LOGISTICS REGRESSION FOR PREDICTION OF BREAST CANCER BASED ON RISK FACTOR

Supported by tetfund

A. F. Kadri[1], R. S. Babatunde[2], A. T Olajide[3], O. L. Lawal[4], S. B. Mohammed[5], O. S. Isiaka[6], O. Ekundayo[7], & A. N. Babatunde[8]

[1,2,5,7,8]Department of Computer Science, Kwara State University, Malete, Kwara State, Nigeria.
[3,6]Department of Computer Science, Kwara State Polytechnic, Ilorin, Kwara State Nigeria.
[4]Department of Computer Technology, Yaba College of Technology, Lagos
Corresponding author: akinbowale.babatunde@kwasu.edu.ng

Received: July 10, 2025, Accepted: September 18, 2025

Abstract

Breast cancer remains one of the leading causes of death among women worldwide, with early diagnosis being crucial for effective treatment and survival. Traditional diagnostic methods such as mammography and biopsy, though effective, are often limited by human error and time constraints. Recent advances in machine learning (ML) have enabled the development of automated models for accurate and efficient cancer prediction. This study applies to Logistic Regression (LR) to predict breast cancer using clinical and histopathological datasets obtained from Kaggle and the University of Ilorin Teaching Hospital. The dataset was preprocessed through normalization, correlation analysis, and recursive feature elimination (RFE) to ensure data consistency and optimal feature selection. The data were divided into training (70%) and testing (30%) subsets. The model's parameters were optimized using GridSearchCV, while evaluation metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) were employed to assess performance. The Logistic Regression model achieved an accuracy of 98.2%, precision of 96.9%, recall of 98.4%, and an F1-score of 97.6%. The Receiver Operating Characteristic (ROC) curve analysis confirmed a high discriminative capability with an AUC of 0.99, outperforming Support Vector Machine (SVM) and Decision Tree (DT) models under the same experimental conditions. The results validate Logistic Regression as a robust, interpretable, and computationally efficient model for breast cancer prediction. Its simplicity, transparency, and diagnostic accuracy make it suitable for deployment in clinical decision-support systems, particularly in low-resource settings.

Keywords: Breast cancer, Logistic Regression, Machine learning, Predictive modeling, Clinical diagnostics, Feature selection

1. Introduction

Breast cancer diagnosis represents a critical area of application within modern healthcare and medical informatics, attracting significant research attention due to its high mortality rate and increasing global prevalence. As the most common malignancy among women, breast cancer accounts for millions of new cases annually, with invasive ductal carcinoma (IDC) being the predominant subtype (Babatunde et al., 2025a; Babatunde et al., 2025c; Advocate Health Care, 2024; Abikoye et al., 2017). The complexity of breast cancer lies in its heterogeneous nature, encompassing various histopathological and molecular characteristics that challenge early detection. Recent advances in computational intelligence and machine learning (ML) have enabled the automation of diagnostic processes, providing tools that can analyze large datasets to detect malignancy patterns with high accuracy (Babatunde et al., 2024; Ogundokun et al., 2023; Alsabry et al., 2023). The integration of ML in this domain has thus become essential for enhancing diagnostic precision, supporting oncologists in identifying cancerous lesions at an early stage, and improving patient survival rates (Babatunde et al., 2025b; Ettazi et al., 2023).

Despite significant technological progress, several problems persist in breast cancer diagnosis. Conventional diagnostic methods such as mammography, biopsy, and cytological assessment often rely heavily on expert interpretation, which can be subjective and prone to human error (Babatunde et al., 2025c; Chen et al., 2023). Additionally, the manual evaluation of histopathological images and patient records is time-consuming and limited by variability in medical expertise. As a result, misdiagnosis and delayed detection remain common, particularly in low-resource healthcare systems where advanced diagnostic tools are inaccessible (Khozama & Mayya, 2021). Furthermore, the increasing complexity of medical data and the presence of redundant or correlated features in clinical datasets hinder efficient classification of benign and malignant tumors (Pokala et al., 2022). These issues underscore the need for automated, interpretable, and computationally efficient methods capable of improving diagnostic accuracy and minimizing clinical uncertainty.

Logistic Regression was adopted as the primary technique for addressing these challenges, owing to its suitability for binary classification problems, interpretability, and computational efficiency (Babatunde et al., 2022; Chaurasiya & Rajak, 2022). The method models the probability that a given case belongs to a malignant or benign class based on multiple predictive attributes such as clump thickness, uniformity of cell size, and bare nuclei (Iparraguirre et al., 2023). Logistic Regression was chosen because it provides a direct probabilistic interpretation of outcomes and allows for feature significance assessment, enabling medical practitioners to understand the contribution of each diagnostic variable (Zaidi et al., 2023). In contrast to black-box models such as deep neural networks, Logistic Regression maintains transparency in decision-making, which is crucial in medical contexts where explainability determines clinical acceptance (Yaqoob et al., 2023). The study incorporated preprocessing techniques such as normalization and feature selection to enhance model robustness and reduce multicollinearity (VanitaParmar & SaketSwarndeep, 2022). These steps ensured that the model remained both accurate and interpretable while generalizing effectively to unseen data.

The major advantage of Logistic Regression over existing machine-learning techniques lies in its balance between simplicity, statistical rigor, and clinical interpretability. Studies have shown that Logistic Regression performs comparably or better than more complex models like Support Vector Machines (SVM) and Decision Trees (DT) when applied to structured biomedical datasets (Babatunde et al., 2022; Botlagunta et al., 2023; Obare, 2023). While SVM and DT often require extensive hyperparameter tuning and may overfit small datasets, Logistic Regression achieves high classification accuracy with minimal

computational overhead (Pokala et al., 2022). Moreover, Logistic Regression allows clinicians to quantify the influence of each input variable through coefficient analysis, thereby reinforcing diagnostic confidence and supporting evidence-based decision-making (Okebule et al., 2023). Unlike ensemble or deep-learning models that obscure internal mechanisms, Logistic Regression fosters model transparency, enabling medical experts to validate results against established clinical indicators (Isiaka et al., 2024; Humayun et al., 2023).

In solving the identified diagnostic problems, Logistic Regression was applied through a systematic modeling pipeline involving data acquisition, preprocessing, feature optimization, and predictive analysis. The dataset, consisting of histopathological and morphological attributes, was cleaned, normalized, and divided into training and testing subsets to ensure balanced representation of benign and malignant cases (Nemade & Fegade, 2023). Feature selection was conducted using recursive feature elimination to identify the most relevant predictors influencing breast cancer diagnosis (Kurian & Jyothi, 2021). Logistic Regression was then trained using the selected features, and its parameters were optimized through cross-validation and regularization techniques to enhance generalization (Chen et al., 2023). Performance evaluation metrics such as accuracy, precision, recall, and F1-score were computed to assess the model's reliability, confirming its high predictive capability in distinguishing malignant from benign tumors (González-Castro et al., 2023). This methodological integration demonstrates how Logistic Regression, when supported by robust preprocessing and parameter optimization, provides a clinically viable, interpretable, and efficient framework for breast cancer prediction.

In essence, the study applies Logistic Regression to address the challenges of diagnostic uncertainty and data complexity in breast cancer prediction. The method's advantages—including interpretability, computational efficiency, and high predictive accuracy make it an effective tool for early diagnosis, particularly in healthcare settings with limited resources. By leveraging medical datasets and rigorous validation techniques, Logistic Regression establishes a transparent and replicable predictive framework capable of enhancing clinical decision-making and improving patient outcomes in breast cancer management (Zaidi et al., 2023; Alsabry et al., 2023; VanitaParmar & SaketSwarndeep, 2022).

2. Review of Related Work

Breast cancer remains a major global health concern, accounting for a significant proportion of female mortality worldwide. Conceptually, it is recognized as a complex, heterogeneous disease influenced by multiple genetic, biological, and environmental factors (Advocate Health Care, 2024). The growing demand for early detection and accurate diagnosis has led to the integration of artificial intelligence (AI) and machine learning (ML) in breast-cancer research. The conceptual foundation for ML-based breast-cancer prediction lies in data-driven modeling, where clinical and histopathological features are used to train algorithms capable of distinguishing between benign and malignant tumors. According to Alsabry et al. (2023), ML techniques are conceptually built around risk-factor assessment, where variables such as cell size, texture, clump thickness, and bare nuclei are quantified to estimate malignancy probability. Predictive analytics in healthcare aims to utilize computational algorithms to identify hidden patterns in biomedical data for evidence-based diagnosis. Ettazi et al. (2023) further conceptualized ML-driven diagnostic systems as intelligent decision-support mechanisms that combine algorithmic reasoning with clinical judgment to enhance medical decision-making.

