www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1033-1037

e-ISSN : 2583-1062
Impact Factor : 7.001

BREAST CANCER DETECTION USING MACHINE LEARNING

Omana Gajbhiye[1], Dr. Rakhi Gupta[2], Nashrah Gowalkar[3]

[1]Dept. of Information & Technology, Kishinchand Chellaram College, Mumbai, India
omanagajbhiye21@gmail.com

[2]Head of the Dept., Dept. of Information & Technology, Kishinchand Chellaram College, Mumbai, India
rakhi.gupta@kccollege.edu.in

[3]Asst Professor, Dept. of Information & Technology, Kishinchand Chellaram College, India
nashrah.gowalkaer@kccollege.e du.in

DOI: https://www.doi.org/10.58257/IJPREMS36854

ABSTRACT

Breast cancer remains one of the most dangerous diseases affecting women worldwide, with approximately 12% of women diagnosed with breast cancer at some point in their lives. Early detection is key to effective treatment and survival, but current diagnosis is not always reliable. The research uses machine learning (ML) technology to improve the accuracy of cancer diagnosis. By training machine learning algorithms on features extracted from cancer patient data, such as malignant and benign cells, the research is designed to create classification models that can distinguish between malignant (cancerous) and benign (non-cancerous). Various supervised learning algorithms are used to improve detection accuracy, including K-Nearest Neighbors (KNN), Random Forest, and Logistic Regression. The accuracy of cancer diagnosis is over 95%. This research helps advance cancer diagnosis and highlights the role of AI and machine learning in diagnosis.

Keywords: Machine learning, Algorithms, Malignant, Benign, Diagnosis, Random forest, supervised learning, technology, logistic regression, worldwide.

1. INTRODUCTION

Breast cancer is one of the most common cancers affecting women worldwide, with approximately 12% of women diagnosed with breast cancer at some point in their lives. Early diagnosis of breast cancer is important to improve treatment outcomes and reduce mortality. However, despite advances in treatment and technology, many cases are still diagnosed at an advanced stage, when the chances of survival are diminished. [1] Breast cancer has become the most common cancer among women in India, with more than 1.6 million new cases reported each year, accounting for 14% of all breast cancers diagnosed in women in India, according to GLOBOCAN 2020 data.

[2] Unfortunately, most cancers in India are diagnosed at an advanced stage due to lack of awareness, delayed access to healthcare, and lack of routine screening days, especially in the districts. [3] Traditional diagnostic methods such as mammography and biopsy, while effective, have limitations, especially in terms of false positives and the need for surgery. These challenges have led researchers to explore the potential of machine learning (ML) technology to improve cancer diagnosis. Contracts. This research focuses on developing a machine learning model for breast cancer classification. The research aims to create a robust classification algorithm that can distinguish malignant from benign cases by leveraging features extracted from patient mobile phone data. This approach has the potential to improve diagnostic accuracy and facilitate early intervention, which is particularly important in resource-limited countries such as India, where timely diagnosis voluntarily improves patient outcomes.

2. LITERATURE REVIEW

Machine learning-based breast cancer diagnosis is an active area of research that has recently gained attraction. In this literature review, I will provide a brief summary of several articles on this topic. Bharati S. conducted a study on breast cancer diagnosis using machine learning: a comparative study.

[4] Their study showed that random forest provides greater accuracy in detecting malignant and benign tumors compared to other algorithms due to its ability to handle large data. The authors concluded that ML- based diagnostic systems can complement traditional methods and reduce diagnosis by up to 10%. Classification of breast cancer cells using vector machines: An Indian case study by Gupta R [5] used SVM based breast cancer classification using data from Indian patients. Their research shows that SVM models can achieve high accuracy when trained on local data, making them suitable for use in the Indian healthcare system. The role of AI and machine learning in breast cancer diagnosis in India by Rao and Verma. [6] Both have explored the use of machine learning to develop low-cost diagnostic tools for cancer diagnosis. Their research suggests that machine learning-based diagnostics can be incorporated into existing medical procedures to help radiologists make more accurate diagnoses. By focusing on locally available data and tailoring the process to the Indian population, these systems can improve early cancer

detection in limited areas. [7] AI in Healthcare: Customizing ML for Breast Cancer Detection in India, by Nair. This includes addressing the unique challenges of cancer screening in India, such as the high cost of diagnostic equipment, the scarcity of radiologists, and ensuring that accurate diagnostic tools are easy to obtain and use. [8]

