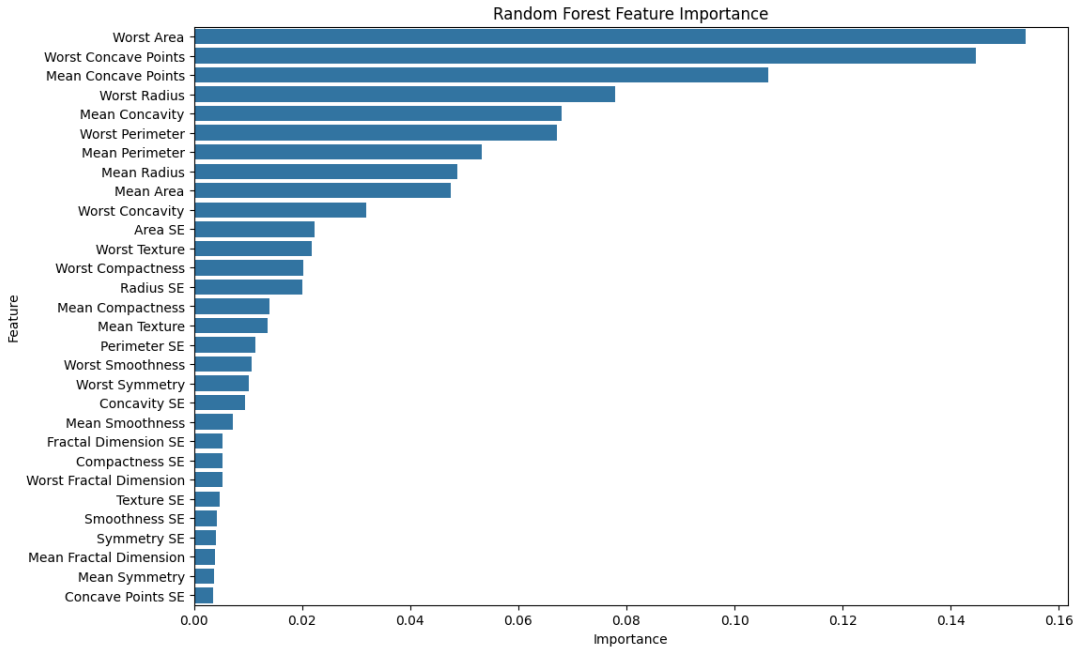

Figure 5. Random Forest Feature Importance

As shown in Figure 5, Worst Area emerged as the most influential feature in the Random Forest model, indicating that the maximum recorded tumor area significantly contributes to distinguishing malignant from benign cases. This is followed by Worst Concave Points and Mean Concave Points, emphasizing that tumor boundary irregularity plays a central role in malignancy prediction. Features related to tumor size and geometry, including Worst Radius, Mean Concavity, Worst Perimeter, and Mean Perimeter, also demonstrated high importance scores. These findings are clinically meaningful, as malignant tumors typically exhibit larger size and more irregular structural patterns.

 ISSN 2579-7069

In contrast, features such as Mean Symmetry, Mean Fractal Dimension, and several standard error (SE) attributes showed relatively lower importance values, suggesting that while they contribute to the model, their impact on classification is comparatively limited.

For comparative interpretability, Logistic Regression coefficients were also analyzed, as depicted in Figure 6. Unlike Random Forest, Logistic Regression assigns signed coefficients that indicate the direction and magnitude of influence for each feature. As shown in Figure 6, features such as Worst Concavity, Worst Compactness, and Worst Symmetry exhibited the largest positive coefficients, indicating a strong association with malignant classification. Conversely, features such as Texture SE and Mean Radius displayed strong negative coefficients, suggesting inverse contributions within the linear decision boundary. While Logistic Regression provides clearer interpretability due to its linear formulation, it does not capture nonlinear feature interactions as effectively as Random Forest.

