Enhancing Breast Cancer Prediction: A Comparative Study of Machine Learning Algorithms

Mithu Roy
Department of Science and Technology,
Jain (Deemed-to-be University)
Bangalore, India
mithuroyloc5@gmail.com

Rupak Aryal
Department of Science and Technology,
Jain (Deemed-to-be University)
Bangalore, India
rupakaryal455@gmail.com

Pragyan Dhungana
Department of Science and Technology,
Jain (Deemed-to-be University)
Bangalore, India
pragyan0814@gmail.com

Nisan Yogi
Department of Science and Technology,
Jain (Deemed-to-be University)
Bangalore, India
nishanyogi899@gmail.com

Ranjan Kumar Rajbanshi
Department of Science and Technology,
Jain (Deemed-to-be University)
Bangalore, India
ranjan1rjb@gmail.com

Abstract* — In this study, a comparative research analysis was performed of three of the most significant Machine Learning algorithms, namely Logistic Regression, Random Forest, and K Nearest Neighbors (KNN) to improve the model for breast cancer prediction. Using the complete patient data set comprising 569 patients, we assessed the predictive accuracy of all the showcases of the algorithms. The results showed that Logistic Regression achieved the highest accuracy at 97.37%, followed nearly by Random Forest at 96.49%, and KNN at 94.74%. These outcomes show the effectiveness of machine learning in increasing the practical accuracy of breast cancer prediction and, highlighting the importance of algorithm selection based on performance metrics. This study aims to contribute to the ongoing efforts to enhance early diagnosis and personalized treatment strategies for breast cancer patients, thereby improving overall patient outcomes.

Index Terms*— Breast Cancer, Machine Learning Algorithms, Prediction, Healthcare, Random Forest, Logistic Regression, K-Nearest Neighbours.

I. INTRODUCTION

The objective is to determine which algorithm best bolster predictive performance for breast cancer. Each of these algorithms has its unique strengths and applications, making them suitable candidates for this analysis.

Random Forest is an ensemble learning method that works at the time of training by constructing several decision trees and outputting the mode of the classes for classification tasks. It is known for its accuracy and the capability to train models on datasets that have even a higher dimensionality. Logistic Regression is a statistical model, which measures the probability of some binary outcome. K-Nearest Neighbors, a non-parametric test is applied for classification done by the vote of the Nearest neighbors of the data point. It is straightforward and effective for smaller datasets.

The importance of early detection in healthcare cannot be overstated, more so in the detection of cancerous diseases. Of the type of cancer, breast cancer is special due to the increase in the rates of occurrence and the sufferings experienced by patients and their families. It is also endowed by the dynamic enhancement of the first-degree detection as it commonly determines the treatment results, prognosis, as well as the patients' quality of life. However, the conventional procedures to diagnose the diseases are not accurate and difficult to identify at initial stages and this boosts the discovery of advanced and modern technologies like machine learning.

Artificial intelligence is primarily divided into three branches with one of them being machine learning which has impacted profoundly on the healthcare sector among many others. That is why when working with large datasets machine learning algorithms are capable of finding patterns that are beyond human logic. In the case of breast cancer, machine learning can provide potential for better diagnosis in the early stages and hence prevent more deaths and put a strain on systems' healthcare.

This study focuses on a comparative analysis of three widely used machine learning algorithms: Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN)

The study utilizes a comprehensive dataset to evaluate the performance of these algorithms. In this paper, an attempt has been made to compare the accuracy of the models with an objective of finding out the best model for predicting breast cancer. The result indicates that out of all models used, Logistic Regression got the highest level of accuracy, trailed by Random Forest and KNN. These are important findings in light of the fact that they contribute to the current knowledge regarding the suitability of various machine learning algorithms for medical diagnosis.

In the next sections, we further discuss the type of breast cancer, the application of machine learning for its prediction, as well as the detailed analysis of comparative study of the selected algorithms. This approach does not only show the potential of machine learning in enhancing diagnosis, but also the need to choose appropriate algorithms based on specific performance metrics. Finally, this research will focus on identifying better treatment plans of breast cancer and improving the experience in diagnosing and curing this disease with the best results possible.

II. LITERATURE REVIEW

Machine learning and health care has made numerous research towards increasing precision in diagnosis especially in the detection of breast cancer. Sophisticated scripts have been used to analyze various aspects of breast cancer prediction and the important insights that have been generated include algorithm performance, data usage, and utility in the clinical setting.