S. Gc works on extracting features such as flexibility, diversity and compactness. They use SVM classification to analyze the performance. Their results showed a maximum of 95% difference and 86% compactness. According to the obtained results, SVM can be considered as a suitable method for cancer prediction. [9] Nithya uses three classification methods including decision tree, k-nearest neighbor, and naive Bayes for different datasets. The authors also studied the measurement error rate. The purpose of this application is to feature the dataset. Together, these data show that significant progress has been made in breast cancer diagnosis using machine learning. However, more research is still needed to improve the accuracy and performance of these algorithms and to ensure their stability and reliability in clinical settings.

3. METHODOLOGY

The methodology for detecting breast cancer using machine learning typically involves the following steps:

A. Dataset Description

We have obtained Breast Cancer Wisconsin (Diagnostic) Dataset from Kaggle. Here 569 Patient's Data Was used for analysis, each instances have 32 Attributes with Diagnosis and Features. Each instance has a parameter of the cancerous and non-cancerous cells and we will predict the cancer just by the input of features.

The values of features is in Numeric Format. The 'Target' means the patient Who is having Whether 'Benign' or 'Malignant' Cancer state.

Type of Patients

Patient Type	Target
Benign	1
Malignant	0

B. Correlation Matrix (Heat map)

To find correlation between each feature and target we visualize heatmap using the correlation matrix.

Feature relationships

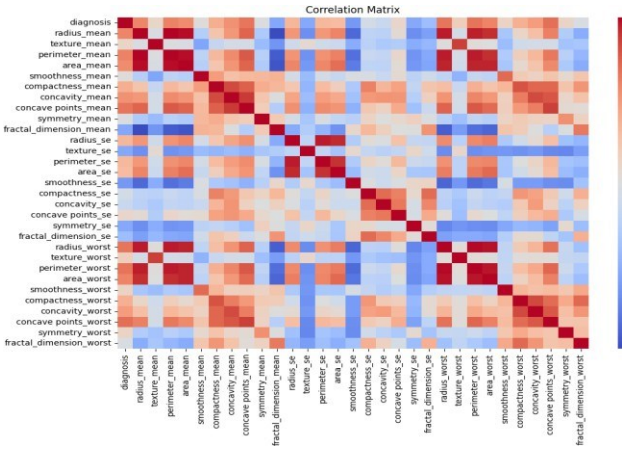
- We can see certain blocks of dark red and blue, which indicate groups of highly correlated features.
- For example, the features related to size measurements (like radius_mean, perimeter_mean, and area_mean) tend to be strongly correlated with each other (red block in the top left)
- Similarly, concavity_mean and concave points_mean show a strong positive correlation, as indicated by the red in that part of the matrix.

Color scale:

- Redder areas (dark red) represent high positive correlations.
- Blue areas (dark blue) represent high negative correlations.
- Lighter colors indicate weaker correlations (closer to 0)

fractal_dimension_mea n	positive correlatio n	indicates a weaker positive correlation between smoothness and fractal dimension features.
Radius_mean & texture_mean	Weak positive correlatio n	A light blue shade shows a weak negative correlation (when on increases the other decreases)

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1033-1037
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001



Qualitative Analysis of correlation matrix

Feature pair	Type of correlatio n	Description
Radius_mean & perimeter_mean	Strong positive correlatio n	The features related to size are highly positively correlated (red color indicates a strong relationship)
Concavity_mean & concave points_mean	Strong positive correlatio n	Indicates a strong relationship between the shapes of tumors (red color indicating both features increase together)
Smoothness_mean &	Weak	Lighter red