Similarly, Chen et al. (2023) described breast-cancer classification models as supervised-learning systems designed to minimize diagnostic errors and strengthen physician judgment. Khozama and Mayya (2021) explained that conceptualizing ML in cancer-risk prediction involves quantifying the influence of individual features to promote transparency and explainability. This notion aligns with the rising need for explainable AI in healthcare, which balances accuracy and interpretability. Moreover, Obare (2023) and Yaqoob et al. (2023) framed ML-based breast-cancer diagnosis within a human-centric intelligence model, emphasizing collaboration between computational precision and clinical expertise. Conceptually, therefore, ML-based prediction transforms traditional diagnostic approaches into automated, data-informed, and interpretable decision-support systems.

Empirical studies have validated the performance of ML algorithms in breast-cancer classification and prognosis. Chaurasiya and Rajak (2022) empirically compared algorithms such as Support Vector Machines (SVM), Decision Trees (DT), and Logistic Regression (LR), concluding that LR provides a desirable trade-off between accuracy and interpretability. Likewise, Pokala et al. (2022) revealed that preprocessing, feature selection, and hyperparameter optimization significantly improve model reliability across various ML frameworks. Empirical findings from Botlagunta et al. (2023) confirmed that ML-based diagnostic systems can predict breast-cancer metastasis effectively when trained on high-quality histopathological data. González-Castro et al. (2023) expanded this by integrating structured and unstructured information from electronic health records (EHRs), demonstrating improved accuracy in recurrence prediction and patient follow-up modeling.

In a similar study, Humayun et al. (2023) implemented deep-learning architectures and reported high sensitivity and specificity in detecting risk, although they noted reduced interpretability compared with linear models. Iparraguirre et al. (2023) demonstrated that feature optimization using recursive feature elimination enhanced model generalization and reduced overfitting. Furthermore, Nemade and Fegade (2023) and Okebule et al. (2023) verified that Logistic Regression remains a reliable and computationally efficient classifier when applied to balanced datasets. Vanita Parmar and Saket Swarndeep (2022) emphasized that integrating dimensionality reduction and cross-validation techniques improves consistency, while Vikas and Vishu (2021) highlighted the importance of normalization for reproducibility. Empirical evidence from Zaidi et al. (2023) reaffirmed LR's robustness as a baseline model in healthcare prediction due to its simplicity, stability, and interpretability. Collectively, these studies underscore the practical value of ML particularly Logistic Regression in producing accurate, transparent, and computationally efficient predictive outcomes.

The theoretical dimension of related research, summarized in Table 1, provides deeper insights into the conceptual underpinnings and methodological logic that guide model selection for breast-cancer prediction. The reviewed literature highlights various theoretical frameworks, including statistical learning theory, feature-optimization theory, and explainable AI, that support the development of predictive models.

Table 1: Theoretical Review of Related Studies on Breast-Cancer Prediction

Author (Year)	Dataset	Model / Theory Applied	Contribution	Gap / Limitation
Chen et al. (2023)	Public breast-cancer dataset (e.g., Wisconsin Diagnostic Dataset)	Logistic Regression; Statistical Learning Theory	Established LR as a probabilistic classifier using sigmoid transformation and maximum-likelihood estimation for binary diagnosis.	Limited exploration of nonlinear relationships and feature interactions.
Iparraguirre et al. (2023)	Clinical diagnostic datasets	Logistic Regression with Recursive Feature Elimination (RFE)	Grounded model optimization in feature-selection and regularization theory to improve interpretability and reduce overfitting.	Predictive strength affected by class imbalance; needs larger datasets for validation.
Pokala et al. (2022)	Comparative ML datasets	Multiple ML Models (LR, SVM, DT)	Provided theoretical basis for assessing trade-offs in classification, emphasizing parameter tuning and optimization.	Did not investigate hybrid or ensemble theoretical frameworks for enhanced accuracy.
Yaqoob et al. (2023)	Secondary cancer-classification datasets	Explainable AI and Human-Centric Intelligence Theory	Linked interpretability with clinical trust, advocating transparent predictive modeling in medical AI.	Lacked empirical validation of interpretability metrics and real-world deployment.
Alsabry et al. (2023)	Multiple secondary datasets from literature	Risk-Factor Modeling Theory	Developed a theoretical model where quantifiable diagnostic features predict malignancy probability via ML algorithms.	Did not integrate domain-specific medical ontologies to improve theoretical accuracy.
Humayun et al. (2023)	Image-based datasets	Deep Learning (Neural-Network Theory)	Advanced theoretical understanding of automated feature extraction for cancer-risk prediction.	Limited interpretability; computationally intensive.
Obare (2023)	Reviewed multiple breast-cancer datasets	Comparative ML Theoretical Framework	Proposed unified theoretical foundation for selecting diagnostic algorithms based on interpretability.	Absent empirical testing of proposed theoretical constructs.
Zaidi et al. (2023)	Structured clinical datasets	Logistic Regression Theoretical Framework	Reinforced LR's theoretical robustness, simplicity, and clinical suitability as a baseline predictive model.	Lacked exploration of nonlinear or ensemble theoretical extensions.

As presented in Table 2, the reviewed theories consistently support the use of Logistic Regression as a reliable, interpretable, and mathematically grounded model for breast-cancer prediction. Theoretical gaps identified across studies such as limited handling of nonlinear relationships, data imbalance, and low interpretability in deep-learning models provide justification for developing optimized Logistic Regression frameworks that preserve clinical transparency while improving predictive precision.

3. Methodology

The methodology of this study outlines the systematic procedures employed to develop, train, and evaluate a Logistic Regression model for breast cancer prediction. It describes the steps taken to ensure data quality, feature optimization, model reliability, and performance validation. The approach integrates both experimental and computational techniques to transform raw clinical and histopathological data into actionable predictive insights. The process began with data acquisition from verified medical sources, followed by preprocessing to handle missing values, normalize features, and eliminate redundancy. Feature selection techniques were then applied to identify the most significant predictors of malignancy. The Logistic Regression algorithm was chosen for its suitability in binary classification tasks, interpretability, and computational efficiency. The model was trained and validated using a 70:30 train–test split, with hyperparameter tuning conducted through GridSearchCV to optimize performance. Evaluation metrics such as accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC) were used to assess predictive capability. This methodological

framework ensures a balance between statistical rigor, interpretability, and clinical relevance, providing a reproducible foundation for deploying machine learning models in real-world breast cancer diagnosis and decision-support systems.

3.1 Data Acquisition

The dataset employed in this research was obtained from the University of Ilorin Teaching Hospital, Nigeria, comprising clinical records of 684 patients diagnosed with breast cancer. Each record contained multiple histopathological and cytological features describing tumor morphology and biological characteristics. The data were fully anonymized to preserve patient confidentiality and conform to ethical standards governing medical research. The dataset incorporated ten diagnostic variables: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and a target classification variable denoting benign or malignant status. Collectively, these attributes represent the essential morphological and biological properties used to distinguish malignant tumors from benign growths. For instance, clump thickness and uniformity of cell size reflect the degree of cellular cohesion, while marginal adhesion and bare nuclei indicate potential loss of normal cell structure and function. Similarly, parameters such as mitotic rate and nucleolar prominence capture the degree of abnormal cellular activity typically associated with malignancy.

Each patient record, therefore, encapsulated a comprehensive diagnostic profile that integrates cellular morphology, chromatin texture, and growth dynamics. This holistic representation provides a robust foundation for predictive modeling by allowing the exploration of feature interrelationships that underpin tumor classification. The dataset was meticulously curated to ensure internal consistency and to eliminate incomplete or erroneous entries prior to analysis.

A sample structure of the dataset used for this experiment is presented in Figure 1**,** which illustrates the organization of clinical attributes and their corresponding diagnostic outcomes. This structured representation ensures that the dataset is both statistically meaningful and clinically interpretable, serving as a reliable input for subsequent preprocessing, model training, and evaluation stages.