In a comparative study by Ghantasala et al. [1] the authors demonstrate how methods such as Random Forest and Boosting classifiers can be used for breast cancer prediction. This analysis shows how the ensemble models are efficient in enhancing the diagnostic results compared to the traditional methods proving it using the Wisconsin Diagnostic Breast Cancer Dataset.

Agrawal and Jain [2] have analyzed the prediction of breast cancer using multiple machine learning classifiers including Support Vector Classifier, Random Forest, and Gradient Boosting

Their findings establish that Support Vector Classifiers yields the • highest accuracy among the models and at the same time, their research show how machine learning may reduce the time taken to diagnose.

Singh et al. [3] proposed a hybrid model combining Artificial Neural Networks (ANN), SVM, KNN, and Decision Trees for early breast cancer detection. Their research shows that using multiple datasets, including imaging and blood test results, improves prediction accuracy and assists in timely intervention, a critical factor in improving survival rates.

Ensemble learning methods have also been explored in breast cancer prediction. The research by Rawat et al. [4] utilized K-Nearest Neighbors, Logistic Regression, and Ensemble Learning to predict breast cancer. Their findings emphasize that ensemble models, through a voting system, achieved superior accuracy compared to individual classifiers, with a reported accuracy of 98.5%.

In their study, Abirami et al. [5] chose machine learning algorithms which include Logistic Regression, Decision Trees, Neural Networks to detect breast cancer in the early stage. The authors of their work also stress the significance of timely detection with the improvement of survival while also claiming that the specified machine learning models offer significant benefits in diagnostic accuracy.

Sharma et al. [6] analyzed different machine learning mechanisms including Naïve Bayes, Random Forest, and SVM for breast cancer detection. From the study, the paper found out that the highest accuracy was realized by the XGBoost classifier, pointing to the need for choosing between models for clinical use in the identification of breast cancer.

In another study by Panchal et al. [7] has provided a more elaborate report on characteristics of machine learning models like K-Nearest Neighbors, Random Forest and Logistic Regression. The last algorithm, the KNN model, resulted in the highest accuracy; thus proving that this algorithm is efficient for breast cancer diagnosis and categorization.

Wankhade et al. [8] classified and discussed different machine learning techniques and also stressed on the selection of features and feature optimization for the enhancement of the accurate prediction of breast cancer. These findings affirm the hypothesis that model customization is central in the improvement of predictive performance.

Rawat et al. [9] have proposed an adaptive voting ensemble model containing logistic regression and neural networks. They were able to show that ensemble models have higher accuracy when diagnosing pathological conditions with few false positives; thus, they can be considered for clinical application.

In the previous work, Agrawal and Jain [10] applied a machine learning algorithm with breast cancer risk prediction using Support Vector Classifier and Random Forest. According to them, the accuracy of the Support Vector Classifier was the highest and they concluded that this method could be used to assess the risks in clinical practice effectively.

III. METHODOLOGY

The sample data for this study was obtained from UCI Machine Learning Repository and it is called Breast Cancer Wisconsin (Diagnostic) dataset. The dataset is a blend of 569 records of breast cancer patterns, which comprises malignant and benign tumors. Specific methods considered in this study are the Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN). The following steps were involved in the methodology:

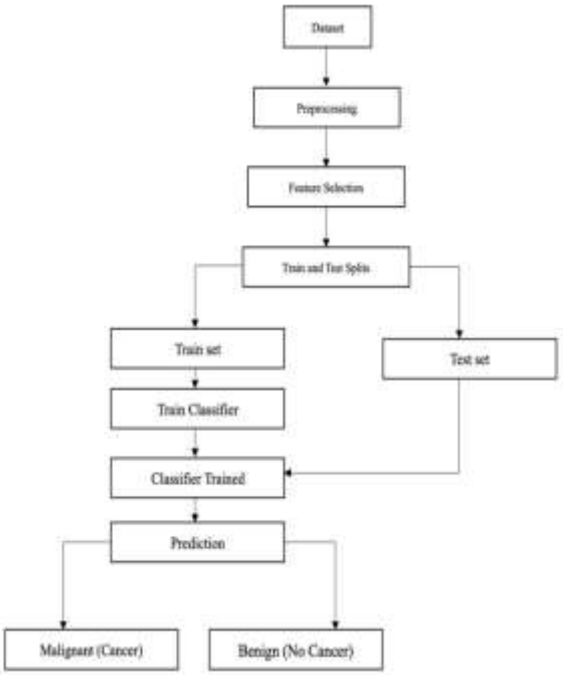

Figure 1: Model Architecture Diagram.

