2025 13th International Conference on Intelligent Embedded, MicroElectronics, Communication and Optical Networks (IEMECON) | 979-8-3315-7201-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/IEMECON69302.2025.11365819

Study on Prediction of Breast Cancer Tumor Detection, based on Biological Properties

Subhadra Bhattacharyya
Dept. of ECE
Institute of Engineering & Management
University of Engineering & Management
Kolkata, India
Subhadra.Bhattacharyya2023@iem.edu.in

Sounak Maiti
Dept. of ECE
Institute of Engineering & Management
University of Engineering & Management
Kolkata, India
Sounak.Maiti2023@iem.edu.in

Amit Kumar Das
Dept. of ECE
Institute of Engineering & Management
University of Engineering & Management
Kolkata, India
amit.das@iem.edu.in

***Abstract*—This study implements a logistic regression model for breast cancer classification using the Wisconsin Breast Cancer Dataset containing 569 samples with 30 morphological features. The model achieved 94.95 percent training accuracy and 92.98 percent test accuracy in classifying malignant from benign tumors. Performance was evaluated using Receiver Operating Characteristics (ROC) curve analysis (Area Under the Curve is 0.93) and confusion matrix visualization. SHapley Additive exPlanations values were integrated to provide model interpretability by identifying the most influential features for classification decisions. The results demonstrate effective machine learning classification combined with explainable Artificial Intelligence techniques for medical diagnosis applications.**

***Keywords*—Breast cancer classification, logistic regression, SHAP, explainable AI, Cancer diagnosis.**

I. INTRODUCTION

Nowadays Breast cancer is a very common and Life-threatening disease which is affecting women throughout the whole world. Early classification of this disease can help to improve the survival rates. By using Machine Learning it can become very easier for the healthcare professionals to take the right decision as early as possible. In this research paper we developed a Machine Learning model, that can classify the cells, whether it is malignant or benign based on the dataset4. It contains 569 data points and 31 numeric features describing the characteristics of the cells present in the breast part. We have implemented Logistic Regression to train our classification model. By using this model, we achieved approximately 94.95% accuracy on training data and 92.98% accuracy on testing data. This enhances the model transparency and provides clinical accuracy by linking the predictive data with the biological characteristics of the cells.

II. METHODOLOGY

The methodology used in this study is structured into various key phases: data acquisition, preprocessing, dimensionality reduction, model training, evaluation, and interpretability. The following steps summarizing the approach:

A. Dataset Acquisition

The dataset used for developing this system is called Breast Cancer Wisconsin Diagnostic Dataset (WBCD), which is available directly through the Scikit-learn library. It contains 569 instances with 30 numeric features describing characteristics of cell nuclei present in digitized images of breast tissue. It is denoted by binary process:

- [0] for malignant tumors.
- [1] for benign tumors.

Also, there are many attributes for understanding the dataset like radius, texture, perimeter, area, smoothness, concavity, compactness, fractal dimension, symmetry etc. Since these attributes are used for describe the cells so it is represented as mean like mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean concavity like this.

B. Data Preprocessing

The data set is very complicated but after framed it looks simple and every attribute makes it easier. At first the dataset is loaded from Sklenar then it is framed by using "features" command from panda's library, here is the figure of the end result.

Table 1: Head of the framed dataset

Feature	Sample 0	Sample 1	Sample 2	Sample 3	Sample 4
Mean Radius	17.99	20.57	19.69	11.42	20.29
Mean Texture	10.38	17.77	21.25	20.38	14.34
Mean Perimeter	122.8	132.9	130	77.58	135.1
Mean Area	1001	1326	1203	386.1	1297
Mean Smoothness	0.1184	0.08474	0.1096	0.1425	0.1003
Mean Compactness	0.2776	0.07864	0.1599	0.2839	0.1328
Mean Concavity	0.3001	0.0869	0.1974	0.2414	0.198
Mean Concave Points	0.1471	0.07017	0.1279	0.1052	0.1043
Mean Symmetry	0.2419	0.1812	0.2069	0.2597	0.1809
Mean Fractal Dimension	0.07871	0.05667	0.05999	0.09744	0.05883

After that we checked the dataset again if there any missing values present or not then we dropped that column and proceed further.

