International Journal of Global Perspective in Academic Research

Journal homepage: https://ijgpar.org/index.php/journal/index

# Research on Prediction Model of Benign and Malignant Breast Cancer Based on Machine Learning

**Lu Gan**

Northern Arizona University, USA

**Abstract:** This study uses three machine learning models to perform classification analysis on breast cancer data, aiming to improve diagnostic accuracy. After data preprocessing and standardization, the performance of each model was comprehensively evaluated using classification and regression metrics. The results showed that the logistic regression model performed the best, with an accuracy of 0.9825, while SVM and random forest also showed good performance. The classification effects of each model were visualized through ROC curves and confusion matrices, demonstrating that logistic regression has high application value in breast cancer diagnosis.

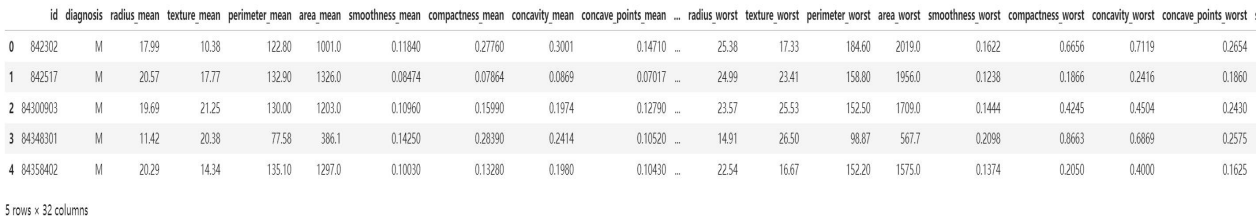
**Keywords**: Breast Cancer; Machine Learning; Logistic Regression; Support Vector Machine; Random Forest

## 1 Introduction

In today's medical field, breast cancer is one of the most common malignant tumors worldwide, and early accurate diagnosis is critical for the prognosis and treatment choices of patients. With the development of machine learning technology, using data-driven models to assist in breast cancer diagnosis has become an effective method. Traditional diagnostic methods rely on the expertise of doctors, while machine learning models can analyze large datasets to help doctors quickly identify benign and malignant tumors. This not only improves diagnostic accuracy but also reduces misjudgments caused by human error[1]. In this study, we use three classic machine learning algorithms—Support Vector Machine (SVM), Random Forest, and Logistic Regression—to analyze and classify a breast cancer dataset. The aim is to evaluate the performance of each model and compare their real-world applicability through multiple metrics. Through this analysis, we hope to provide more technical support for automated breast cancer diagnosis, ultimately enhancing the efficiency and accuracy of clinical diagnosis.

## 2 Data Collection and Preparation

In this study, the dataset used is named data.csv, and it can be downloaded from the following link: https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset?resource=download. As shown in Figure 1, the breast cancer dataset contains a total of 569 records, which are used to distinguish between benign and malignant tumor characteristics. The dataset includes various morphological features of tumors from patients. Each sample records 10 features such as radius, texture, perimeter, area, smoothness, compactness, concavity, etc., with each feature further divided into mean value (_mean), standard error (_se), and maximum value (_worst). The diagnosis results are labeled as "B" for benign tumors and "M" for malignant tumors. The "Id" represents the sample number. This dataset is widely used in training machine learning models, helping researchers develop and evaluate classification algorithms aimed at improving the accuracy of breast cancer diagnosis through data analysis and prediction, providing valuable support in medical diagnostics.