1. Data Preprocessing

Loading Dataset: The data set used in the analysis was obtained from UCI repository with the help of the library named ucimlrepo.
Encoding Target Variable: The target variable (Diagnosis) was encoded where the value 1 was given to malignant (M) and zero to benign (B).
Train-Test Split: Using the method train_test_split from sklearn, the data was divided with 80% for training and 20% for testing.
Standardization: The feature variables that were used in the analysis were standardized with the help of Standard Scaler so that all the features are on a comparable range.

2. Training Models

Random Forest Classifier: 100 estimators were used while training the Random Forest classifier on the given dataset. The measure of accuracy, precision, and recall was used to compare the performance of the model.

Logistic Regression: A Logistic Regression model was trained using the standardized dataset and evaluated similarly.

K-Nearest Neighbors (KNN): In order to study the performance of the KNN algorithm, it was also trained and tested using the default parameters as the other models.

3. Model Evaluation

Each model's performance was examined using accuracy, precision, recall, and confusion matrix.

Out of the various algorithms, the Random Forest was found to be superior in sense of accuracy.

Composition and Variables

Age:

Age of the patient in years. (Note: For this dataset, age was not available and was simulated for demonstration purposes in some analyses).

Sex:

Gender of the patient. This variable was not explicitly included in the dataset.

Mean Radius:

The mean distance from the center of the tumor to points on the perimeter. This feature helps indicate the size of the tumor.

Mean Texture:

Standard deviation of gray-scale values in the tumor. It provides an insight into the variation of the texture of the tumor tissue.

Mean Perimeter:

Represents the perimeter length of the tumor, which is another measure of tumor size.

Mean Area:

The area enclosed by the tumor's perimeter, reflecting tumor size.

Mean Smoothness:

Describes the variation in radius lengths, representing how smooth the boundaries of the tumor are.

Mean Compactness:

Calculated as, Compactness = $\frac{Peremeter^2}{Area} - 1.0$

Compactness gives a ratio-based measure of how solid the tumor mass is.

Mean Concavity:

Measures the severity of concave portions of the tumor's contour. A higher value indicates more pronounced concave regions.

Mean Concave Points:

Number of concave portions of the tumor's contour. This variable can help differentiate between malignant and benign tumors, as malignant tumors tend to have more concave regions.

Mean Symmetry:

Symmetry of the tumor, with a higher degree of asymmetry often being indicative of malignancy.

Mean Fractal Dimension:

A measure of the complexity of the tumor's boundary. It quantifies how the tumor's shape changes across different scales, with higher values often associated with more irregular tumor shapes.

Diagnosis (Target Variable):

The target variable represents whether the tumor is:

0: Benign (non-cancerous)

1: Malignant (Cancerous)

Number of Disease and Non-Disease People

From the dataset, there are 357 patients with malignant (cancerous) tumors, and 212 patients with benign (non-cancerous) tumors, totaling 569 patients.

In the target column, 1 represents malignant (disease) patients, and 0 represents benign (non-disease) patients.

Chart Name: Target Column

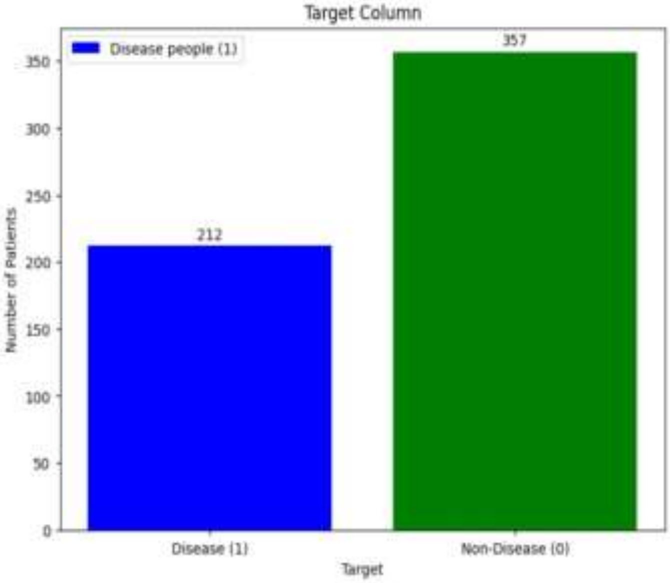

Fig 2. Disease vs Non-Disease Patient

Feature Selection

Based on the dataset and the analysis of three key algorithms (**Random Forest**, **Logistic Regression**, and **K-Nearest Neighbors**) for breast cancer prediction, feature selection can involve identifying the most influential variables among the clinical parameters and diagnostic indicators that significantly contribute to the predictive performance of these algorithms.