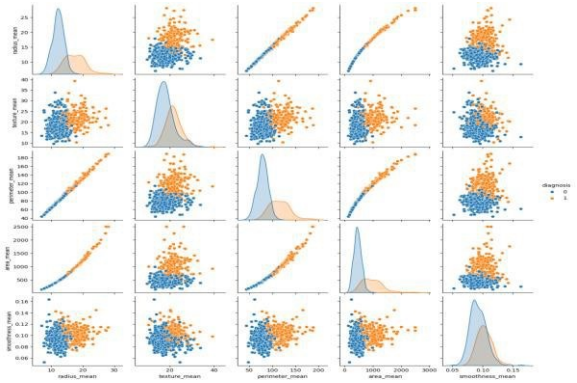
Quantitative Analysis

Feature pair		Correlatio n value	Color indicatio n
Radius_mean	& perimeter_mean	0.9	Dark red (strong positive)
Concavity_mean	& concave points_mean	0.85	Dark red (strong positive)
Smoothness_mean	& fractal_dimension_mea n	0.3	Light red (weak positive)
Radius_mean	& texture_mean	-0.2	Light blue (weak negative)

C. Data visualization

Pair plot of breast cancer data. Basically ,the pair plot is used to show the numeric distribution in the scatter plot.

The pair plot showing malignant and benign tumor data distributed in two classes. It is easy to differentiate in the pair plot.



Pairplot of features [radius_mean, texture_mean, perimeter_mean, area-mean, smoothness_mean]

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1033-1037

e-ISSN : 2583-1062
Impact Factor : 7.001

D. Section headings

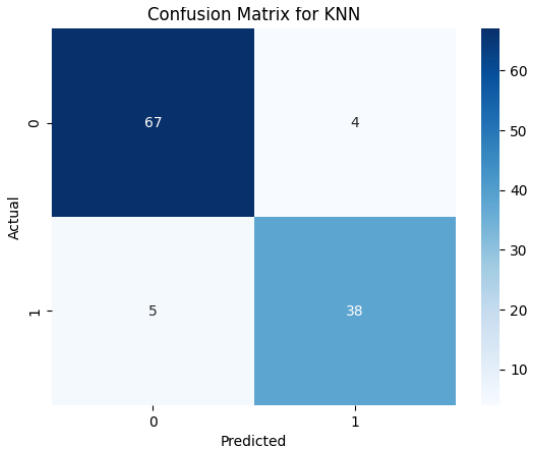
We used Google Colab as a coding platform. Our method includes Supervised learning algorithms and classification techniques like Random Forest, KNN(K-nearest neighbour) and Logistic Regression.

Model selection is the most important step in machine learning. For this research, I'm using supervised learning. We used all methodologies to predict the result and noted their accuracy.