| | id | diagnosis | radius_mean | texture_mean | perimeter_mean | area_mean | smoothness_mean | compactness_mean | concavity_mean | concave points_mean | ... | radius_worst | texture_worst | perimeter_worst | area_worst | smoothness_worst | compactness_worst | concavity_worst | concave points_worst |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 842302 | M | 17.99 | 10.38 | 122.80 | 1001.0 | 0.11840 | 0.27760 | 0.3001 | 0.14710 | ... | 25.38 | 17.33 | 184.60 | 2019.0 | 0.1622 | 0.6656 | 0.7119 | 0.2654 |
| 1 | 842517 | M | 20.57 | 17.77 | 132.90 | 1326.0 | 0.08474 | 0.07864 | 0.0869 | 0.07017 | ... | 24.99 | 23.41 | 158.80 | 1956.0 | 0.1238 | 0.1866 | 0.2416 | 0.1860 |
| 2 | 84300903 | M | 19.69 | 21.25 | 130.00 | 1203.0 | 0.10960 | 0.15990 | 0.1974 | 0.12790 | ... | 23.57 | 25.53 | 152.50 | 1709.0 | 0.1444 | 0.4245 | 0.4504 | 0.2430 |
| 3 | 84348301 | M | 11.42 | 20.38 | 77.58 | 386.1 | 0.14250 | 0.28390 | 0.2414 | 0.10520 | ... | 14.91 | 26.50 | 98.87 | 567.7 | 0.2098 | 0.8663 | 0.6869 | 0.2575 |
| 4 | 84358402 | M | 20.29 | 14.34 | 135.10 | 1297.0 | 0.10030 | 0.13280 | 0.1980 | 0.10430 | ... | 22.54 | 16.67 | 152.20 | 1575.0 | 0.1374 | 0.2050 | 0.4000 | 0.1625 |

5 rows × 32 columns

**Figure 1 Basic Information of Breast Cancer Data**

The 10 features in the breast cancer dataset are categorized as _mean (average), _se (Standard Error), and _worst (maximum value) to comprehensively describe the tumor's morphological characteristics. _mean provides an overall trend of the features, _se reflects the uncertainty or variability of the measurements, and _worst captures the most extreme values. This classification method helps the model analyze the tumor from multiple perspectives, combining the average trend, variability, and extreme cases, thus enhancing the model's diagnostic capabilities. By using these classifications, the model can more accurately distinguish between benign and malignant tumors, improving the prediction's accuracy and reliability[2].

## 3 Basic Statistical Analysis of the Data

### 3.1 Correlation Classification and Selection of Feature Values

Since the breast cancer dataset contains 30 feature values, we categorized these features into four main groups for more efficient analysis and visualization. First, for basic morphological features, we selected the tumor's average radius (radius_mean), perimeter (perimeter_mean), area (area_mean), and texture (texture_mean) for data visualization. These features reflect the overall size and shape of the tumor. Second, based on features of morphological irregularity, we chose compactness (compactness_mean), concavity (concavity_mean), concave points (concave_points_mean), and smoothness (smoothness_mean), which help analyze the tumor's edge complexity[3]. Additionally, extreme value features record the tumor's most extreme measurements, so we selected the worst radius (radius_worst), worst perimeter (perimeter_worst), worst area (area_worst), and worst texture (texture_worst). Finally, for standard error features like radius standard error (radius_se), perimeter standard error (perimeter_se), and area standard error (area_se), these provide information on the variability of the measurements. These selected features help better understand and predict the diagnosis of breast cancer.

### 3.2 Relationship Between Basic Morphological Features and Diagnostic Results

We selected radius_mean, perimeter_mean, area_mean, and texture_mean from the breast cancer data as the basic morphological features. The chart displays the relationship between these four basic morphological features (radius_mean, perimeter_mean, area_mean, and texture_mean) and the diagnostic results (B for benign, M for malignant). Each chart uses a box plot to show the distribution of feature values across different diagnostic categories. It is evident that the median and overall distribution of radius_mean, perimeter_mean, and area_mean are significantly larger in malignant tumors (M) than in benign tumors (B), indicating that these features are larger in malignant cases. The distribution difference in texture_mean is relatively smaller, but there is still a slight upward trend in malignant tumors. The outliers in the box plots further highlight the extreme distributions of feature values in each category.