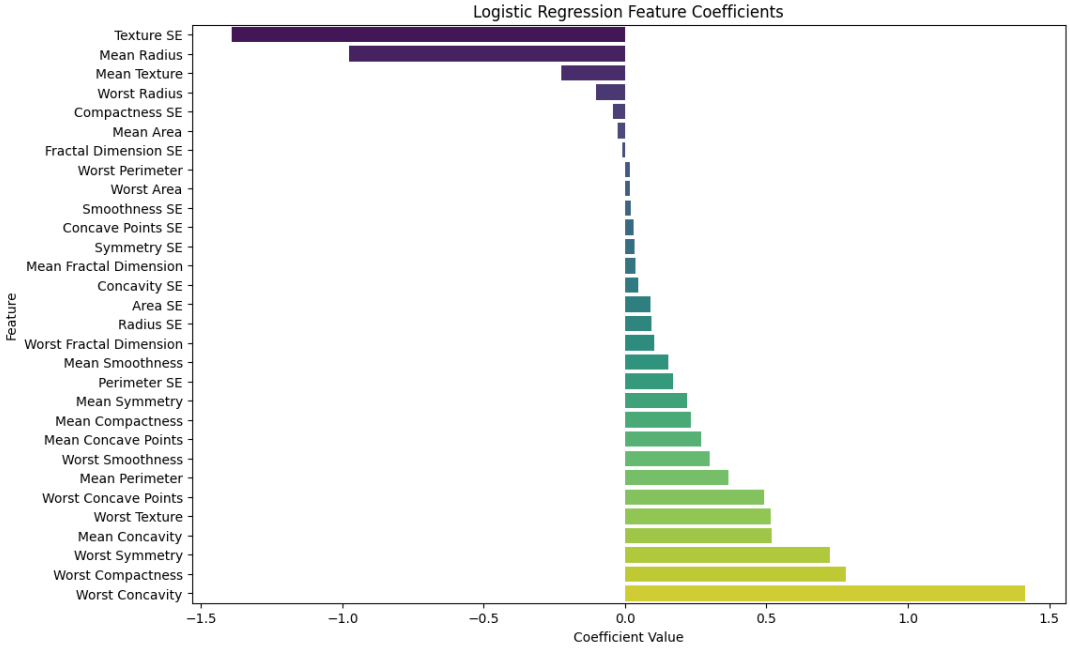

Figure 6. Logistic Regression Feature Importance

Overall, the feature importance analysis presented in Figure 5 and Figure 6 demonstrates that tumor size (area, radius, perimeter) and boundary irregularity (concave points, concavity, compactness) are the most decisive predictors in breast cancer classification. These results are consistent with established oncological knowledge and further reinforce the robustness and interpretability of the Random Forest model as a clinically relevant diagnostic support tool.

5. Discussion

The experimental results indicate that all four evaluated models Random Forest, SVM, KNN, and Logistic Regression achieved strong classification performance, with accuracy ranging from 0.95 to 0.96. These findings confirm that structured morphological features provide substantial discriminative power for benign–malignant tumor classification. Among the models, Random Forest demonstrated the most balanced performance, particularly achieving the highest recall for malignant tumors and the lowest false-negative rate. Given that minimizing false negatives is critical in cancer diagnostics, this suggests that ensemble learning effectively captures nonlinear interactions among tumor features.

SVM achieved perfect precision for malignant cases but exhibited lower recall, indicating a more conservative decision boundary. KNN performed competitively but showed slightly higher sensitivity to false negatives, likely due to its distance-based mechanism in high-dimensional spaces. Logistic Regression achieved comparable overall accuracy, suggesting partial linear separability of the dataset, though its linear structure limits modeling of complex feature interactions. Overall, ensemble-based approaches appear more robust for structured breast cancer datasets.

These findings are consistent with prior studies reporting the effectiveness of machine learning in addressing limitations of conventional imaging modalities, including variability in sensitivity, operator dependency, and accessibility constraints [11], [12]. ML techniques have been shown to improve predictive accuracy and reduce diagnostic variability [13], which is supported by the high accuracy achieved in this study.

The superior performance of Random Forest aligns with previous evidence highlighting the strength of ensemble methods in high-dimensional medical datasets [6], [17]. Similarly, SVM's competitive results reflect its established effectiveness in structured classification tasks [15]. The importance of morphological features such as tumor area, radius, and concavity identified in this study further corroborates earlier research emphasizing their clinical relevance [13]. The observed sensitivity–specificity trade-offs also reflect known challenges in medical ML applications, including class imbalance and generalization issues [31].

Several limitations should be acknowledged. The dataset represents a structured and controlled environment, which may not fully reflect real-world clinical variability. Moderate class imbalance may influence sensitivity–precision trade-offs. The study focused exclusively on classical ML algorithms applied to structured features, without incorporating imaging-based deep learning models. Additionally, external validation using multi-center datasets was not performed, limiting broader clinical generalizability.