C. Exploratory Data Analysis (EDA)

Correlation heatmaps graph (Fig. 2) was showed to identify relationships between various kind of features. Principal Component Analysis (PCA) was built in this project to split the data into two dimensions for visualization and to observe the separation of classes. Feature distributions were also analyzed using class-wise grouping to identify statistically significant differences.

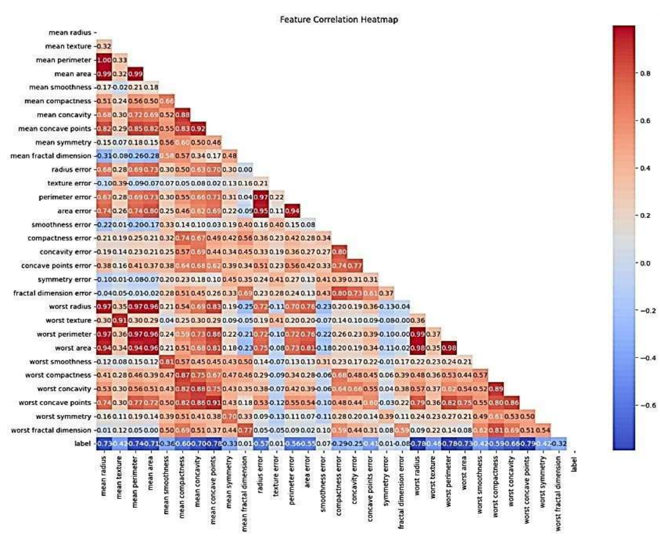

Fig. 2: Correlation Heatmap

D. Model Training

The dataset was split into 80% of training data and 20% of testing data with a fixed random state for reproducibility. A Logistic Regression model is selected due to its interpretability, simplicity, and strong performance in binary classification tasks. The model is trained on the training subset using supervised learning. The trained data achieved approximately 95% accuracy which indicates the processes are well executed throughout the development of this system.

E. Evaluation and Interpretability

Model performance was measured using accuracy scores, Receiver Operating Characteristics (ROC-AUC) curves, and confusion matrices. Moreover, SHapley Additive exPlanations (SHAP) was used to interpret the contribution of individual features toward each prediction, providing insight into the model's decision-making process.

ROC curves are played very important role in this evaluation because it is necessary for evaluating binary classification model since the model is highly dependent on binary classification. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The curve helps visualize the trade-off between correctly identifying positive cases (TPR) and incorrectly classifying negative cases as positive (FPR) across different thresholds, with a model that hugs the top-left corner indicating better performance.

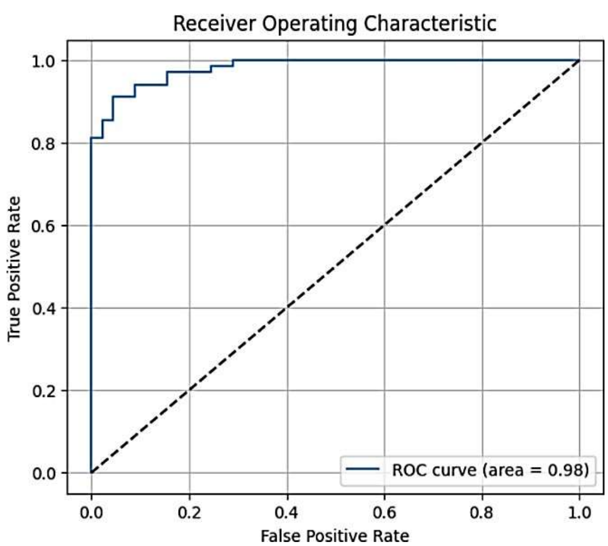

Fig. 3: Receiver Operating Curve

III. DATA PREPROCESSING

In machine learning, data preprocessing is a very crucial step for classification using the dataset. In this process cleaning, modeling and to realize the data there is various types of techniques used to increase the capability of classification and modeling. This process is done under the Non-Disclosure Agreement (NDA) while working with anonymized medical data.