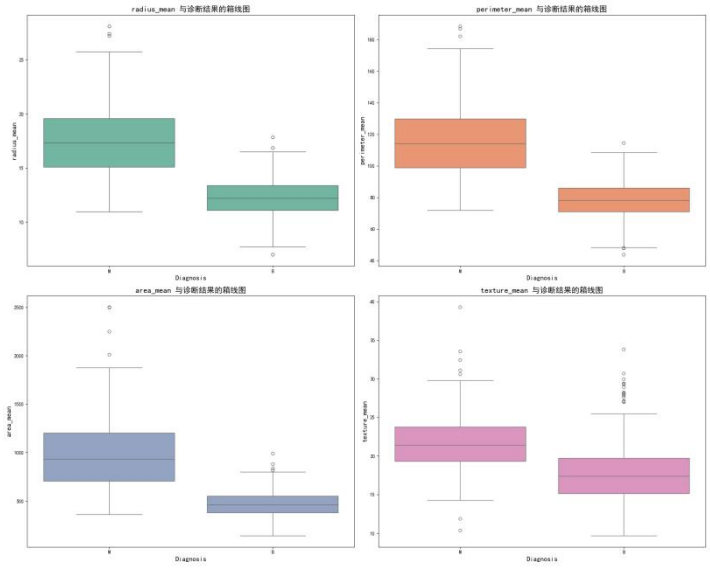



Figure 2 Data Comparison Display in the Scatter Plot

### 3.3 Relationship Between Morphological Irregularity Features and Diagnostic Results

Figure 3 shows the distribution of four morphological irregularity features (compactness_mean, concavity_mean, concave_points_mean, and smoothness_mean) in benign (B) and malignant (M) tumors within the breast cancer dataset. The violin plot illustrates that the distribution ranges for compactness_mean, concavity_mean, and concave_points_mean are significantly higher in malignant tumors, indicating that these features have higher values in malignant cases. This suggests that tumors with greater compactness, concavity, and concave points are more likely to be malignant. In contrast, the distribution of smoothness_mean shows a smaller difference between benign and malignant tumors, indicating that smoothness is less effective in distinguishing between diagnoses[4]. These observations highlight the importance of morphological irregularity features in differentiating between benign and malignant tumors.

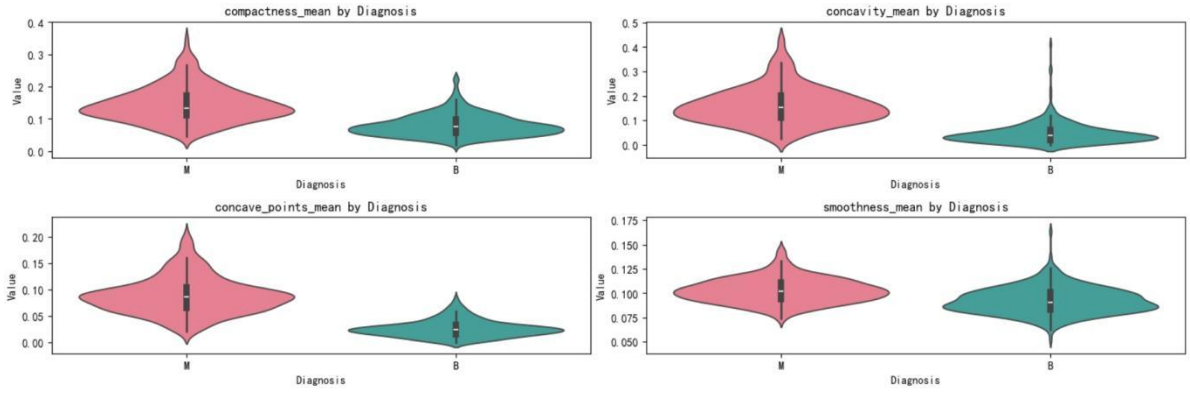



Figure 3 Violin Plot Data Comparison Display

### 3.4 Relationship Between Maximum Value Features and Diagnostic Results

This kernel density plot shows the distribution of four maximum value features (radius_worst, perimeter_worst, area_worst, and texture_worst) in benign (blue) and malignant (red) tumors within the breast cancer dataset. Overall, the feature values for malignant tumors are significantly higher than those for benign tumors, with particularly notable differences in the radius_worst and area_worst features. These differences suggest that these features hold significant reference value in distinguishing between benign and malignant tumors[5].