The results demonstrate the feasibility of integrating machine learning-based decision-support systems into breast cancer screening workflows. In particular, models with high malignant recall, such as Random Forest, can function as early warning tools to reduce missed cancer diagnoses. Feature importance analysis enhances interpretability, strengthening clinician trust. Moreover, structured ML models require relatively low computational resources, making them suitable for implementation in resource-limited healthcare settings.

Future research should explore deep learning approaches, particularly CNN-based models for imaging datasets, and hybrid models combining structured and imaging features. Expanding datasets to include multi-center and demographically diverse populations would improve robustness. Incorporating advanced feature selection, imbalance-handling strategies, and explainable AI techniques such as SHAP or LIME may further enhance transparency and clinical adoption.

6. Conclusion

This study provides a comprehensive comparative analysis of four widely used machine learning algorithms Random Forest, SVM, KNN, and Logistic Regression for structured breast cancer tumor classification. The empirical results demonstrate consistently high predictive performance across models, with accuracy values reaching up to 0.96. Among the evaluated approaches, Random Forest achieved the most balanced and clinically favorable performance, particularly in terms of malignant tumor recall and minimized false-negative predictions. Given that missed malignant cases pose the greatest diagnostic risk, this finding underscores the superiority of ensemble-based approaches in structured medical datasets.

Beyond predictive performance, this research reinforces the importance of interpretable and computationally efficient models in clinical environments. While deep learning models dominate image-based detection research, the results confirm that classical machine learning algorithms remain highly effective for structured diagnostic data. The feature importance analysis further highlights the critical role of tumor morphology specifically area, concavity, radius, and perimeter in malignant classification, aligning computational findings with established oncological knowledge. By combining performance evaluation, interpretability analysis, and robust validation, this study contributes to the development of reliable and transparent machine learning-based decision-support systems in breast cancer diagnostics.

Future research should extend this work by integrating deep learning models for imaging-based datasets, exploring hybrid approaches that combine structured and image-derived features, and validating models across larger, multi-center, and demographically diverse datasets. Additionally, incorporating explainable AI techniques and advanced imbalance-handling strategies will further strengthen clinical applicability and trust. Ultimately, the integration of robust and interpretable machine learning models into early breast cancer detection workflows holds significant promise for improving diagnostic accuracy, reducing clinical burden, and enhancing patient outcomes.

7. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization A.A. and H.G.; Methodology, A.A. and H.G.; Software, A.A.; Validation, A.A.; Formal Analysis, A.A.; Investigation, H.G.; Resources, A.A.; Data Curation, H.G.; Writing Original Draft Preparation, A.A.; Writing Review and Editing, A.A. and H.G.; Visualization, H.G. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. M. Sy, F. Mbaye, B. Keneme, and P. M. Sembene, "Association between SNPs C1236T and C3435T of the ABCB1 Gene and the Neoadjuvant Chemotherapy in Breast Cancer," *Ann. Microbiol. Infect. Dis.*, vol. 6, no. 1, pp. 12–24, 2024, doi: 10.22259/2637-5346.0601002.

[2] S. Zhang, Z. Jin, L. Bao, and P. Shu, "The Global Burden of Breast Cancer in Women From 1990 to 2030: Assessment and Projection Based on the Global Burden of Disease Study 2019," *Front. Oncol*., vol. 14, no. 1, pp. 1–13, 2024, doi: 10.3389/fonc.2024.1364397.

[3] A. Ghoneim, B. Abdelmaksoud, F. Alruwaily, and N. B. Elsharkawy, "Measurement of the Extent of Awareness of Al jouf Region Residents with the Importance and Methods of Breast Cancer Early Detection in Females," *Asian Pacific J. Cancer Prev*., vol. 24, no. 3, pp. 945–951, 2023, doi: 10.31557/APJCP.2023.24.3.945.

[4] I. Aqeel and M. Iram, "The Impact of Breast Self-Examination Education on Breast Self-Examination Practice Among the Women in Rural Community Lahore Pakistan," *Int. J. Sci. Res. Publ.*, vol. 12, no. 1, pp. 278–283, 2022, doi: 10.29322/IJSRP.12.01.2022.p12136.