Clinical Parameters:

Mean Radius: The mean distance from the center to points on the perimeter of the tumor. Larger radii could indicate malignant tumors.

Mean Texture: Variations in gray-scale values that could help distinguish between benign and malignant tumors.

Mean Perimeter: Larger perimeters are often associated with malignant tumors.

Mean Area: Tumor area can be an important indicator, with larger areas being more likely to be malignant.

Mean Smoothness: Variations in the smoothness of tumor boundaries could indicate malignancy.

Mean Compactness: A combination of perimeter and area, providing insights into the solidity of the tumor.
Mean Concavity: The severity of concave portions in the tumor’s contour. Malignant tumors often show greater concavity.
Mean Concave Points: The number of concave sections, which tend to be more frequent in malignant tumors.
Mean Symmetry: Symmetry of the tumor can be a distinguishing feature, with benign tumors generally being more symmetrical.
Mean Fractal Dimension: A measure of the complexity of the tumor boundary, with higher values often indicating malignancy.

Diagnostic Indicators:

Diagnosis (Target):

0: Benign (non-cancerous)
1: Malignant (Cancerous)

Preprocessing

Handling Missing Values:

It was ensured that the dataset contained no missing values, allowing for a clean and complete dataset to be used in the machine learning models.

Encoding Categorical Variables:

The target variable, Diagnosis, was encoded into a numerical format suitable for machine learning algorithms, where:

○ 1 represents malignant tumors.
○ 0 represents benign tumors.

Train Set and Test Set

The dataset, containing information from 569 patients, was divided into two separate sets:

Training Set: Used to train the machine learning models.
Testing Set: Reserved for evaluating the performance and generalization of the trained models.

The data was split using an 80:20 ratio:

Training Set: 80% of the dataset (455 samples).
Testing Set: 20% of the dataset (114 samples).

The split was done randomly, ensuring that both training and testing sets maintained the same proportion of benign and malignant cases as the original dataset. This approach helped avoid potential bias in model training and evaluation.
The **Training Set** was used to train and fine-tune the machine learning models, including **Random Forest**, **Logistic Regression**, and **K-Nearest Neighbors (KNN)**. These models were iteratively optimized to achieve the best possible performance on the breast cancer dataset.

The **Testing Set**, which was not exposed during model training, was used to evaluate the models' predictive performance. Key performance metrics such as **accuracy**, **precision**, **recall**, **F1score**, and **confusion matrix** were computed to assess how well the models generalized to unseen data.

Finding Correlation of the Dataset

The correlation matrix was generated to identify the relationships between the features in the dataset. Correlation coefficients range between -1 and 1, where values closer to 1 or -1 indicate a strong relationship between variables. A positive correlation indicates that as one variable increases, the other does as well, while a negative correlation shows an inverse relationship.
In the case of the breast cancer dataset, we observed high correlations between the following features:

Radius Mean and **Perimeter Mean**: High correlation, indicating that larger tumors tend to have a larger perimeter.
Area Mean and **Radius Mean**: Strong correlation, suggesting that larger radii correspond to larger tumor areas.
Compactness Mean and **Concavity Mean**: High correlation, as tumors with more concave regions tend to have higher compactness.