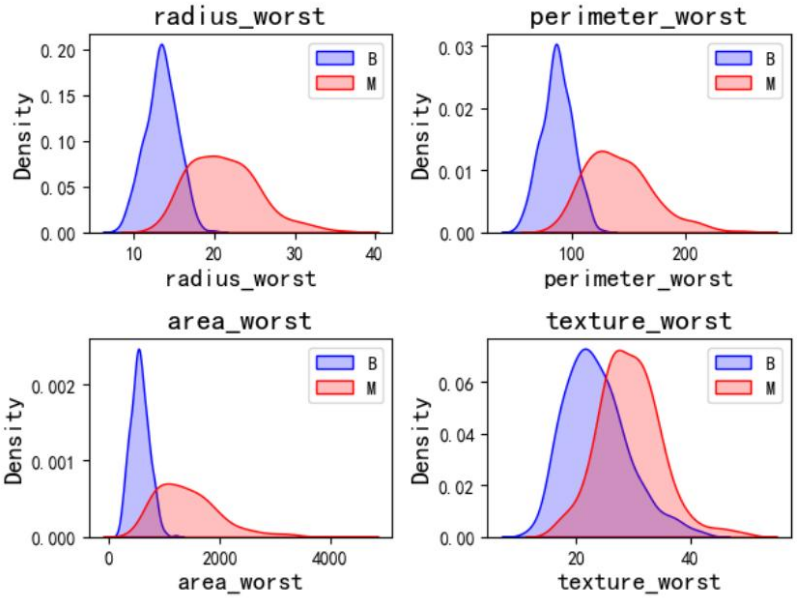



Figure 4 Kernel Density Display of Maximum Value Features for Benign and Malignant Tumors

### 3.5 Relationship Between Standard Error Features and Diagnostic Results

Figure 5 shows a scatter plot of the pairwise relationships between three standard error features (radius_se, perimeter_se, and area_se) in the breast cancer dataset, with colors distinguishing between benign (B) and malignant (M) tumors. The plot reveals a strong positive correlation between these features, especially between radius_se and perimeter_se, as well as perimeter_se and area_se, where the scatter points are nearly linearly distributed. The correlation between radius_se and area_se is also high, though slightly weaker. The values of these standard error features tend to be higher in malignant tumors, indicating that measurement uncertainty is more pronounced in malignant cases. This uncertainty may reflect tumor irregularities or complexity, which could influence the diagnosis.

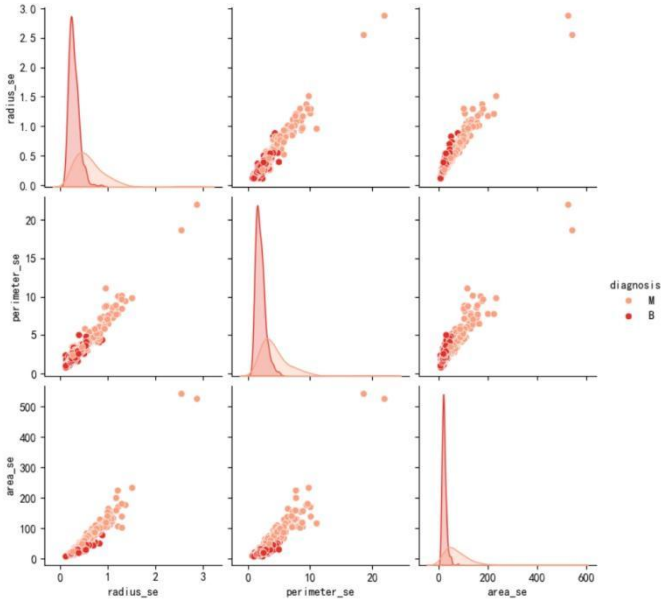



Figure 5 Scatter Plot Data Comparison Display of Relationships

## 4 Machine Learning Prediction Models

### 4.1 Selection of Machine Learning Prediction Models

For the breast cancer dataset, we selected three models for training: Support Vector Machine (SVM), Logistic Regression, and Random Forest. The choice of these models is based on their respective strengths and adaptability to