[5] C. Dai, "Study on Breast Cancer Image Detection and Classification Based on Residual Connected Convolutional Neural Network (CNN)," *Theor. Nat. Sci*., vol. 33, no. 1, pp. 224–230, 2024, doi: 10.54254/2753-8818/33/20240914.

[6] Q. Ruan, "A Comparative Study of Seven Machine Learning Algorithms for Breast Cancer Detection and Diagnosis," *Acad. J. Med. Heal. Sci*., vol. 4, no. 3, pp. 25–33, 2023, doi: 10.25236/AJMHS.2023.040305.

[7] Y.-W. Wang, T.-T. Kuo, Y. Chou, Y. Su, S. Huang, and C.-J. Chen, "Breast Tumor Classification using Short-ResNet with Pixel-based Tumor Probability Map in Ultrasound Images," Ultrason. Imaging, vol. 45, no. 2, pp. 74–84, 2023, doi: 10.1177/01617346231162906.

[8] A. J. Zaylaa and S. Kourtian, "Advancing Breast Cancer Diagnosis through Breast Mass Images, Machine Learning, and Regression Models," *Sensors*, vol. 24, no. 7, pp. 2312, 2024, doi: 10.3390/s24072312.

[9] K. T. Anyachebelu, S. H. Hosea, M. U. Abdullahi, and M. A. Ibrahim, "Comparative analysis of machine learning algorithms for breast cancer prediction," *Dutse J. Pure Appl. Sci*., vol. 9, no. 4b, pp. 71–82, 2024, doi: 10.4314/dujopas.v9i4b.8.

[10] A. S. Aloufi, R. A. Alharbi, H. S. Alshammari, R. S. Alzahrani, L. M. Alghamdi, R. A. Alotaibi, N. A. Almutairi, S. M. Alsharif, F. A. Alqahtani, and M. A. Alharthi, "Accuracy of Abbreviated Breast MRI in Diagnosing Breast Cancer in Women With Dense Breasts Compared With Standard Imaging Modalities," *Saudi J. Med. Med. Sci.*, vol. 13, no. 1, pp. 7–17, 2025, doi: 10.4103/sjmms.sjmms_58_24.

[11] K. Ramesh, A. Truong, Y. Wang, M. Rusckowski, and M. Gkikas, "Ligand-Specific Nano-Contrast Agents Promote Enhanced Breast Cancer CT Detection at 0.5 mg Au," *Int. J. Mol. Sci*., vol. 23, no. 17, pp. 9926, 2022, doi: 10.3390/ijms23179926.

[12] T. Borowiec, R. Matkowski, B. Cybulska-Stopa, T. Kuniej, A. Kołodziejczyk, D. Dupla, and A. Maciejczyk, "Baseline Cross-Sectional Imaging of Locally Advanced High-Risk Breast Cancer Facilitates Highly Customized Radiation Therapy in Surgically Inaccessible Anatomical Areas," *Front. Oncol.*, vol. 15, no. 1, pp. 1–10, 2025, doi: 10.3389/fonc.2025.1556122.

[13] A. Rezazadeh, Y. Jafarian, and A. Kord, "Explainable Ensemble Machine Learning for Breast Cancer Diagnosis Based on Ultrasound Image Texture Features," *Forecasting*, vol. 4, no. 1, pp. 262–274, 2022, doi: 10.3390/forecast4010015.

[14] A. Batool and Y.-C. Byun, "Toward Improving Breast Cancer Classification Using an Adaptive Voting Ensemble Learning Algorithm," *IEEE Access*, vol. 12, no. 1, pp. 12869–12882, 2024, doi: 10.1109/ACCESS.2024.3356602.

[15] M. R. Savalia and J. V. Verma, "Classifying Malignant and Benign Tumors of Breast Cancer," *Int. J. Reliab. Qual. E-Healthcare*, vol. 12, no. 1, pp. 1–19, 2023, doi: 10.4018/IJRQEH.318483.

[16] D. Chandra, S. Islam, and D. Pandey, "Dengue Prediction Using Time Series Forcasting I.E ARIMA Model," *Int. J. Health Sci. (Qassim).,* vol. 6, no. S8, pp. 4517–4526, 2022, doi: 10.53730/ijhs.v6nS8.13221.