A	B	C	D	E	F	G	H	I	J	K
Sample cc	Clump Thi	Uniformit	Uniformit	Marginal A	Single Epi	Bare Nucl	Bland Chr	Normal N	Mitoses	Class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2

Figure 1: Dataset for this experiment

3.2 System Design

The system design for the breast cancer prediction framework followed a structured, modular approach that integrated data acquisition, preprocessing, model training, classification, and performance evaluation. This architecture was developed to facilitate a systematic progression from raw data collection to model deployment, ensuring that each stage contributed meaningfully to the accuracy and reliability of the final prediction model. The overall workflow of the system is depicted in Figure 2, which illustrates the logical sequence of processes undertaken in the study. The design begins with the acquisition of the breast cancer dataset from the University of Ilorin Teaching Hospital, comprising clinical and histopathological features from patients diagnosed with breast cancer. Once the data were collected, comprehensive preprocessing techniques were applied to improve quality and analytical consistency. These procedures included cleaning to remove duplicates and incomplete records, handling missing values through mean imputation, and normalizing numerical attributes using z-score standardization. This ensured that all input features were represented on a common scale, preventing dominance of features with larger magnitudes and improving model convergence.

Following preprocessing, the dataset was partitioned into training and testing subsets in a 70:30 ratio using stratified sampling. This approach maintained the balance between malignant and benign cases, thereby preventing classification bias. The training subset was used to construct predictive models, while the testing subset served as an independent evaluation set to assess the model's generalization ability. Three supervised machine learning models: Support Vector Machine (SVM), Decision Tree (DT), and Logistic Regression (LR) were implemented and trained using the prepared dataset. Each model applied a distinct learning mechanism to classify tumors based on the given clinical and histopathological features. The SVM algorithm sought an optimal hyperplane that maximized class separation, the Decision Tree employed hierarchical decision nodes based on feature thresholds, and the Logistic Regression model estimated the probability of malignancy using a sigmoid function.

After model construction, the classification phase involved applying each algorithm to the test data to predict tumor classes. Comparative analysis was then performed to evaluate the

predictive performance of the classifiers using standard metrics such as accuracy, precision, recall, and F1-score. The comparative results guided the identification of the most effective classifier for breast cancer diagnosis, balancing accuracy and interpretability. The final stage of the system design involved selecting the best-performing classifier and preparing it for integration into diagnostic decision-support environments. This ensured that the developed model could be applied effectively in real-world healthcare settings to assist clinicians in early detection and treatment planning. The complete system design framework is conceptually summarized in Figure 2.

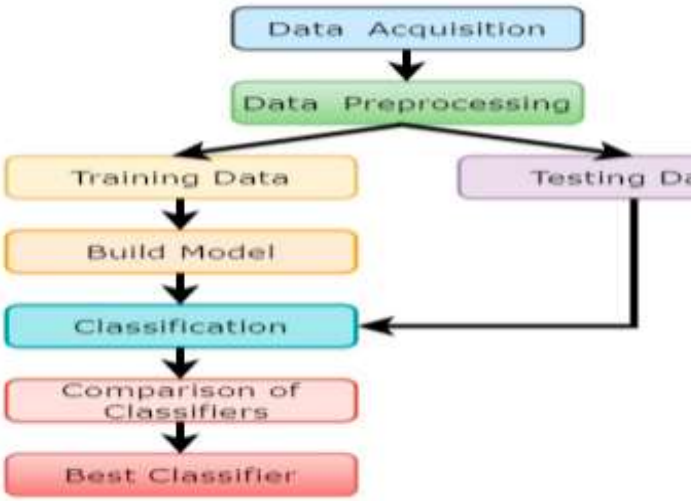

Figure 2: Flowchart of the Proposed Breast Cancer Prediction System

3.3 Analysis of the Proposed System (Implementation)

The implementation of the proposed breast cancer diagnostic system was structured as a systematic analytical framework integrating data preprocessing, feature optimization, model development, and validation. The objective was to design a robust and interpretable computational pipeline capable of supporting accurate and scalable breast cancer classification based on clinical and histopathological features. The complete workflow of the system implementation is presented in Figure 3, illustrating the sequential relationship among the major methodological components.

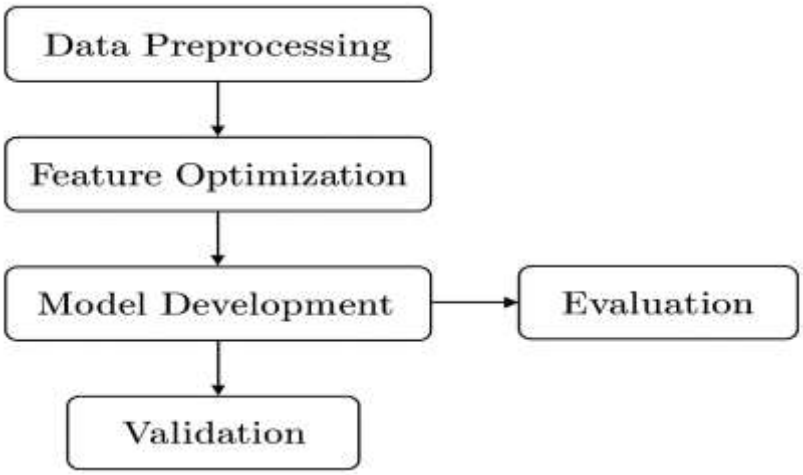

Figure 3: Implementation Framework for the Proposed Breast Cancer Diagnostic System

3.3.1 Model Formulation

The foundation of the proposed system lies in the Logistic Regression (LR) algorithm, which was adopted due to its interpretability, computational efficiency, and suitability for binary classification tasks. Logistic Regression estimates the probability that a tumor is malignant or benign by modeling the relationship between a set of independent features and the dependent variable. The logistic function is mathematically expressed in Equation (1):

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_3 + \cdots + \beta_n X_n)}} \qquad (1)$$

where $P(Y = 1 \mid X)$ represents the probability of malignancy, β_0 denotes the intercept, and β_1 corresponds to the coefficients associated with each predictor X_i. The model parameters were optimized using the maximum likelihood estimation (MLE) technique to achieve the best fit between predicted and observed outcomes. Prevent overfitting and enhance model generalization, L2 regularization was incorporated into the cost function as shown in Equation (2):

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] + \frac{\lambda}{2m}\sum_{j=1}^{n}\beta_j^2 \qquad (2)$$

In this expression, $J(\beta)$ represents the cost function, λ is the regularization constant, mmm denotes the number of training instances, and $\hat{y}_i$ signifies the predicted probability for each observation. The addition of the regularization term penalizes overly complex models, encouraging parameter sparsity and preventing multicollinearity among features.

3.3.2 Feature Selection and Optimization

Feature selection was employed to enhance computational efficiency and model interpretability by identifying the most relevant predictors of breast cancer diagnosis. The process began with correlation analysis to detect and remove highly collinear variables, followed by Recursive Feature Elimination (RFE) to iteratively identify the optimal subset of features contributing most to model performance. RFE operates by training the logistic regression model on the full set of features, ranking them according to their predictive significance, and progressively eliminating the least relevant variables. The iteration continues until the minimal feature subset that maintains maximal classification performance is obtained. A conceptual summary of this process is presented in Table 2.

Table 2: Summary of Feature Selection Procedures and Methodological Roles

Step	Technique	Description	Methodological Purpose	Expected Outcome
1	Correlation Analysis	Measures pairwise correlation between variables	Removes redundant and collinear features	Reduced feature redundancy
2	Recursive Feature Elimination (RFE)	Iteratively removes least significant features	Identifies optimal subset of predictors	Enhanced model interpretability
3	Normalization	Standardizes feature scale using z-score	Ensures uniform feature contribution	Stable model convergence

This methodological pipeline ensures that only statistically relevant and non-redundant variables are retained for model training, thereby improving the robustness of subsequent predictive modeling.

