A Novel Breast Cancer Screening System using Logistic Regression

Prasad S Sase
Computer science and Engineering Department,
Shri JJT University, Jhunjhunu, India
prasadsase@gmail.com

Debabrata Swain
Computer Science and Engineering Department,
Pandit Deendayal Energy University, Gandhinagar, India
debabrata.swain7@yahoo.com

Shailesh Kumar
Computer science and Engineering Department, Shri JJT University, Jhunjhunu, India
shaileshkumar1610@gmail.com

Varsha Prasad Sase
Department of Engineering Sciences,
Sinhgad Institute of Technology and Science,
Narhe, Pune, India
sasevarsha@gmail.com

Abstract

Breast cancer is appearing as one of the most life-threatening illnesses, mostly affecting the females throughout the world. Approximately 67000 female death cases were recoded due to this disease in 2022. Advance diagnosis of the disease, is the only way through which the patient lives can be saved. The disease at its primary stage does not give any symptom about its presence. The proposed approach used logistic regression for the early-stage detection of breast cancer disease. The logistic regression gives a probabilistic output which is mostly used for clinical decision-making process. The breast cancer dataset produced by Wiscon hospital is used for the training and validation of the model. For the feature selection process correlation method is used. The highest accuracy reported by the model is 96.5%.

Keyword- Breast cancer disease, Logistic regression, Min-Max scaling, Corelation score, ROC curve

1. Introduction

Cancer appears to be one of the greatest clinical challenges, resulting with substantial number of death cases every year through out the world. It's an indicator of a huge number of ailments, occurring due to the abnormal growth of the muscles that damages any organ of the human body drastically. Breast cancer has become one of the most common types of cancer, that is found with the females now a days in the society, causing a large number of death cases in the society. As per the world-wide statistics in 2022 year around 6,70,000 females' mortalities have reported because of breast cancer, demanding an efficient and effective screening method for early diagnosis. One of the most challenging problems in controlling the effect of breast cancer disease depends upon its initial stage screening.

The detection gets delayed and chances of survival becomes less in this disease because at its preliminary stage the illness shows either no or clinically unimportant indications.

Mostly the breast cancer disease starts its infection from the milk secretion portion of the breast known as lobule and duct [1]. The medical case studies shows that the most of the malignant tumours approximately 85% growth starts from the duct whereas the rest of cases found the presence in the lobule region [2].

The ductal network is responsible for transporting milk toward the nipple, whereas the lobules consist of glandular sacs involved in milk secretion. Malignant transformation within these structures can lead to aggressive tumour growth if not identified at an early stage [3].

Over the past decade, breast cancer incidence has surpassed that of lung cancer, making it the most commonly diagnosed cancer globally. Clinically, disease progression is categorized into five stages, ranging from stage

0 to stage 4. Stage 0 corresponds to non-invasive ductal carcinoma in situ, where abnormal cells are confined within the ducts. Stages 1 through 3 represent increasing tumour size and regional spread, while stage 4 is characterized by metastasis to distant organs such as the lungs, liver, bones, or brain. Timely detection of the disease when it has not spread to the other organs, helps the patient to get recovery from the disease and enhance the probability of survival [4].

Traditional breast cancer detection system greatly depends upon the different clinical imaging methods such as breast X-rays. These images are closely monitored by the experienced radiologist to discover mass or abnormal growth of muscle. The detection outcome greatly suffered with heterogeneity found in different female's breast denseness or hormonal changes. During the complex or high-risk cases, different invasive methods like biopsies or histopathological tests are being followed. Even though the tests are more effective but still, these are more costly, time consuming and creates a lot of mental stress within the patient.

In the current decade Artificial intelligence (AI) and Machine learning (ML) based methods have gained importance in solving many critical medical domain problems. These methods are capable of capturing patterns and derive hidden relations between different bio markers, that was extremely difficult to understand by other traditional methods used by medical science. By leveraging the pre-processed medical data, ML models can support in enhancing the prediction, minimize human error, and making faster clinical decisions.