[17] Y. Lai and P. Huang, "Machine Learning Identifies the Association Between Second Primary Malignancies and Postoperative Radiotherapy in Young-Onset Breast Cancer Patients," *PLoS One*, vol. 20, no. 2, pp. e0316722, 2025, doi: 10.1371/journal.pone.0316722.

[18] H. Bhukya and M. Sadanandam, "RoughSet based Feature Selection for Prediction of Breast Cancer," *Wirel. Pers. Commun*., vol. 130, no. 3, pp. 2197–2214, 2023, doi: 10.1007/s11277-023-10378-4.

[19] V. More, S. Srivastav, S. Gaud, S. Suman, and P. Soni, "Machine Learning Based Multiple Disease Prediction using Streamlit," *Int. J. Res. Appl. Sci. Eng. Technol*., vol. 12, no. 4, pp. 590–593, 2024, doi: 10.22214/ijraset.2024.59817.

[20] D. Chen, R. Parsa, A. Hope, B. Hannon, E. Mak, L. Eng, F.-F. Liu, N. Fallah-Rad, A. M. Heesters, and S. Raman, "Physician and Artificial Intelligence Chatbot Responses to Cancer Questions from Social Media," *JAMA Oncol.*, vol. 10, no. 7, p. 956, 2024, doi: 10.1001/jamaoncol.2024.0836.

[21] K. Sun and J. Pan, "Model of Storm Surge Maximum Water Level Increase in a Coastal Area Using Ensemble Machine Learning and Explicable Algorithm," *Earth Sp. Sci*., vol. 10, no. 12, pp. 1–21, 2023, doi: 10.1029/2023EA003243.

[22] C. Walker, P. Ramuhalli, V. Agarwal, N. Lybeck, and M. Taylor, "Development of Short-Term Forecasting Models Using Plant Asset Data and Feature Selection," *Int. J. Progn. Heal. Manag*., vol. 13, no. 1, pp. 1–13, 2022, doi: 10.36001/ijphm.2022.v13i1.3120.

[23] J.-H. Choi, Y. Choi, K. Lee, K.-H. Ahn, and W. Y. Jang, "Explainable Model Using Shapley Additive Explanations Approach on Wound Infection after Wide Soft Tissue Sarcoma Resection: 'Big Data' Analysis Based on Health Insurance Review and Assessment Service Hub," *Medicina (B. Aires)*., vol. 60, no. 2, pp. 327, 2024, doi: 10.3390/medicina60020327.

[24] J. Chi, J. Wang, H. Tang, S. Wang, and Z. Chen, "Development of a Risk Prediction Model for Linezolid-Induced Thrombocytopenia Based on the Machine Learning Algorithm," *J. Clin. Pharmacol*., vol. 65, no. 11, pp. 1451–1459, 2025, doi: 10.1002/jcph.70054.

[25] R. B. Amalo, Y. Tuti, and A. R. Dewi, "Diagnostic Challenges in a Case of Suspected Breast Cancer with Low FDG Uptake and an Incidental Thyroid Lesion: A Case Report and Literature Review," *World J. Nucl. Med*., vol. 24, no. 04, pp. 373–376, 2025, doi: 10.1055/s-0045-1809310.

[26] T. Zhou, "Comparison of Machine Learning Algorithms and Feature Importance Analysis for Star Classification," *Appl. Comput. Eng*., vol. 30, no. 1, pp. 261–270, 2024, doi: 10.54254/2755-2721/30/20230111.

[27] K. Cheng, J. Wang, J. Liu, X. Zhang, Y. Shen, and H. Su, "Public Health Implications of Computer-Aided Diagnosis and Treatment Technologies in Breast Cancer Care," *AIMS Public Heal*., vol. 10, no. 4, pp. 867–895, 2023, doi: 10.3934/publichealth.2023057.

[28] D. Ozcelik and Ö. Taştan, "A Weakly Supervised Clustering Method for Cancer Subgroup Identification," *Balk. J. Electr. Comput. Eng*., vol. 10, no. 2, pp. 178–186, 2022, doi: 10.17694/bajece.1033807.

[29] Y. Anusha, R. Visalakshi, and K. Srinivas, "Imbalanced Data Classification Using Improved Synthetic Minority Over-Sampling Technique," *Multiagent Grid Syst*., vol. 19, no. 2, pp. 117–131, 2023, doi: 10.3233/MGS-230007.