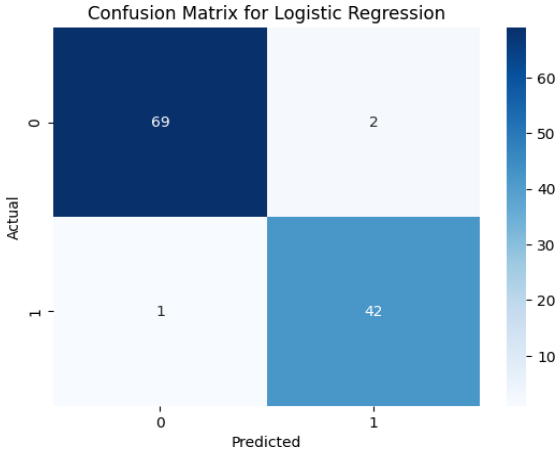
Comparing Accuracy of Models

Models	Accuracy
KNN	0.92
Logistic Regression	0.97
Random Forest	0.96

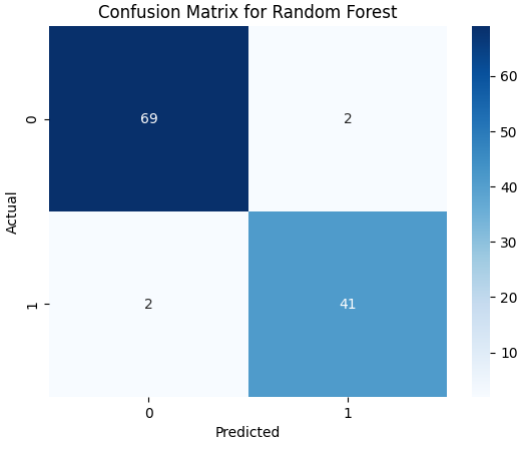
E. Confusion matrix of Models (i)KNN



(ii) Logistic Regression



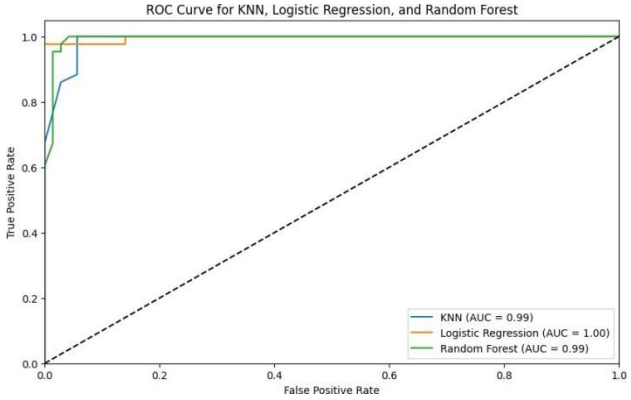
(iii) Random Forest



www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1033-1037
e-ISSN : 2583-1062
Impact Factor : 7.001

Confusion matrix is used for evaluating the performance of a model. The matrix compares the actual target values with predicted values by machine learning model.

F. ROC curve for each model



4. CONCLUSION

This paper examines various machine learning techniques for breast cancer diagnosis. The aim of our study was to analyze Wisconsin breast cancer data by demonstrating and evaluating machine learning predictions. From this study, we found that logistic regression achieved the highest accuracy among algorithms such as logistic regression, random forest classifier, and K-nearest neighbor (KNN), achieving 97% accuracy in breast cancer diagnosis.

However, it is important to preprocess the dataset before running the algorithm. In the future, we aim to evaluate the efficiency and scalability of these algorithms using larger datasets.

5. REFERENCES

[1] Sankaranarayanan, R.,Cancer survival in Africa, Asia, the Caribbean and Central America: Database from the International Agency for Research on Cancer. GLOBOCAN 2020.

[2] Dikshit, R.,(2012).Cancer mortality in India: A nationally representative survey.The Lancet, 379(9828), 1807-1816.

[3] Chaudhary, A., & Pachori, R. B. (2018). Automatic diagnosis of breast cancer using empirical mode decomposition and Hilbert transform. Expert Systems with Applications, 92, 87-102.

[4] Bharati, S., et al. (2020). Breast Cancer Detection Using Machine Learning Techniques: A Comparative Study. International Journal of Medical Informatics.

[5] Gupta, R., et al. (2018). Classification of Breast Cancer Using Support Vector Machines: An Indian Case Study. Journal of Healthcare Informatics.

[6] Rao, P., & Verma, S. (2019). The Role of AI and ML in Low-Cost Breast Cancer Detection in India. Indian Journal of Medical Research.

[7] Nair, K., et al. (2019). AI in Healthcare: Tailoring ML for Breast Cancer Detection in India. Journal of Medical Systems

[8] S. Gc, R. Kasaudhan, T. K. Heo, and H.D. Choi, “Variability Measurement for Breast Cancer Classification Mammographic adaptive and convergent systems (RACS), Prague, Czech Republic, 2015, pp. 177–182.

[9] B. Nithya, V. Ilango, 2017, “Relative Analysis of categorization Methods in R Environment with two Different Datasets.”, Intl J Scientific Research and Computer Science, Engineering and Information Technology (IJSRCSEIT), vol 2, Issue 6, ISSN: 2456- 3307.