In this work, logistic regression model has been used as a useful tool breast cancer screening. Logistic regression is primarily used for binary classification problem, based on probabilistic outcome. The probability score estimated by the model enables the system to distinguish between benign and malignant records. Due to the presence of corelated features and probabilistic nature of medical data, the logistic regression model helps to bifurcate the data records with less precision. For the relevant features selection, and redundant feature elimination, the multicollinearity and corelation analysis has been performed to enhance model reliability.

2. Literature

The author [5] employed 2 different AI methods for the screening of BCD. There were 17-features used by the models for the prediction. The decision tree classifier has reported the highest performance of 93% during validation.

There were 4 different machine learning methods [6] used for the screening of BCD. There were 11 number of biomarkers considered during the classification process. Among all different classifiers Support vector machine outperformed than other applied classifiers with an accuracy score of 95%.

For prior prediction [7] of BCD the Knn classifier was applied. The features were extracted from the mammogram image. The highest performance accuracy of 84% was obtained during the validation of the model.

The perceptron model [8] was used for the identification of the illness. The total records used in this case was 116. During data preprocessing label encoding was used to convert alphabet values to numeric. The perceptron model has reported the highest prediction accuracy of 85%.

There were 11 features [9] used for the screening of the disease. Various ML models were used during the implementation of the system. The K nearest neighbour model has reported boosted performance of 95% among all applied models.

The author [10] has established a breast cancer analysis system using gradient boosting model. During data analysis it was found that the dataset is not balanced. The total number of features used in this were 10. The highest accuracy shown by the model was 71%.

The author [11] has created a breast cancer screening system using RF and DT algorithm. The total number inputs used in the model was 13. The RF algorithm has shown the accuracy of 93%.

3. Proposed System

The block diagram of the proposed screening system is demonstrated in the following figure-1.

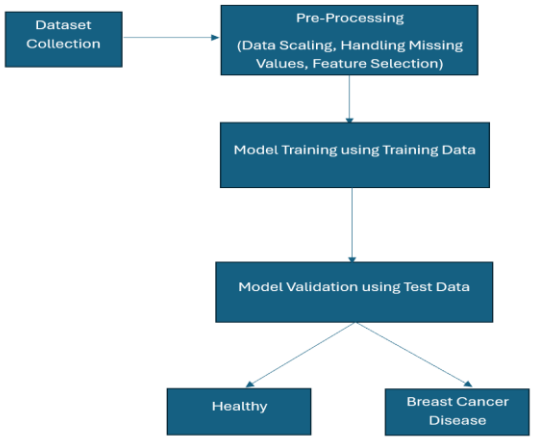

Fig 1. Proposed System Architecture

3.1 Dataset Collection

In this study the Breast cancer dataset of Wiscon is considered [12]. The dataset comprises of total 569 number of records. Each record in the dataset is either belongs to malignant (M) or benign (B) class type. There is total 32 features available in the dataset. Few features are of numeric whereas others are of type categorical. These features are mainly representing different biomarkers of breast whose status gives vital information about the status of breast cancer disease. The features are shown in the following table-1. The dataset is prepared after analysing the digital images of Wiscon hospital. During the analysis all the features present in the dataset were extracted from the image samples.