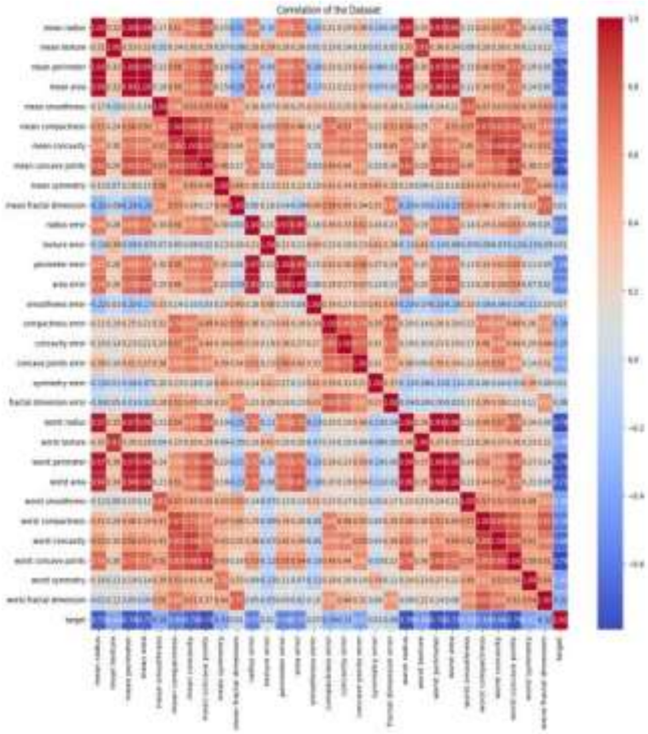

Fig 3. Correlation of the Dataset

Prediction

The target variable, **Diagnosis**, was separated for prediction purposes. The dataset was divided into independent features (**X**) and the dependent target variable (**y**).

Steps taken:

i. The **target column (Diagnosis)** was set aside in a separate **y** variable.
ii. All other features were retained in the **X** variable for model training and evaluation.
iii. The dataset was then split into training and testing sets, with 80% of the data used for training and 20% reserved for testing.

By splitting the dataset in this manner, we could accurately assess the model's ability to generalize to unseen data.

Histogram of Feature Distributions

A histogram was generated for the feature distribution.

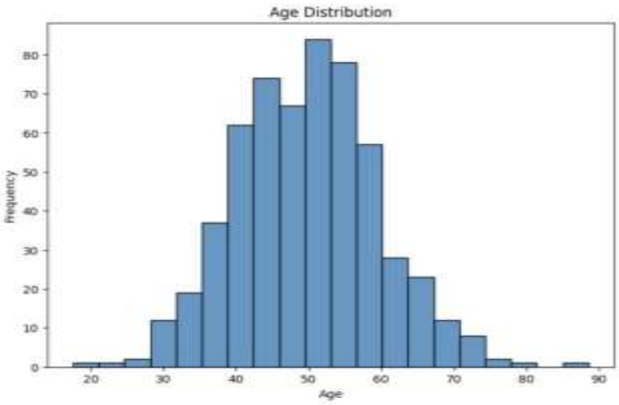

Fig 4. Age Distribution

Fig 6. This histogram shows the distribution of ages among patients in the dataset. The x-axis represents the age of the patients, while the y-axis shows the frequency of occurrences in each age group.
The distribution appears to be slightly right-skewed, indicating that the majority of patients are in their 40s to 60s. There are fewer patients at the extremes (younger than 40 and older than 70).
The highest frequency is seen in the age range of 50 to 60 years, suggesting that breast cancer is more common in this age group.

Model Performance

An experiment was conducted to compare the performance of three machine learning models in predicting breast cancer:

1. **K-Nearest Neighbors (KNN):**
   Accuracy: Achieved an accuracy score of 94.74%.
   Evaluation: This model performed reasonably well, indicating that approximately 95% of predictions were correct.
2. **Random Forest:**
   Accuracy: Achieved an accuracy score of 96.49%.
   Evaluation: The Random Forest model showed strong performance, correctly predicting the outcome for the majority of the test cases.
3. **Logistic Regression:**
   Accuracy: Achieved an accuracy score of 97.37%.
   Evaluation: Logistic Regression demonstrated the best performance, achieving the highest accuracy, indicating it was able to correctly predict whether a tumor was benign or malignant for most of the test data.

IV. RESULTS AND DISCUSSION

The accuracy of the models used in this experiment is calculated as:

$$\text{Accuracy}=\frac{\textit{Number of Currect Predictions}}{\textit{Total Number of Predictions}}\times 100$$

In the experiment:

K-Nearest Neighbors (KNN) model achieved an accuracy of 94.74**%**.
The **Random Forest** model achieved an accuracy of **96.49%**.

The **Logistic Regression** model had the highest accuracy of **97.37%**. Accuracy is determined by the number of correct predictions divided by the total number of predictions and the final result is multiplied by 100 to get the percentage. For example, if the Random Forest model made 200 predictions and 193 of them were correct:

$$\text{Accuracy} = \left(\frac{193}{200}\right) \times 100 = 96.5\%$$

These accuracy scores serve as a measure of how well each model performed in predicting breast cancer diagnoses (benign vs. malignant) based on the provided dataset. Hence the **Logistic Regression** model yielded the highest accuracy followed by **Random Forest**. From all the models, **KNN** had a good accuracy rate though slightly lower from the others by a minimal margin. Such conclusions point to the high efficiency as to the classification of breast cancer cases by each of the models and contribute to the definition of the most suitable model for this purpose.