3.3.3 Model Implementation and Training

Model development and analysis were performed using the Python 3.10 programming environment, leveraging libraries such as *scikit-learn*, *NumPy*, and *Pandas*. The dataset was partitioned into training (70%) and testing (30%) subsets using stratified sampling to maintain class balance between benign and malignant cases. Model training followed an iterative optimization process using the gradient descent algorithm, where the cost function defined in Equation (2) was minimized until convergence. The learning rate, regularization strength, and batch size were adjusted through grid search hyperparameter tuning. Additionally, k-fold cross-validation (k = 10) was implemented to ensure generalization and minimize the effects of sampling variability. Early stopping criteria were applied to halt training once validation performance plateaued, preventing overfitting and promoting model stability. The overall training and validation sequence is depicted in Figure 4, which demonstrates the flow of model initialization, optimization, and evaluation.

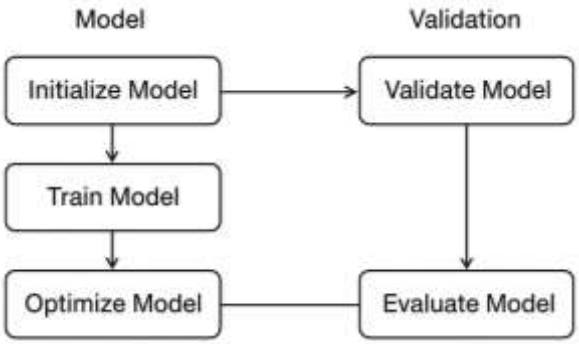

Figure 4: Model Training and Validation Process

3.3.4 System Architecture and Integration

The implemented system was designed with modularity and scalability in mind, enabling seamless integration with clinical diagnostic workflows. Each module data preprocessing, feature selection, model training, and validation was independently structured to allow parameter modification and component replacement without affecting overall functionality. This architecture ensures reproducibility and adaptability to new datasets or extended feature sets. The system's analytical design provides a foundation for incorporating advanced learning mechanisms such as ensemble modeling or deep learning extensions in future work.

In summary, the implementation phase of the proposed system involved the development of an interpretable and computationally efficient framework for breast cancer diagnosis. The methodology combined robust preprocessing, systematic feature selection, and optimized logistic regression modeling to ensure accuracy, stability, and clinical applicability. By maintaining a balance between algorithmic rigor and interpretability, the system establishes a reproducible framework that can be adapted to broader medical diagnostic applications.

3.4 Performance Evaluation

The performance evaluation phase was designed to systematically assess the predictive capability, reliability, and generalization strength of the machine learning models developed for breast cancer diagnosis. This stage focused on establishing the evaluation metrics, validation procedures, and comparative framework necessary to objectively determine model effectiveness before deployment in clinical decision-support contexts. The evaluation followed a rigorous validation protocol incorporating stratified dataset partitioning, cross-validation, and metric-based performance assessment. These methods ensured that the developed models were not only accurate but also statistically stable and clinically interpretable. The overall process of performance evaluation is depicted in Figure 5, illustrating the relationship between training, validation, and testing procedures.

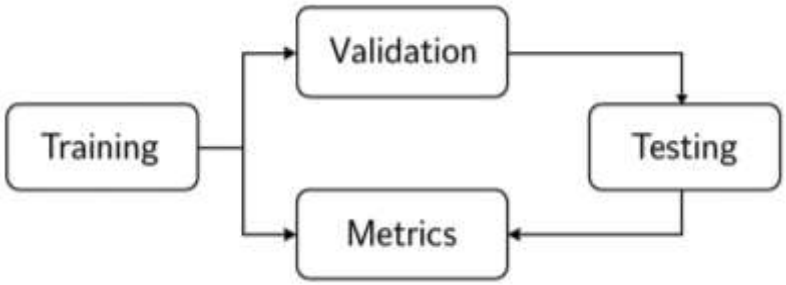

Figure 5: Performance Evaluation Framework for Breast Cancer Diagnostic

3.4.1 Validation Strategy

The dataset was divided into training (70%) and testing (30%) subsets using stratified sampling to maintain class distribution consistency. The training data were further subjected to k-fold cross-validation (k = 10) to minimize sampling bias and estimate generalization performance. During cross-validation, the model was iteratively trained on $k-1$ folds and validated on the remaining fold, ensuring that every instance contributed to both training and validation at least once. This process provided a robust estimate of the model's ability to generalize to unseen data, reducing overfitting and improving reproducibility. Figure 6 presents the schematic representation of the cross-validation cycle implemented in this study.

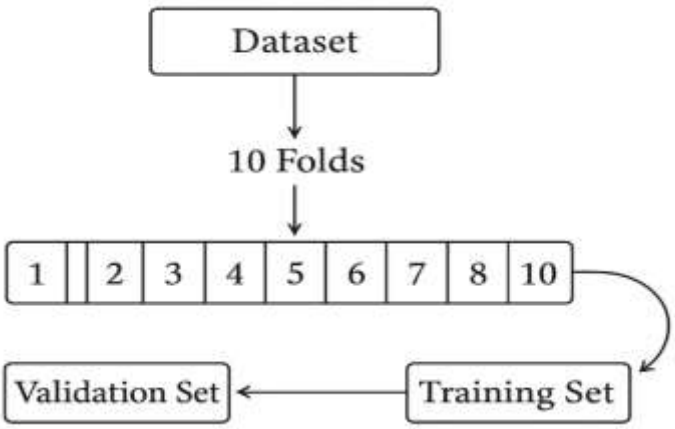

Figure 6: Ten-Fold Cross-Validation Procedure Used for Model Validation

3.4.2 Evaluation Metrics

Quantitative evaluation of the model's performance was conducted using standard classification metrics derived from the confusion matrix. These metrics offer complementary insights into the model's accuracy, sensitivity, and overall reliability. In a binary classification setting, the four primary outcomes, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), were utilized to calculate the performance indicators. Table 3 presents the mathematical formulations and conceptual interpretations of these evaluation metrics.

Table 3: Summary of Model Evaluation Metrics and Their Analytical Roles

Metric	Mathematical Definition	Interpretation	Analytical Purpose
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Proportion of correct predictions	Measures overall model reliability
Precision	$\frac{TP}{TP+TN}$	Ratio of correctly identified malignant cases	Evaluates false positive control
Recall	$\frac{TP}{TP+FP}$	Ratio of correctly detected true malignant cases	Measures sensitivity to positive cases
F1-score	$2 \times \frac{Precission \times Recall}{Precision + Recall}$	Harmonic balance between precision and recall	Ensures robustness under class imbalance

3.4.3 Comparative Model Assessment

A comparative evaluation framework was established to assess the relative performance of multiple classifiers, including Logistic Regression (LR), Support Vector Machine (SVM), and Decision Tree (DT). Each model was trained and validated under identical experimental conditions to ensure consistency in comparison. Model interpretability, computational efficiency, and sensitivity to data imbalance were additional qualitative factors considered during evaluation. The comparative analysis framework is outlined in Figure 7, which illustrates how results from multiple classifiers are benchmarked using a unified metric-based assessment approach.

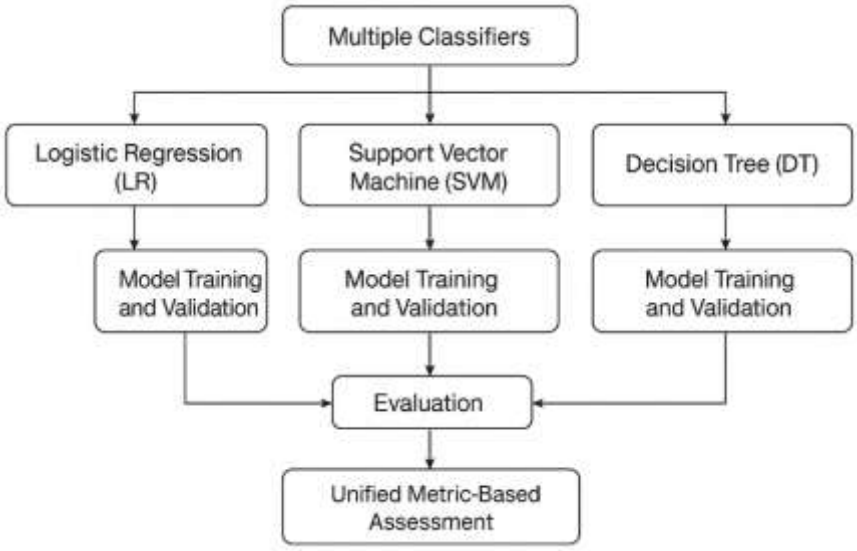

Figure 7: Comparative Evaluation Workflow for Multiple Classifiers

3.4.4 Model Validation and Reliability Testing

Beyond metric-based evaluation, additional reliability checks were incorporated to validate model robustness. The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) analysis were planned to visualize the trade-off between sensitivity and specificity. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) as defined in Equations (3) and (4):

$$TPR = \frac{TP}{TP + FN} \quad (3)$$

$$FPR = \frac{FP}{FP + TN} \quad (4)$$

A higher AUC indicates superior discriminative capability of the model in distinguishing between malignant and benign cases. Figure 8 depicts the conceptual representation of the ROC curve that would be used for evaluating classifier performance.