Table 1: Feature List

Feature Name	Base Characteristic	Description
ID	N/A	Patient unique identification number.
Diagnosis	N/A	Target variable: M (Malignant) or B (Benign).
radius_mean	Radius	**Mean value** of Radius. Significance: Size of the tumor.
radius_se	Radius	**Standard Error (Variability)** of Radius. Significance: Size of the tumor.
radius_worst	Radius	**Worst (Largest) value** of Radius. Significance: Size of the tumor.
texture_mean	Texture	**Mean value** of Texture. Significance: Standard deviation of gray-scale values (Roughness/Irregularity).
texture_se	Texture	**Standard Error (Variability)** of Texture. Significance: Standard deviation of gray-scale values (Roughness/Irregularity).
texture_worst	Texture	**Worst (Largest) value** of Texture. Significance: Standard deviation of gray-scale values (Roughness/Irregularity).
perimeter_mean	Perimeter	**Mean value** of Perimeter. Significance: Size of the tumor.
perimeter_se	Perimeter	**Standard Error (Variability)** of Perimeter. Significance: Size of the tumor.
perimeter_worst	Perimeter	**Worst (Largest) value** of Perimeter. Significance: Size of the tumor.
area_mean	Area	**Mean value** of Area. Significance: Size of the tumor.
area_se	Area	**Standard Error (Variability)** of Area. Significance: Size of the tumor.
area_worst	Area	**Worst (Largest) value** of Area. Significance: Size of the tumor.
smoothness_mean	Smoothness	**Mean value** of Smoothness. Significance: Local variation in radius lengths (Irregularity of contour).
smoothness_se	Smoothness	**Standard Error (Variability)** of Smoothness. Significance: Local variation in radius lengths (Irregularity of contour).
smoothness_worst	Smoothness	**Worst (Largest) value** of Smoothness. Significance: Local variation in radius lengths (Irregularity of contour).
compactness_mean	Compactness	**Mean value** of Compactness. Significance: Measure of shape/density: (Perimeter^2 / Area) - 1.0.
compactness_se	Compactness	**Standard Error (Variability)** of Compactness. Significance: Measure of shape/density: (Perimeter^2 / Area) - 1.0.
compactness_worst	Compactness	**Worst (Largest) value** of Compactness. Significance: Measure of shape/density: (Perimeter^2 / Area) - 1.0.
concavity_mean	Concavity	**Mean value** of Concavity. Significance: Severity of concave portions of the contour (Indentation).
concavity_se	Concavity	**Standard Error (Variability)** of Concavity. Significance: Severity of concave portions of the contour (Indentation).
concavity_worst	Concavity	**Worst (Largest) value** of Concavity. Significance: Severity of concave portions of the contour (Indentation).
concave points_mean	Concave Points	**Mean value** of Concave Points. Significance: Number of concave portions of the contour (Notch count).
concave points_se	Concave Points	**Standard Error (Variability)** of Concave Points. Significance: Number of concave portions of the contour (Notch count).
concave points_worst	Concave Points	**Worst (Largest) value** of Concave Points. Significance: Number of concave portions of the contour (Notch count).
symmetry_mean	Symmetry	**Mean value** of Symmetry. Significance: Symmetry of the cell nuclei.
symmetry_se	Symmetry	**Standard Error (Variability)** of Symmetry. Significance: Symmetry of the cell nuclei.
symmetry_worst	Symmetry	**Worst (Largest) value** of Symmetry. Significance: Symmetry of the cell nuclei.
fractal dimension_mean	Fractal Dimension	**Mean value** of Fractal Dimension. Significance: Complexity of the cell shape.
fractal dimension_se	Fractal Dimension	**Standard Error (Variability)** of Fractal Dimension. Significance: Complexity of the cell shape.
fractal dimension_worst	Fractal Dimension	**Worst (Largest) value** of Fractal Dimension. Significance: Complexity of the cell shape.

3.2 Data Pre-Processing

During the initial phase the id column was completely removed from the dataset as the column gives some arbitrary information. Also, this feature does not play any role in the prediction of breast cancer disease. In the next step during the pre-processing, the target column values like M and B were converted to 1 and 0. For this task label encoding method is used. For feature scaling, the Min-Max scaler method is used [13]. This method has brought the values of different features between 0 and 1. The Min-Max scaler is demonstrated in the following equation1.

$$K\_norm = (K - K\_min) / (K\_max - K\_min) \quad (1)$$

Where, K corresponds to the original feature value, K_min is the minimum value, and K_max is the maximum value of that feature in the dataset.

During this phase it is also found that the dataset is not containing any missing values. Therefore, no data imputation operation was carried out during the pre-processing stage. After analysing the data distribution, it is found that the dataset contains 63% benign records and 37% malignant records. For training and testing 70% and 30% of the total number of records are used.

3.3 Feature Selection

This phase is mainly conducted to find the useful or relevant features out of the total number of features present in the dataset. For finding the relevant features, first the heat map is drawn representing the correlation score between the different features as shown in the following figure-2.