classification tasks. SVM excels in handling high-dimensional data and clear boundary classification problems, and its flexible kernel functions can effectively manage complex data distributions[6]. Logistic Regression is simple, easy to interpret, and particularly suitable for binary classification problems, providing intuitive probability outputs that help understand the confidence of diagnoses[7]. Random Forest, by integrating multiple decision trees, improves model accuracy and robustness, prevents overfitting, and can automatically handle feature selection and missing values[8]. This combination of models allows for a comprehensive evaluation and optimization of prediction performance in breast cancer classification tasks.

**4.2 Mathematical Principles and Formulas of Machine Learning Prediction Models**

*4.2.1 Mathematical Principles and Formula of the Support Vector Machine (SVM) Model*

Mathematical Principle: The goal of the Support Vector Machine (SVM) is to find the optimal hyperplane that maximizes the margin between classes for classification. For linearly separable data, SVM seeks a hyperplane that can completely separate the data, ensuring that the points closest to the decision boundary (called support vectors) are as far from the hyperplane as possible. For a linearly separable binary classification problem, assume we have a feature vector $x$ ,and its corresponding label is $y$. The decision boundary can be represented as:

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

where $w$ is the weight vector, and $b$ is the bias term.The objective is to maximize the margin $\frac{2}{\|w\|}$ while satisfying the constraint:

$$y_i (w \cdot x_i + b) \geq 1$$

for all training examples $(x_i, y_i)$.

*4.2.2 Mathematical Principles and Formula of Logistic Regression*

Mathematical Principle: Logistic regression is used to solve binary classification problems. It is based on the linear regression model but applies the Sigmoid function to constrain the output between [0,1], providing the probability that a sample belongs to a particular class. The core formula of logistic regression is:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

where $P(y = 1|x)$ is the probability that the sample belongs to class 1, w is the weight vector, $x$ is the feature vector, $b$ is the bias, and $e$ is the base of the natural logarithm. The Sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$ transforms the linear combination $w \cdot x + b$ into a probability value between 0 and 1.

*4.2.3 Mathematical Principles and Formula of Random Forest*

Mathematical Principle: Random Forest is an ensemble learning algorithm based on multiple decision trees. It builds several decision trees using random sampling and random feature selection, and the final prediction result is obtained through voting (for classification) or averaging (for regression) [8]. The basic formula for Random Forest classification is as follows:

Suppose we have $N$ decision trees, and the prediction result of each tree is $h_i(x)$ (where $i = 1,2, \ldots, N$). The final prediction result for Random Forest classification is determined by majority voting:

$$H(x) = \text{mode}\{h_1(x), h_2(x), \ldots, h_N(x)\}$$

where $H(x)$ is the final predicted class, and mode represents the most frequent class among the predictions of all trees.

**5 Data Preprocessing**

During data processing, the first step is to convert the category labels in the diagnosis column into numerical values before standardization or model training. We convert 'B' to 0 and 'M' to 1. Then, we remove the 'id' column and select all columns except diagnosis from the cleaned dataset (data_cleaned) as the feature variables X, and set the diagnosis column as the target variable y. Here, X contains the tumor's numerical features, while y represents the diagnostic result (benign or malignant). Next, we split the dataset into 70% training and 30% testing sets, using random_state=42 to ensure the split is reproducible. Finally, we use StandardScaler to standardize the feature data in both the training and testing sets, adjusting the data to have a mean of 0 and a standard deviation of 1[9]. This eliminates the differences in scale between features, improving the effectiveness of model training and prediction accuracy.