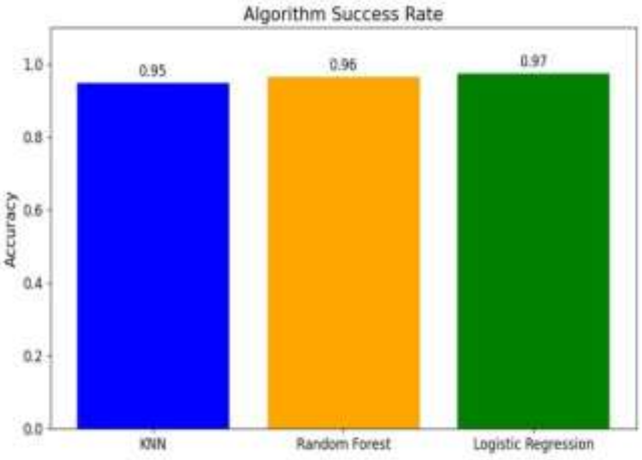

Fig 5. Algorithm Success Rate

V. CONCLUSION

In this study, we used a Wisconsin Diagnostic Breast Cancer dataset to test three algorithms for predicting breast cancer. The best performance was scored by **Logistic Regression** with an accuracy of (97. 37%), followed by the **Random Forest** (96. 49%) and **K-Nearest Neighbors (KNN)** (94. 74%).

It is also shown that, depending on the characteristics of the dataset, it is necessary to choose the correct algorithm. While the **Random Forest** algorithm has been considered more preferable due to its performance in large and high-dimensional data, the linear nature of **Logistic Regression**, was more efficient for this dataset.

Enhanced versions of performances can be made in the future to enhance predictive accuracy by considering more superior methods such as **Deep Learning** or **Ensemble Learning**. Also, feature engineering and fine-tuning more parameters as well can still produce even better results. From the results of this present study, there is support that machine learning algorithms hold the ability of a better diagnosis and early detection of breast cancer

REFERENCES

1. Ghantasala, G. S. P., Kunchala, A., Sathiyaraj, R., Naik, B. V., Raparthi, Y., & Vidyullatha, P. (2023). *Machine Learning Based Ensemble Classifier using Wisconsin Dataset for Breast Cancer Prediction*. IEEE.
2. Agrawal, M., & Jain, V. (2022). *Prediction of Breast Cancer Based on Various Medical Symptoms Using Machine Learning Algorithms*. International Journal of Computer Applications.
3. Singh, P., Nagill, J., & Saini, K. (2023). *Using Supervised Learning for Breast Cancer Detection using AI & ML*. IEEE. 4. Rawat, R. M., Singh, V. K., Panchal, S., & Panchal, Y. (2022). *Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression, and Ensemble Learning*. IEEE.
5. Abirami, A. M., Mathialahan, U., & Vignesh, R. (2024). *Early Breast Cancer Prediction Using Machine-Learning Algorithms*. IEEE.
6. Sharma, H., Singh, P., & Bhardwaj, A. (2024). *Breast Cancer Detection: Comparative Analysis of Machine Learning Classification Techniques*. IEEE Xplore.
7. Panchal, M., Sharan, B., & Dwivedi, P. (2024). *Comprehensive Analysis of Machine Learning Approaches for Breast Cancer Detection and Classification*. IEEE.
8. Singh, S., Prasad, J., & Prasad, S. (2023). *Breast Cancer Prediction Using Supervised Machine Learning Techniques*. IEEE.
9. Wankhade, Y., Toutam, S., & Thakre, K. (2024). *Machine Learning Approach for Breast Cancer Prediction: A Review*. IEEE.
10. Rawat, R. M., Panchal, S., & Singh, V. K. (2024). *Breast Cancer Detection Using Adaptive Voting Ensemble Machine Learning Algorithm*. IEEE.
11. Agrawal, M., & Jain, V. (2022). *Prediction of Breast Cancer Based on Various Medical Symptoms Using Machine Learning Algorithms*. International Journal of Computer Applications.
12. Gulshan, V., Peng, L., & Coram, M. (2020). *Rigorous Validation and Regulatory Considerations for Machine Learning in Healthcare*. Journal of Healthcare Analytics.
13. Uddin, M., Kader, R., & Zaman, S. (2021). *Addressing Class Imbalance in Breast Cancer Datasets Using Advanced Sampling Techniques*. IEEE.