The dataset used in this project, is already pre-cleaned and anonymized from Scikit-lean's breast cancer module. The basic requirements of NDA are already fulfilled. Each row of the dataset indicates a case of an individual patient to avoid complex encoding.

Data Preprocessing steps:

- Bunch objects are conversed into a Pandas-Data Frame for increasing the visualization of data.
- For classification purpose the name of the target variable is denoted by binary digits.
  [0] – for malignant
  [1] – for benign
- The verification process is done to check if there is any missing value or null entry. Due to this process, there is no need of imputation.
- To assess multicollinearity among the features, the correlation matrix was computed and visualized.
- Reduction of dimensionality is done by using principal component analysis (PCA) and also to increase the visualization of class separability in 2D space. As a result, it helps to figure out which

principal component contained the most discriminative information.

IV. MODEL ANALYSIS

In this model the use of logistic regression is very crucial. This is a unique type of algorithm, which is useful for binary classification. This model helps to find out the accuracy of the model.

A. Model Selection and Training

Logistic Regression was chosen to identify clarity, simplicity at the linearly separable data. After splitting the dataset into training and testing ratio (80:20), the model was trained using the training subset using a constant regularization (L2 penalty) and the maximum 10,000 iterations to ensure the convergence.

B. Performance Metrics

The model passed the maximum percentage of accuracy:

a. Training Accuracy: 94.95%
b. Testing Accuracy: 92.98%

The categorization breakdown was displayed in a confusion matrix, showing good reliability for clinical application with extremely few false positives and false negatives. The discriminating capacity of the model was confirmed by the ROC-AUC score, which was calculated to be substantially above 0.9 (Fig. 2).

The calculated metrics evaluated as:

c. Precision and Recall for each label-class.
d. F1 Score.
e. ROC Curve representation.

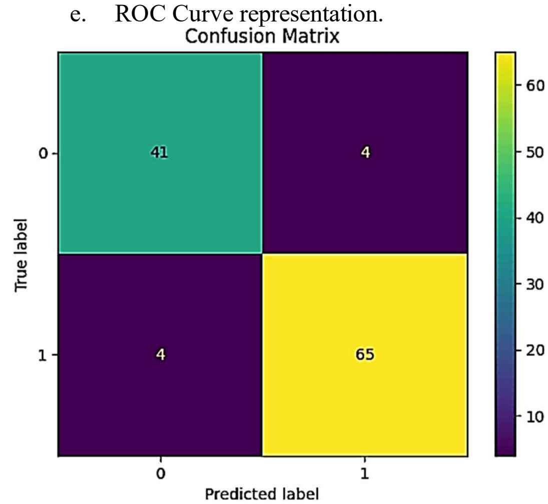

Fig. 4: Confusion Matrix

C. Feature Contribution & Interpretability

To make the betterment of the model, Shapley Additive explanations (SHAP) values were included. The most significant factors influencing the forecasts may be found using SHAP plots; parameters like "worst perimeter," "mean concavity," and "worst concave points" are of great significance. Clinical decision-making trust and model transparency are enhanced by this interpretability.

D. Usage of SHAP

SHapley Additive exPlanations is a post-hoc, model-agnostic technique for model interpretability. It assigns an important value to each feature in a model's prediction, showing how much each feature contributes to a specific outcome.

We have also used SHAP for interpreting the value of the feature of this model. At first, we have created a SHAP explainer object which will help us to interpret the model which is developed by using logistic regression. After calculating the values of each feature by using SHAP we finally plot the data for visualization.