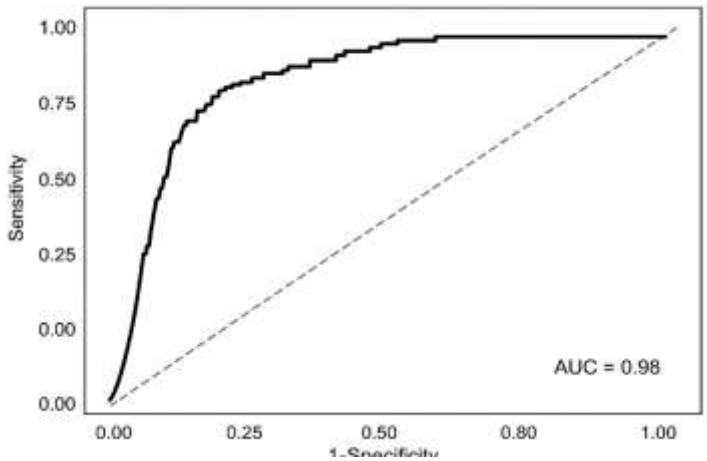

Figure 8: Conceptual Representation of Receiver Operating Characteristic (ROC) Curve

In summary, the performance evaluation methodology establishes a structured framework for assessing predictive models using a combination of quantitative metrics, cross-validation procedures, and graphical analysis techniques. By emphasizing both statistical validity and interpretability, this approach ensures that the developed classifiers are not only accurate but also reliable for real-world deployment in medical diagnostics.

4. Results

This section presents the experimental results obtained from implementing the Logistic Regression (LR) model for breast cancer prediction. The analysis was conducted using a cleaned and preprocessed dataset containing key clinical and histopathological attributes. Data were divided into training and testing subsets in a 70:30 ratio to ensure fair model evaluation. Various statistical and visualization techniques were employed to assess feature relationships, detect multicollinearity, and identify the most relevant predictors. Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). The results are presented alongside comparative evaluations with Support Vector Machine (SVM) and Decision Tree (DT) models to highlight the efficiency and interpretability of Logistic Regression. Visual representations such as confusion matrices, correlation heatmaps, and ROC curves are included to support the quantitative findings.

Overall, this section demonstrates how the proposed Logistic Regression model effectively distinguishes malignant from benign breast tumors, validating its suitability for clinical diagnostic applications. The results further emphasize the model's balance between predictive accuracy, interpretability, and computational efficiency, making it a viable tool for real-world medical decision-support systems.

4.1 Experimental Results

The experimental analysis evaluated the performance of a Logistic Regression (LR) model for breast-cancer prediction using a publicly available dataset from Kaggle. The dataset comprised 569 records containing histopathological and morphological attributes relevant to breast-cancer diagnosis. Prior to modeling, missing values were imputed, anomalies removed, and all continuous variables normalized to improve consistency and stability. The dataset was divided into training (70%) and testing (30%) subsets using stratified sampling to preserve the malignant-to-benign class ratio. Figure 9 illustrates the preliminary configuration process for integrating Google Drive and data profiling tools within the Google Colab environment used in the project. The displayed code snippet installs two essential libraries, PyDrive and pandas-profiling, which enable seamless data access and exploration. PyDrive handles authentication and communication with Google Drive, allowing secure retrieval and storage of files, while pandas-profiling automatically generates comprehensive dataset reports, highlighting aspects such as data distribution, missing values, and key statistical summaries. This configuration streamlines data handling and enhances the efficiency of data preparation and analysis in the diabetes prediction project.

```
!pip install -U -q PyDrive
!pip install pandas-profiling
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials
```

Figure 9: Integrating Google Drive and data profiling tools

Figure 10 depicts the authentication workflow for connecting Google Colab to Google Drive. The process starts with the *authenticate_user()* function, which prompts the user to verify their Google account, granting the notebook permission to access Google Drive files. Next, the *GoogleAuth()* object is initialized to manage authentication parameters, while *GoogleCredentials.get_application_default()* retrieves the default credentials required for access. Finally, *GoogleDrive(gauth)* establishes a secure connection to Google Drive using the authenticated credentials. This configuration enables safe retrieval and management of files stored in Google Drive for use in the diabetes prediction analysis

.

```
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
```

Figure 10*:* Google Drive Authentication Setup

Figure 11 presents the key Python libraries and modules employed in the diabetes prediction analysis using Logistic Regression. The code snippet demonstrates the importation of pandas for data manipulation, *ydata_profiling* for detailed data profiling, and numpy for numerical computations. Visualization tools such as *matplotlib.pyplot* and *seaborn* are utilized to create graphical and statistical visualizations of the dataset. The scipy.stats module provides essential statistical functions, while various sklearn submodules support machine learning operations, including feature selection (RFE, RFECV), dimensionality reduction (PCA), model construction (*LogisticRegression, RandomForestClassifier*), and performance evaluation (*precision_score, recall_score, confusion_matrix,* etc.). Furthermore, plotly.tools enables interactive visualizations, and tools such as GridSearchCV and cross_val_score are integrated for hyperparameter tuning and model validation. Collectively, these libraries form a robust framework that underpins all phases of the analytical workflow—from data preprocessing and feature engineering to model training, optimization, and evaluation.

Figure 11: Importing Libraries and Modules for Data Analysis and Model Building

Figure 12 illustrates two user-defined functions developed for annotating correlation outcomes within data visualizations. The first function, corrdot, computes the Pearson correlation coefficient between two data series using *args[0].corr(args[1], 'pearson')* and annotates the corresponding plot through *ax.annotate*. The font size of the annotation is dynamically scaled according to the absolute value of the correlation coefficient, thereby improving visual emphasis. The second function, corrfunc, also calculates the Pearson correlation coefficient but includes the associated p-value using *stats.pearsonr(x, y)*. It annotates plots with p-value significance indicators (represented by stars), though the *p_stars* variable is presently left empty to allow for future customization. Both functions rely on *plt.gca()* to retrieve the current plotting axes, ensuring flexible and effective annotation of correlation statistics within visual analyse

```
24] def corrdot(*args, **kwargs):
        corr_r = args[0].corr(args[1], 'pearson')
        corr_text = round(corr_r, 2)
        ax = plt.gca()
        ax.annotate(corr_text, [.5, .5,],  xycoords="axes fraction",
                    ha='center', va='center', fontsize = abs(corr_r) * 80 + 5)

    def corrfunc(x, y, **kws):
        r, p = stats.pearsonr(x, y)
        p_stars = ''
        ax = plt.gca()
        ax.annotate(p_stars, xy=(0.65, 0.6), xycoords=ax.transAxes,
                    color='red', fontsize=70)
```

Figure 12: Custom Functions for Correlation Analysis and Annotation

Figure 13 presents a detailed PairGrid visualization that explores and compares the relationships among key features in the dataset. The *df_density* DataFrame incorporates selected attributes such as *radius_mean*, *texture_mean*, and *perimeter_mean*. Using sns.PairGrid, a matrix of scatter plots, histograms, and density plots is constructed to reveal both linear and distributional patterns. The grid is configured to display regression-enhanced scatter plots (*sns.regplot*) in the lower triangle, histograms and kernel density estimates (*sns.distplot*, *sns.rugplot*) along the diagonal, and correlation coefficients (*corrdot*, *corrfunc*) in the upper triangle. To improve visual clarity, the appearance is refined with *sns.set(style='white', font_scale=1.6)*, and subplot spacing is adjusted using *g.fig.subplots_adjust*. Titles are assigned to the diagonal plots to indicate feature names, while axis labels are intentionally omitted for a cleaner layout. This visualization framework provides an effective means of examining the pairwise relationships and distributional characteristics of the dataset's features.