**6 Experimental Results**

**6.1 ROC Curve Analysis of the Three Models**

Figure 7 presents the ROC curves for the three models (SVM, Logistic Regression, and Random Forest). Each model

has an AUC value of 1.00, indicating very high classification performance on the breast cancer dataset. The ROC curve shows the classification results at different thresholds, with the x-axis representing the False Positive Rate and the y-axis representing the True Positive Rate. An AUC of 1.00 signifies that the models can perfectly differentiate between benign and malignant tumors. By comparing the curves, it is clear that all models closely approach the top-left corner, indicating excellent performance with nearly no misclassifications.

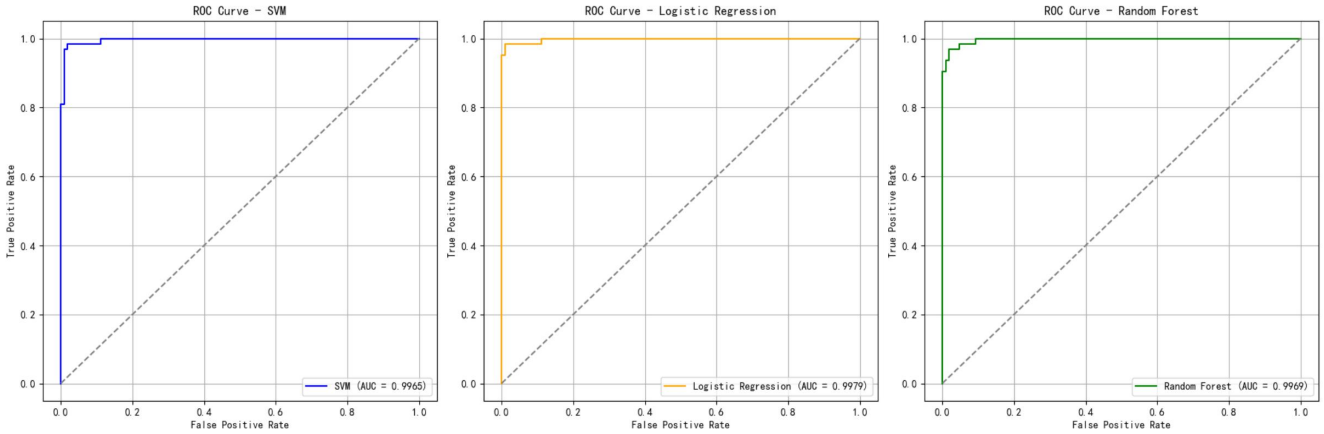



Figure 7 ROC Curve

### 6.2 Evaluation Performance Metrics for Each Model

Table 1 evaluates the prediction performance of each model using Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination ($R^2$) [10]. The results show that Logistic Regression performs the best across all three metrics, with the lowest MSE and MAE of 0.0175, and the highest $R^2$ of 0.9246, indicating the strongest fitting ability. SVM follows with an MSE of 0.0233 and $R^2$ of 0.8994. Random Forest performs slightly worse, with both MSE and MAE at 0.0292, and an $R^2$ of 0.8743. Overall, Logistic Regression demonstrates the best prediction accuracy and fitting performance.

Table 1

| Model | MSE | MAE | $R^2$ |
|---|---|---|---|
| **SVM** | 0.0233 | 0.0234 | 0.8994 |
| **Random Forest** | 0.0292 | 0.0292 | 0.8743 |
| **Logistic Regression** | **0.0175** | **0.0175** | **0.9246** |

### 6.3 Random Forest Classification Model Prediction Results Analysis

Table 2 presents the classification performance metrics of the three models (SVM, Random Forest, and Logistic Regression) on the breast cancer dataset. The models were evaluated using Precision, Recall, and F1 scores, analyzing both class labels (0 for benign and 1 for malignant) [11]. The results show that Logistic Regression performed best overall in terms of accuracy, precision, recall, and F1 score, with an accuracy of 0.9825, indicating the strongest classification ability. The SVM model performed slightly worse but still maintained high accuracy and strong metrics. The Random Forest model had a slightly lower recall for malignant tumors (1), with a recall rate of 0.94, but its overall performance was still very close to the other models. Overall, Logistic Regression was the best-performing model in this experiment.