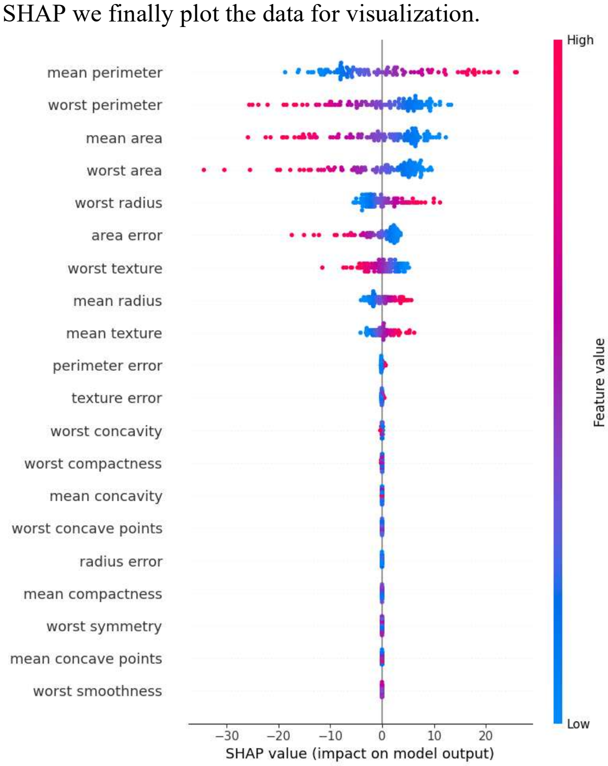

Fig. 5: Visualization of SHAP

Features are ordered from top to bottom by their average absolute SHAP value (most important at top) In this figure positive SHAP values (red) push the prediction toward malignant (1) and negative SHAP values (blue) push the prediction toward benign (0). Red dots indicate high feature values and blue dots indicate low feature values and each dot represents one patient. The spread shows how the feature affects different patients differently.

V. EFFICIENCY OF THE MODEL

In machine learning, efficiency refers to the model's computing performance, resource use, and generalizability across datasets in addition to prediction accuracy. In a number of characteristics, the breast cancer classification model created in this work has great efficiency.

A. Computational Efficiency

The Logistic Regression approach scales effectively with structured datasets and is computationally lightweight. Since training took less than a second on a typical CPU-based system without requiring GPU acceleration, the model is well-suited for deployment in low-resource environments such as clinics or embedded medical devices.

B. Model Complexity

The model maintains modest complexity and great interpretability through the use of logistic regression. It needs, little adjustment and prevents overfitting instead of increasing complexity. By shrinking the feature space while maintaining crucial information, the application of PCA for dimensionality reduction significantly improved computing efficiency.

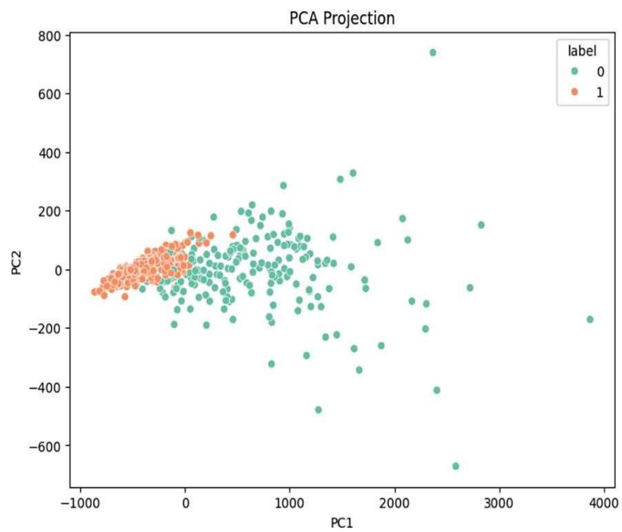

Fig. 6: Principal Component Analysis Projection

A. Generalization Capability

With a low performance difference between training and testing accuracy (94.95% vs. 92.98%), the model showed good generalization. When applied to unseen data, this supports its dependability and shows resistance against overfitting.