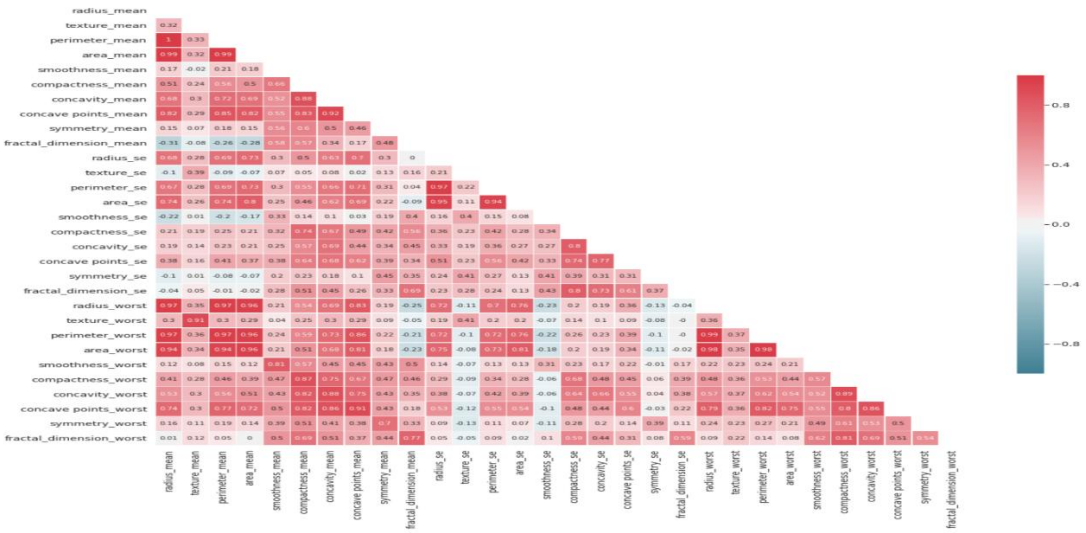

Fig. 2: Feature Correlation Map

The heat map shows that radius_mean, perimeter_mean, and area_mean features are having high correlation score. All these 6 features are representing the physical size of the cell. As these 6 features are representing the similar kind of data therefore only one feature is considered here by dropping the rest 4 features (perimeter_mean, perimeter_se, area_mean, area_se). All the mean and worst features are having high correlation with each other. Therefore, we kept the ten mean features and dropped the ten worst features from the dataset. The features like compactness, concavity, and concave points are having high correlation with each other. Therefore, the features related to concavity, and concave points are dropped from the dataset. In this way out of 30 input features 18 features are dropped from the dataset. The count of final set of input features remaining in the dataset is 12. The correlation plot of the final set of features is shown in the following figure-3.

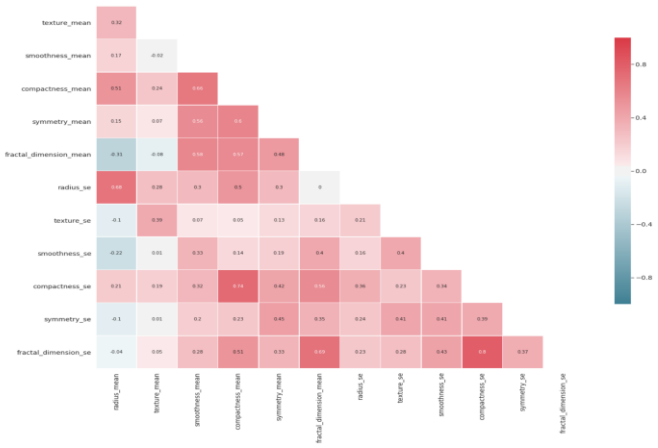

Fig. 3 Final Feature set Correlation plot

3.4 Model Training

Logistic Regression

It is supervised linear classification algorithm [14] that is used to classify a data point to a specific class by taking the linear combination of the input features. The outcome is a probability score that is based on sigmoid function. The output returned by the function is either 0 or 1. It classifies a point to the output class 1 if the probability score is more than the threshold value (0.5), otherwise it will be labelled as 0. The following equation 2 and 3 demonstrates the behaviour of logistic regression. The loss function finds the difference between predicted and actual output value demonstrated in equation 4. For classifying the points, the algorithm finds the optimum value of slope and intercept at which the error will be minimum.