[30] V. Kumar, G. S. Lalotra, P. Sasikala, D. S. Rajput, R. Kaluri, K. Lakshmanna, M. Shorfuzzaman, A. Alsufyani, and M. Uddin, "Addressing Binary Classification Over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques," *Healthcare*, vol. 10, no. 7, p. 1293, 2022, doi: 10.3390/healthcare10071293.

[31] B. H. Aubaidan, R. A. Kadir, M. T. Lajb, M. Anwar, K. N. Qureshi, B. A. Taha, and K. Ghafoor, "A Review of Intelligent Data Analysis: Machine Learning Approaches for Addressing Class Imbalance in Healthcare – Challenges and Perspectives," *Intell. Data Anal.*, vol. 29, no. 3, pp. 699–719, 2025, doi: 10.1177/1088467X241305509.

[32] T. Xia, T. Dang, J. Han, L. Qendro, and C. Mascolo, "Uncertainty-Aware Health Diagnostics via Class-Balanced Evidential Deep Learning," *IEEE J. Biomed. Heal. Informatics*, vol. 28, no. 11, pp. 6417–6428, 2024, doi: 10.1109/JBHI.2024.3360002.

[33] N. Chukwurah, A. B. Ige, V. I. Adebayo, and O. G. Eyieyien, "Frameworks for Effective Data Governance: Best Practices, Challenges, and Implementation Strategies Across Industries," *Comput. Sci. IT Res. J.*, vol. 5, no. 7, pp. 1666–1679, 2024, doi: 10.51594/csitrj.v5i7.1351.

[34] H. Patel, D. Singh Rajput, O. Petru Stan, and L. Cristian Miclea, "A New Fuzzy Adaptive Algorithm to Classify Imbalanced Data," *Comput. Mater. Contin*., vol. 70, no. 1, pp. 73–89, 2022, doi: 10.32604/cmc.2022.017114.

[35] Q. Wang, J. Zheng, and W. Zhang, "Dense Fuzzy Support Vector Machine to Binary Classification for Imbalanced Data," *J. Intell. Fuzzy Syst*., vol. 45, no. 6, pp. 9643–9653, 2023, doi: 10.3233/JIFS-232414.

[36] Y.-H. Zhou, "Study for Iris Classification Based on Multiple Machine Learning Models," *Highlights Sci. Eng. Technol.,* vol. 23, no. 1, pp. 342–349, 2022, doi: 10.54097/hset.v23i.3620.

[37] S. Roobini, M. S. Kavitha, and S. Karthik, "A Systematic Review on Machine Learning and Neural Network Based Models For Disease Prediction," *J. Integr. Sci. Technol*., vol. 12, no. 4, pp. 1–11, 2024, doi: 10.62110/sciencein.jist.2024.v12.787.

[38] D. C. Joo, G. H. Kim, M. W. Lee, B. E. Lee, J. W. Kim, and K. B. Kim, "Artificial Intelligence-Based Diagnosis of Gastric Mesenchymal Tumors Using Digital Endosonography Image Analysis," *J. Clin. Med*., vol. 13, no. 13, pp. 3725, 2024, doi: 10.3390/jcm13133725.

[39] V. V. Sagar, K. S. Varshini, N. C. Kumar, K. Poorvitha, and P. Ashirwad, "Breast Cancer Detection Using Machine Learning Algorithm," *Int. J. Sci. Res. Eng. Manag.*, vol. 8, no. 4, pp. 1–5, 2024, doi: 10.55041/IJSREM31153.

[40] H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, F. Qanouni, and M. Bahaj, "An Ontological Model based on Machine Learning for Predicting Breast Cancer," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 7, pp. 108–115, 2022, doi: 10.14569/IJACSA.2022.0130715.

[41] S. Wang, "Design and Implementation of Breast Cancer Prediction System Based on Machine Learning," *Theor. Nat. Sci*., vol. 32, no. 1, pp. 238–247, 2024, doi: 10.54254/2753-8818/32/20240871.

[42] S. Chaurasiya and R. Rajak, "Comparative Analysis of Machine Learning Algorithms in Breast Cancer Classification," *Wirel. Pers. Commun*., vol. 131, no. 2, pp. 763–772, 2023, doi: 10.1007/s11277-023-10438-9.