B. Real-time Applicability

The model is appropriate for real-time diagnostic applications because to its quick inference time and low processing requirements. It can be included into cloud-based platforms that demand quick responses, embedded diagnostic equipment, or mobile health applications.

VI. CONCLUSION

This study shows how machine learning methods, specifically Logistic Regression, may be applied to accurately and comprehensibly classify breast cancer. The model's excellent classification accuracy of 92.98% on testing data and 95% on training data. The development of a clear and reliable diagnostic tool was aided by many crucial project phases, including data preparation, feature correlation analysis, dimensionality reduction using PCA, and model interpretation using SHAP. The approach is more effective, less complicated, and ideal for use in clinical settings where quick and precise decision- making is essential. Our model confirms that interpretable machine learning models may be extremely helpful in enhancing cancer treatment planning and early diagnosis.

REFERENCES

[1] Predictive Modeling for Early Cancer Detection: A Machine Learning Approach to Prostate, Lung, and Breast Cancer, Chandrayani Rokde; Prajakta Kharwandikar; Amrita Kungwani; Anshu R. Dudhe, 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI), 10.1109/IDICAIEI61867.2024.10842741, 29-30 November, 2024.

[2] Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset, Atajan Rovshenov; Serhat Peker, 2022 3rd International Informatics and Software Engineering Conference (IISEC), 10.1109/IISEC56263.2022.9998248, 15-16 December 2022.

[3] Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, Youness Khourdifi; Mohamed Bahaj, 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), 10.1109/ICECOCS.2018.8610632, 05-06 December 2018.

[4] Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms, Sharmin Ara; Annesha Das; Ashim Dey, 2021 International Conference on Artificial Intelligence (ICAI), 10.1109/ICAI52203.2021.9445249, 05-07 April 2021.

[5] A Survey of Transfer Learning in Breast Cancer Image Classification, Jinghong Xu; Xinyou Dong, 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), 10.1109/IICSPI51290.2020.9332405, 28-30 November, 2020.

[6] Breast Cancer Classification Using Machine Learning, Yajush Tewari; Eshant Ujjwal; Lalit Kumar, 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 10.1109/ICACITE53722.2022.9823932, 28-29 April 2022.

[7] Evaluation of feature extraction methods with ensemble learning for breast cancer classification, Yuki Ono; Yoshihiro Mitani, 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), 10.1109/LifeTech53646.2022.9754789, 07-09 March 2022.

[8] A study on prediction of breast cancer recurrence using data mining techniques, Uma Ojha; Savita Goel, 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, 10.1109/CONFLUENCE.2017.7943207, 12-13 January 2017.

[9] Using Pretrained Vision Transformer for Breast Cancer Binary Classification, Esam Mohammed Asem Othman, 2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA), 10.1109/eSmarTA59349.2023.10293730, 10-11 October 2023

[10] A Comparative Study on Breast Cancer Prediction using Optimized Algorithms, S. Nathiya; J. Sumitha, 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 10.1109/ICOSEC51865.2021.9591787, 07-09 October 2021.

[11] Breast Cancer Prediction of Benign and Malignant Tumors by Classification Algorithms, Santad Promtan; Phungern Khongthong; Chidchanok Choksuchat, 2023 4th International Conference on Big Data Analytics and Practices (IBDAP), 10.1109/IBDAP58581.2023.10271967, 25-27 August 2023.

[12] Analysis of Various Machine Learning Classifiers for Prediction and Classification of Breast Cancer, Sneha; Sanjana Sharma B S; Ruchitha H K; Ashwini Kodipalli; Trupthi Rao; Rohini B R, 2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA), 10.1109/CIISCA59740.2023.00062, 22-23 June 2023.

[13] Machine Learning model for detection of Breast Cancer, Anshuman; Upendra Kumar, 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 10.1109/ISCON52037.2021.9702416, 22-23 October 2021.

[14] Using Machine Learning Methods in Early Diagnosis of Breast Cancer, Begüm Erkal; Tülin Erçelebi Ayyıldız, 2021 Medical Technologies Congress (TIPTEKNO),