Given features $x \in \mathbb{R}^d$ and parameters $\mathbf{w}$ and b, the model computes

$$z = \mathrm{w}^\top x + b \qquad (2)$$

$$p(y = 1 \mid x) = \sigma(z) = \frac{1}{1+e^{-z}} \qquad (3)$$

$$\mathrm{L(w, b)} = -\sum_{i=1}^{N} y_i \log y_i + (1 - y_i) \log(1 - y_i) \qquad (4)$$

4. Result

The performance of the proposed model is evaluated using different performance factors [15] like accuracy, precision, recall, F1-Score, Confusion matrix, ROC curve and PR curve. For evaluation 30% (171 records) of the total dataset is considered. The following table-2 shows the performance details of the model. The table-3 has shown the confusion matrix. The figure-4 and 5 has shown the ROC and PR curve [16] of the model. These curves score shows the efficiency of the models during classification process.

Table-2: Performance Table

Performance Factor	Percentage
Accuracy	96.5%
Precision	96.6%
Recall	96.5%
F1-Score	96.5%

Table-3: Confusion Matrix

TP	**54**
FP	**4**
TN	**111**
FN	**2**

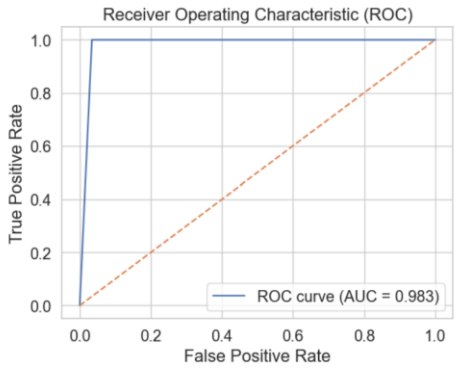

Figure-4 ROC Curve

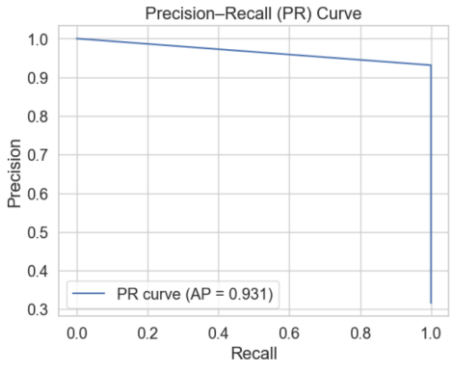

Figure-5 PR Curve

5. Comparative Analysis

Table-4: Comparative Analysis Table

Ref No.	**Model**	**Accuracy**
5	Decision Tree	93%
6	SVM	95%
7	KNN	84%
8	MLP	85%
9	KNN	95%
10	Gradient Boosting	71%
11	Random Forest	93%
Proposed System		**96.5%**

A brief comparative analysis is demonstrated between the proposed method and other literatures is demonstrated in the above table-4. In medical usecase Logistic regression model is used most of the time because the result produced by the model is more clinically interpretable. It also sets the value of the coefficients as odd ratio that represents the importance of the risk parameter. The different risk parameters always the influence the status of the disease based on their effect towards the disease. Compared with decision trees, random forests, gradient boosting, SVMs, KNN, and MLPs, logistic regression is more transparent, easier to audit, and better aligned with epidemiological practice, making it suitable for regulatory and ethical validation in healthcare. It also performs reliably when sample sizes are limited or data are noisy, which is common in hospital settings and rare-disease studies, while many advanced models require large datasets and extensive tuning. Finally, logistic regression provides probabilistic risk estimates that clinicians can communicate and justify, supports hypothesis testing and reproducibility, and can be recalibrated across hospitals or populations, making it a robust and trustworthy choice for medical research and clinical analytics. Another important reason for the better result obtained by the model is that during data pre-processing all the irrelevant and redundant features are removed from the dataset through their correlation score. Always presence of relevant and non-repeating features helps the model to give accurate result with more stability.