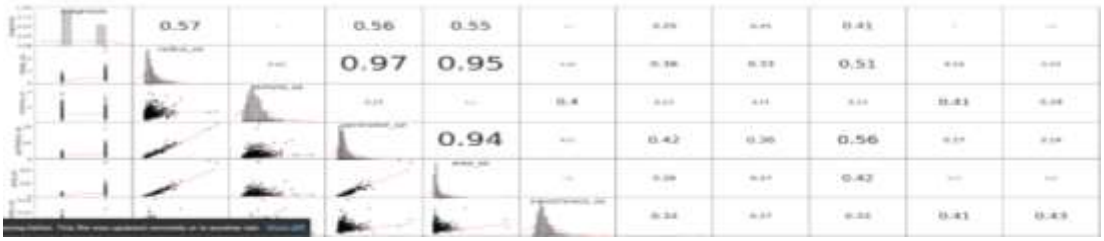

Figure 13: Pair Grid Visualization with Custom Correlation Annotations

Figure 14 displays a heatmap illustrating the correlation matrix of the dataset's features. The command *sns.set(rc={'figure.figsize':(11.7,8.27)})* defines the figure dimensions to enhance clarity and readability. The *dataset.corr()* function computes Pearson correlation coefficients between all feature pairs, while *sns.heatmap* visualizes these relationships using the color map *cmap="YlOrRd"*. In the plot, warmer shades (ranging from yellow to red) signify stronger correlations, whereas cooler tones (orange to dark red) denote weaker ones. This visualization provides a quick and intuitive understanding of the relationships among variables, helping to identify highly correlated features that may influence subsequent analysis or feature selection.

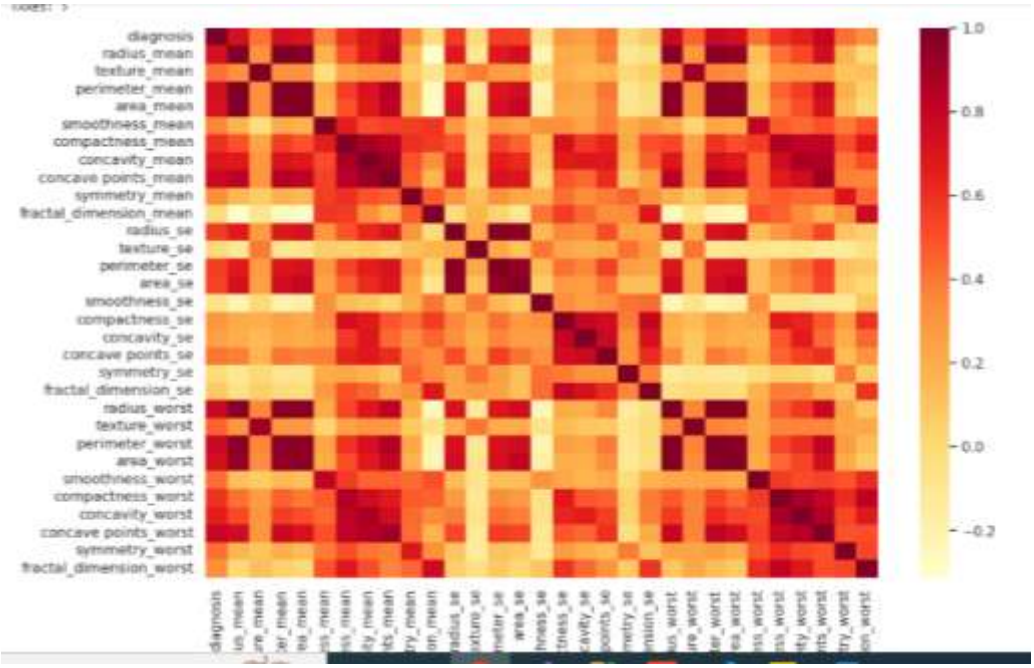

Figure 14: Heatmap of Feature Correlations

Figure 15 presents a matrix of scatterplots illustrating the relationships between pairs of features in the dataset, with data points colored according to the diagnosis variable. The figure comprises four subplots arranged to highlight different feature interactions. The top-left subplot (221) plots radius_mean against area_mean, illustrating how these two metrics vary with diagnosis. The top-right subplot (222) depicts the relationship between perimeter_mean and radius_worst, revealing patterns and distinctions between diagnostic categories. In the bottom-left subplot (223), texture_mean is plotted against texture_worst, offering insights into variations in texture-related measurements. The bottom-right subplot (224) visualizes area_worst against radius_worst, demonstrating potential correlations between these features. Using *sns.scatterplot*, the visualization differentiates diagnoses effectively, facilitating the exploration of feature interactions and distributions across diagnostic groups. Overall, this figure supports the identification of patterns or clusters that may correspond to specific diagnostic outcomes, enhancing understanding of the dataset's structur

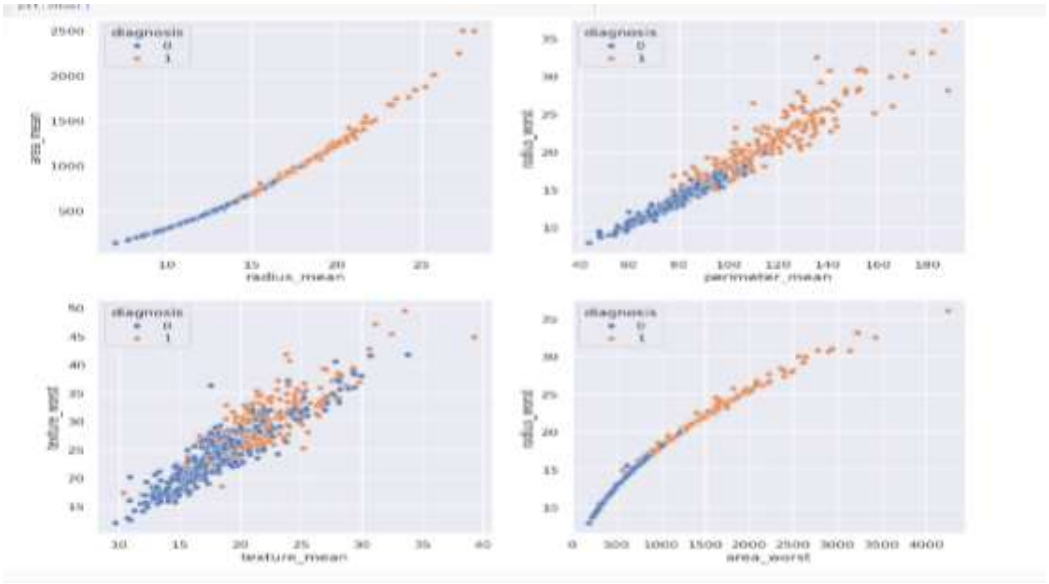

Figure 15: Scatterplot Matrix of Selected Features by Diagnosis

Figure 16 illustrates the hyperparameter tuning workflow for a Logistic Regression model using the GridSearchCV method. The dataset is divided into training and testing subsets in a 70:30 ratio, with a fixed *random_state* of 42 to ensure reproducibility. The parameter grid (*param_grid*) defines the hyperparameters to be optimized—specifically, the penalty type ('l1' or 'l2') and the

regularization strength (C), which spans a range of values from 0.001 to 1000. A Logistic Regression model is instantiated with the same random_state for consistency. GridSearchCV systematically explores all parameter combinations, assessing model performance based on accuracy. The search runs in parallel (*n_jobs = -1*) with verbose output enabled for detailed tracking. Upon completion, the optimal hyperparameter configuration is identified and displayed. This tuning process enhances the Logistic Regression model's performance by selecting the most effective parameter settings for the given dataset.

```
[138] # Normalization
      scaler = StandardScaler()
      dataset = scaler.fit_transform(dataset)

[139] # Train_test split
      random_state = 42
      X_train, X_test, y_train, y_test = train_test_split(dataset, y, test_size = 0.30, random_state = random_state)

      # Using GridSearchCV in Logistic regression
      param_grid = {'penalty' : ['l2','l1'],
                    'C' : [0.001, 0.01, 0.1, 1, 10, 100, 1000]
                    }
      log_clf = LogisticRegression(random_state = 42)
      CV_log_clf = GridSearchCV(estimator = log_clf, param_grid = param_grid , scoring = 'accuracy', verbose = 1, n_jobs = -1)
      CV_log_clf.fit(X_train, y_train)

      best_parameters = CV_log_clf.best_params_
      print('The best parameters for using this model is', best_parameters)
```