Table 2

| Model | Accuracy | Precision (0) | Precision (1) | Recall (0) | Recall (1) | F1 (0) | F1 (1) |
|---|---|---|---|---|---|---|---|
| **SVM** | 0.9766 | 0.98 | 0.97 | 0.98 | 0.97 | 0.98 | 0.97 |
| **Random Forest** | 0.9708 | 0.96 | 0.98 | 0.99 | 0.94 | 0.98 | 0.96 |
| **Logistic Regression** | **0.9825** | **0.99** | **0.97** | **0.98** | **0.98** | **0.99** | **0.97** |

**7 Conclusion**

In this analysis and prediction experiment using the breast cancer dataset, we employed three models—Support Vector Machine (SVM), Random Forest, and Logistic Regression—for classification and performance evaluation. First, we preprocessed the dataset, including feature standardization and converting the target variable to numerical values. Then, the data was split into training and test sets, and each model was trained and tested. We comprehensively evaluated the models using classification metrics such as accuracy, precision, recall, and F1 score, as well as regression metrics like MSE, MAE, and $R^2$.

The experimental results showed that the Logistic Regression model performed best across all metrics, achieving an accuracy of 0.9825 and demonstrating strong classification ability. The SVM model followed closely, also performing excellently. Although the Random Forest model had some errors in classifying malignant tumors, it still maintained high accuracy and robustness overall. The ROC curve and confusion matrix visualizations provided an intuitive display of each model's classification performance at different thresholds. Ultimately, the Logistic Regression model proved to be the best-performing model in this experiment.

**References**


[1] Ara S, Das A, Dey A. Malignant and benign breast cancer classification using machine learning algorithms[C]//2021 International Conference on Artificial Intelligence (ICAI). IEEE, 2021: 97-101.

[2] Rabiei R, Ayyoubzadeh S M, Sohrabei S, et al. Prediction of breast cancer using machine learning approaches[J]. Journal of biomedical physics & engineering, 2022, 12(3): 297.

[3] Laghmati S, Tmiri A, Cherradi B. Machine learning based system for prediction of breast cancer severity[C]//2019 International Conference on Wireless Networks and Mobile Communications (WINCOM). IEEE, 2019: 1-5.

[4] Naji M A, El Filali S, Aarika K, et al. Machine learning algorithms for breast cancer prediction and diagnosis[J]. Procedia Computer Science, 2021, 191: 487-492.

[5] Chaurasia V, Pal S, Tiwari B B. Prediction of benign and malignant breast cancer using data mining techniques[J]. Journal of Algorithms & Computational Technology, 2018, 12(2): 119-126.

[6] Pisner D A, Schnyer D M. Support vector machine[M]//Machine learning. Academic Press, 2020: 101-121.

[7] Rigatti S J. Random forest[J]. Journal of Insurance Medicine, 2017, 47(1): 31-39.

[8] Hilbe J M. Logistic regression models[M]. Chapman and hall/CRC, 2009.

[9] Qiao Y, Li K, Lin J, et al. Robust Domain Generalization for Multi-modal Object Recognition[J]. arXiv preprint arXiv:2408.05831, 2024.

[10] Rawal R. Breast cancer prediction using machine learning[J]. Journal of Emerging Technologies and Innovative Research (JETIR), 2020, 13(24): 7.

Dalal S, Onyema E M, Kumar P, et al. A hybrid machine learning model for timely prediction of breast cancer[J]. International Journal of Modeling, Simulation, and Scientific Computing, 2023, 14(04): 2341023.