6. Conclusion

In this study a machine learning based method is used for the classification of breast cancer disease. During the experimentation it is found that the proposed Logistic regression method has shown the best performance as compared to the other methods used in the literature. As logistic regression is based on probabilistic approach hence it is working efficiently for medical usecase. Also, it is found that if more than one feature are having high correlation score among each other then dropping of any one feature will not impact the result of the classification task.

References

[1] World Health Organization, "Breast cancer," WHO Fact Sheets: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] Y. Feng, M. Spezia, S. Huang, C. Yuan, Z. Zeng, L. Zhang, X. Ji, W. Liu, B. Huang, W. Luo, B. Liu, Y. Lei, S. Du, A. Vuppalapati, H. H. Luu, R. C. Haydon, T.-C. He, and G. Ren, "Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis," Genes & Diseases, vol. 5, no. 2, pp. 77–106, 2018. doi: 10.1016/j.gendis.2018.05.001.

[3] University of Rochester Medical Center, "Breast Cancer Risk Factors," Health Encyclopedia : https://www.urmc.rochester.edu/encyclopedia/content?contenttypeid=85&contentid=P09506

[4] P. Kathale and S. Thorat, "Breast Cancer Detection and Classification," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.367.

[5] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artif. Intell. Med., vol. 34, no. 2, pp. 113–127, 2005, doi: 10.1016/j.artmed.2004.07.002.

[6] R. K. Tripathy, S. Mahanta, and S. Paul, "Artificial intelligence-based classification of breast cancer using cellular images," RSC Advances, vol. 4, no. 18, pp. 9349–9355, 2014, doi: 10.1039/C3RA47489E.

[7] Esmaeili, M., Ayyoubzadeh, S.M., Ahmadinejad, N. et al. A decision support system for mammography reports interpretation. Health Inf Sci Syst 8, 17 (2020). https://doi.org/10.1007/s13755-020-00109-5

[8] T. Khatun et al., "Performance Analysis of Breast Cancer: A Machine Learning Approach," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 1426-1434, doi: 10.1109/ICIRCA51532.2021.9544879.

[9] M. Pal, S. Parija, and G. Panda, "Prediction of breast cancer using tools of machine learning techniques," Onkologia i Radioterapia, vol. 17, no. 4, pp. 1–6, Apr. 2023.

[10] M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast Cancer Prediction via Machine Learning," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 121-124, doi: 10.1109/ICOEI.2019.8862533.

[11] A. Kaur and S. Gupta, "Unveiling Precision in Breast Cancer Prediction with Random Forest and Decision Trees," 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2024, pp. 1232-1236, doi: 10.1109/ICOSEC61587.2024.10722493.

[12] Breast Cancer Wisconsin (Diagnostic) [Dataset]," UCI Machine Learning Repository,1993- https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[13] H. W. Herwanto, A. N. Handayani, A. P. Wibawa, K. L. Chandrika and K. Arai, "Comparison of Min-Max, Z-Score and Decimal Scaling Normalization for Zoning Feature Extraction on Javanese Character Recognition," *2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, Malang, Indonesia, 2021, pp. 1-3, doi: 10.1109/ICEEIE52663.2021.9616665.

[14] X. Zou, Y. Hu, Z. Tian and K. Shen, "Logistic Regression Model Optimization and Case Analysis," *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Dalian, China, 2019, pp. 135-139, doi: 10.1109/ICCSNT47585.2019.8962457

[15] X. Zhang, Y. Sun and H. Xue, "A Study on the Selection of Performance Evaluation Metrics for Machine Learning Methods," *2024 5th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI)*, Jingdezhen, China, 2024, pp. 448-453, doi: 10.1109/ICCBD-AI65562.2024.00081.

[16] M. H. Ferris *et al*., "Using ROC curves and AUC to evaluate performance of no-reference image fusion metrics," *2015 National Aerospace and Electronics Conference (NAECON)*, Dayton, OH, USA, 2015, pp. 27-34, doi: 10.1109/NAECON.2015.7443034.