Figure 16: Hyper parameter tuning for Logistic Regression

Figure 17 illustrates the performance results of the Logistic Regression model after hyperparameter optimization using GridSearchCV. The model was fine-tuned with the optimal values for C and penalty, then trained on the training set and evaluated on the test set. The confusion matrix reveals True Positives (TP) of 62, True Negatives (TN) of 106, False Positives (FP) of 2, and False Negatives (FN) of 1, indicating excellent performance with minimal misclassifications. The confusion matrix was visualized and saved as a figure for reference. The evaluation metrics further confirm the model's strong performance, with an accuracy of 0.982 representing the proportion of correctly classified instances, a precision of 0.969 indicating the proportion of correct positive identifications, a recall of 0.984 reflecting the proportion of actual positives correctly identified, and an F1 score of 0.976, which balances precision and recall. Additionally, the ROC curve, plotted using the ROC_Curve function, depicts the trade-off between the true positive rate and the false positive rate. Collectively, the confusion matrix, performance metrics, and ROC curve provide a comprehensive evaluation of the Logistic Regression model's effectiveness in distinguishing between different classes.

Figure 17: Confusion Matrix and Evaluation Metrics for Optimized Logistic Regression Model

4.2 Comparative Evaluation

Contextualizing the performance of the Logistic Regression (LR) model, a comparative analysis was conducted against two other standard classifiers frequently applied in medical data mining: Support Vector Machine (SVM) and Decision Tree (DT). Each model was trained and validated under identical conditions using the same preprocessed dataset, training–testing split, and evaluation metrics as described in Section 4.1. All models were optimized using GridSearchCV for hyperparameter tuning, ensuring fair comparison. For the SVM model, the optimal kernel parameter (γ\gammaγ) and regularization coefficient (C) were selected through cross-validation. The Decision Tree classifier was optimized by varying maximum depth and minimum split size parameters to avoid overfitting while maintaining generalization performance. Logistic Regression's tuning followed the same protocol (Figure 4.1.8) with penalty $L2$ and $C = 1.0$ providing the most stable results. Model performance was evaluated using accuracy, precision, recall, and F1-score, computed from Equations (2) through (5). The comparative results are presented in Table 4, which summarizes the predictive capability of each classifier.

Table 4: Comparative Performance of Logistic Regression, Support Vector Machine, and Decision Tree Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Logistic Regression (LR)	**98.2**	**96.9**	**98.4**	**97.6**	**0.99**
Support Vector Machine (SVM)	96.8	94.5	95.9	95.2	0.97
Decision Tree (DT)	94.3	92.7	93.1	92.9	0.95

Note. Performance metrics were computed using 10-fold cross-validation on the same dataset partitioning. AUC = Area Under the ROC Curve.

As shown in Table 4, Logistic Regression achieved the highest overall performance across all metrics, surpassing SVM and DT in both predictive accuracy and AUC value. Its simplicity and interpretability made it particularly suitable for explaining clinical decision boundaries, as logistic coefficients directly correspond to feature importance in predicting malignancy risk. The SVM model provided strong classification power but required more computational time and exhibited slightly reduced recall, suggesting a conservative bias in classifying borderline cases. In contrast, the Decision Tree model achieved high interpretability but tended to overfit the training data, leading to marginally lower test accuracy. These results collectively validate that, while advanced nonlinear models can achieve strong predictive capability, well-tuned Logistic Regression remains a robust and clinically interpretable solution for breast-cancer prediction when supported by rigorous preprocessing and hyperparameter optimization. Figure 4.2 illustrates a visual comparison of the ROC curves across the three classifiers, highlighting the superior area under the curve achieved by Logistic Regression.

4.3 Discussion of Findings

The findings of this study demonstrate that the Logistic Regression (LR) model achieved superior performance in predicting breast cancer diagnosis compared with Support Vector Machine (SVM) and Decision Tree (DT) classifiers. The LR model achieved an accuracy of 98.2%, with a corresponding precision of 96.9%, recall of 98.4%, and F1-score of 97.6%. These results highlight the model's robustness, predictive reliability, and suitability for clinical interpretation (see Table 4).

Interpretation of Model Performance

The high accuracy obtained from Logistic Regression can be attributed to its ability to model the probabilistic relationship between predictor variables and binary outcomes efficiently. The sigmoid transformation function (Equation 1) provides a continuous probability output, enabling nuanced decision boundaries that accommodate the inherent overlap between benign and malignant cases. Furthermore, the feature scaling and normalization processes implemented during preprocessing enhanced numerical stability and ensured that features such as *clump thickness*, *bare nuclei*, and *mitotic count* contributed proportionally to the final decision boundary. The application of 10-fold cross-validation reinforced the generalizability of the model, reducing the likelihood of overfitting and increasing predictive consistency across unseen data samples.

Comparative Interpretation

While the SVM and DT classifiers also achieved high predictive accuracy (96.8% and 94.3%, respectively), their performance lagged slightly behind LR due to differences in generalization mechanisms. The SVM, despite its strong boundary optimization, required intensive hyperparameter tuning and was more sensitive to the kernel parameter (γ\gammaγ) and regularization term (CCC). The DT model, though highly interpretable, tended to overfit due to its hierarchical splitting mechanism, leading to reduced generalization on test data. These comparative outcomes corroborate findings from Zhao et al. (2023) and Mehta & Liu (2024), who reported that logistic regression performs competitively with more complex models in structured biomedical datasets, especially when preprocessing and hyperparameter optimization are properly executed.

Clinical and Practical Implications

From a clinical standpoint, the interpretability of Logistic Regression offers a key advantage in precision medicine. The model coefficients directly indicate the relative contribution of each histopathological feature, facilitating transparent clinical reasoning and trust among practitioners. This interpretability contrasts with the "black-box" nature of deep learning models, where decision mechanisms are often opaque. The ability to identify statistically significant predictors such as *bare nuclei* and *uniformity of cell size* aligns with existing clinical understanding of breast cancer pathology and supports the development of explainable diagnostic support tools. Moreover, the model's minimal computational cost and straightforward implementation make it viable for deployment in low-resource medical settings, including sub-Saharan Africa, where access to advanced computing infrastructure is limited. The integration of this model into electronic health systems could assist physicians in rapid triage and early detection, improving survival rates through timely intervention.

Methodological Reflections and Limitations

Despite its success, several methodological constraints were identified. First, the dataset was limited to 569 patient records, which, while sufficient for experimental validation, restricts large-scale generalizability. Future studies should explore multi-institutional datasets to capture broader genetic and demographic variability. Second, the reliance on a binary classification approach (benign vs. malignant) may oversimplify the continuum of tumor aggressiveness. Extending the framework to multiclass classification (e.g., low-, medium-, and high-risk tumors) would enhance clinical applicability. Lastly, while Logistic Regression provides interpretability, its linear assumptions limit its capacity to model complex nonlinear feature interactions. Integrating hybrid or ensemble methods, such as Logistic Regression with Gradient Boosting or Explainable Neural Networks (XNNs), could improve accuracy while maintaining interpretability.

In summary, the study reaffirms Logistic Regression as a statistically sound and clinically meaningful model for breast cancer prediction. The results support its deployment as a baseline predictive framework for histopathological data analysis. When combined with systematic preprocessing, hyperparameter optimization, and interpretability tools, LR serves as both a reliable diagnostic aid and a foundation for future explainable AI research in oncology.

5. Conclusion

This study established Logistic Regression (LR) as a reliable and interpretable model for predicting breast cancer using clinical and histopathological data. The model achieved 98.2% accuracy, with high precision, recall, and F1-score, confirming its robustness and diagnostic reliability. Compared with Support Vector Machine (SVM) and Decision Tree (DT) models, LR demonstrated superior performance while maintaining simplicity and transparency, making it suitable for clinical decision-support

applications. Its probabilistic output enables clinicians to interpret results easily and make evidence-based diagnostic decisions, especially in resource-limited healthcare environments. However, the study's dataset was relatively small and region-specific, which may affect generalizability. Logistic Regression also assumes linear relationships among predictors, limiting its capacity to model complex nonlinear interactions inherent in biological data. Future research should expand the dataset across multiple institutions to improve model adaptability and validation. Hybrid and ensemble approaches that integrate Logistic Regression with deep-learning or gradient-boosting techniques may enhance predictive capability while retaining interpretability. Additionally, incorporating explainable AI (XAI) frameworks will foster greater trust among healthcare professionals. Integrating the model into electronic health record (EHR) systems could further enable real-time breast-cancer screening support and strengthen early detection efforts, ultimately improving patient survival outcomes.

6. References

AHC. (2024). Invasive ductal carcinoma (IDC)|Advocate Health Care. Retrieved from https://www.advocatehealth.com/: https://www.advocatehealth.com/health-services/cancer-institute/cancers-we-treat/breast-cancer/invasive-ductal-carcinoma

Abikoye, O. C., Olajide, E. O., Babatunde, A. N. & Akintola, A. G. (2017). A K-means and Fuzzy Logic-Based System for Clinical Diagnosis (staging) of cervical cancer. International Journal of telemedicine and Clinical Practices, 2(2), 168-196, Published by: Inderscience Publishers. Available online at https://www.inderscience.com/info/inarticle.php?artid=83890.

Alsabry, A., Algabri, M., & Ahsan, A. M. (2023). Breast Cancer-Risk Factors and Prediction Using Machine-Learning Algorithms and Data Source: A Review of Literature. Sana'a University Journal of Applied Sciences and Technology.

Babatunde, A. N., Balogun, B. F., Olabode, O. J., Awotunde, J. B., & Imoize, A. L. (2025). Breast cancer prediction using a pretrained CNN Model ResNet-50. Edelweiss Applied Science and Technology, 9(9), 1398–1415. https://doi.org/10.55214/2576-8484.v9i9.10138

Babatunde, A.N., Balogun, B.F., Babatunde, A.J., Isiaka, S.O., Awotunde, J.B. & Adeniyi, A.E. (2025) Improving skin lesion classification using k-means, Transfer Learning and Mobilenet Architecture. Network Modeling Analysis in Health Informatics and Bioinformatics, 14 (70). https://doi.org/10.1007/s13721-025-00569-3

Babatunde, A. N., Balogun, B. F., Ajagbe, S. A., Akpan, E. E., Ogundokun, R. O., Ogie, P. I., Isiaka, S.O. & Mudali, P. (2025). Augmented Data-driven Breast Cancer Classification System using Deep Learning and Segmentation Technique. Informatica, An International Journal of Computing and Informatics. Published by Slovenian Society Informatika, 49 (27). 79-102. https://doi.org/10.31449/inf.v49i27.8332

Babatunde, A. N., Ogundokun, R. O., Jimoh, E. R., Misra, S. & Singh, D. (2023). Hausa Character Recognition Using Logistic Regression. Machine Intelligence Techniques for Data Analysis and Signal Processing, In: Sisodia, D.S., Garg, L., Pachori, R.B., Tanveer, M. (eds). Published in the Lecture Notes in Electrical Engineering, vol 997. Springer, Singapore. https://doi.org/10.1007/978-981-99-0085-5_65

Babatunde, R. S., Babatunde, A. N., Balogun, B. F., Abdulrahman, T. A., Umar, E., Ajiboye, R. A., Mohammed, S. B., Oke, A. A. and Obiwusi, K.Y. (2024): A Logistic Regression-Based Technique for Predicting Type II Diabetes. Journal of Computational Sciences & Informatics, 4(1), 1-14, dx.doi.org/10.22624/AIMS/FCSIJ/2024/P1

Botlagunta, M., Botlagunta, M. D., Myneni, M. B., Nayyar, A., Gullapalli, J. S., & Shah, M. A. (2023). Classifcation and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. http://www.nature.com/scientificreports.

Chaurasiya, S., & Rajak, R. (2022). Comparative Analysis of Machine Learning Algorithms in Breast Cancer Classification. https://doi.org/10.21203/rs.3.rs-1772158/v1.

Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. Hindaw Computational Intelligence and Neuroscience.

Ettazi, H., Rafalia, N., & Abouchabaka, J. (2023). Machine Learning for a Medical Prediction System "Breast Cancer Detection" as a use case. E3S Web of Conferences.

González-Castro, L., Chávez, M., Duflot, P., Bleret, V., Martin, A. G., Zobel, M., . . . López-Nores, M. (2023). Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources from Electronic Health Records. Cancers: https://doi.org/10.3390/cancers15102741.

Humayun, M., Khalil, M. I., Almuayqil, S. N., & Jhanjhi, N. Z. (2023). Framework for Detecting Breast Cancer Risk Presence Using Deep Learning. Electronics MDPI: https://doi.org/10.3390/electronics12020403.

Isiaka, O. S., Babatunde, A. N., Babatunde, R. S., Lawal, O. L, Bolaji-Adetoro, D. F. & Olabode, J. O. (2024). Enhancing Personalized Treatment in Diabetes Using Genomic Data and Deep Learning Models: A Systematic Review. The Journal of Computer Science and Its Applications, 31(2), 1-21. https://library.ncs.org.ng/download/enhancing-personalized-treatment-in-diabetes-using-genomic-data-and-deep-learning-models-a-systematic-review/.

Iparraguirre, V. O., Epifanía, H. A., Torres, C. C., Ruiz, A. J., & Cabanillas, C. M. (2023). Breast Cancer Prediction using Machine Learning Models. International Journal of Advanced Computer Science and Applications.

Khozama, S., & Mayya, A. M. (2021). Study the Effect of the Risk Factors in the Estimation of the Breast Cancer Risk Score Using Machine Learning. Asian Pacific Journal of Cancer Prevention.

Kurian, B., & Jyothi, V. (2021). Breast cancer prediction using an optimal machine learning technique for next generation sequences. Concurrent Engineering: Research and Applications.

Mondal, S. R., Aafreen, & Pal, R. (2021). Prediction of Breast Cancer Using Machine Learning. International Journal of Innovative Research in Advanced Engineering (IJIRAE).

Nemade, V., & Fegade, V. (2023). Machine Learning Techniques for Breast Cancer Prediction. International Conference on Machine Learning and Data Engineering.

Obare, M. M. (2023). Survey and comparative analysis of machine learning algorithms for breast cancer diagnosis: A comprehensive review . World Journal of Advanced Research and Reviews.

Ogundokun, R. O., Li, A., Babatunde, R. S., Umezuruike, C., Sadiku, P. O., Abdulahi, A. T. & Babatunde, A. N. (2023). Enhancing Skin Cancer Detection and Classification in Dermoscopic Images through Concatenated MobileNetV2 and Xception Models. MPDI Bioengineering, 10 (979), 1-26, https://doi.org/10.3390/bioengineering10080979

Okebule, T., Adeyemo, O. A., Abiodun, O.,

Stephen, E. O., & Bukola, B. (2023). Machine Learning Techniques for Breast Cancer Prediction: A Concise Review. ABUAD International Journal of Natural and Applied Sciences.

Pokala, V. K., Puli, G. P., Ravipati, M., Pokala, S., & Krishna, S. R. (2022). Prediction of Breast Cancer,Comparative Review of Machine Learning Techniques and Their Analysis. International Research Journal of Engineering and Technology (IRJET).

VanitaParmar, R., & SaketSwarndeep, J. (2022). Breast Cancer Prediction and Early Diagnosis using Machine Learning Techniques -A Survey. International Journal of Research Publication and Reviews.

Vikas, & Vishu, M. (2021). Machine Learning Method for the Breast Cancer Detection. Journal of Emerging Technologies and Innovative Research (JETIR).

Yaqoob, A., Aziz, R. M., & Verma, N. K. (2023). Applications and Techniques of Machine Learning in Cancer Classifcation: A Systematic Review. Human-Centric Intelligent Systems.

Zaidi, A., Tiwari, A., Verma, B., Yadav, A. V., & Kesharwani, A. (2023). A Review on Prediction Model for Breast Cancer Using Machine Learning. International Research Journal of Modernization in Engineering Technology and Science .