**Abstract:**

Breast cancer starts when some cells in the breast grow uncontrollably, forming a lump or mass called a tumor. There are two main types of breast tumors: benign (non-cancerous) and malignant (cancerous). Malignant tumors are dangerous because they can spread to other parts of the body through the blood or lymph system. This spreading process is called metastasis.

Benign tumors usually stay in one place and don’t spread, while malignant tumors can grow and move to other areas, making them more serious . Breast cancer is the second leading cause of death in women, with over 500,000 cases diagnosed each year.[4]

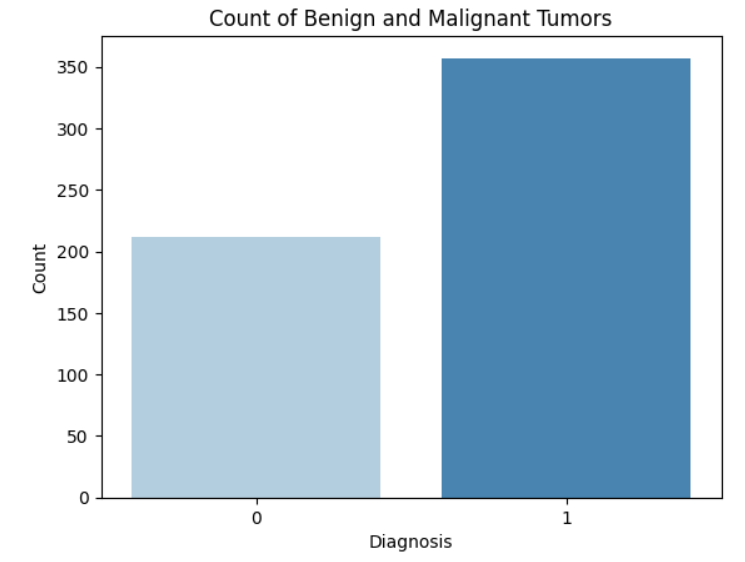
**Keywords:** Breast Cancer Prediction , Machine Learning , Logistic Regression , Random Forest Classification

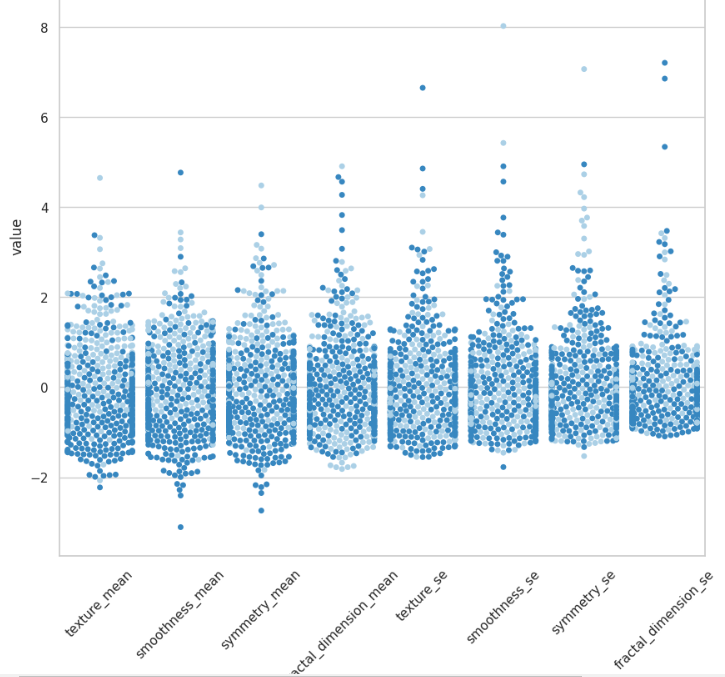
**Introduction:** Breast cancer is one of the most prevalent cancers in women globally and affects the general public's health significantly. Improving patient outcomes and survival rates requires early detection of breast cancer and precise breast cancer prediction. The diagnosis of breast cancer with traditional methods frequently depends on human interpretation, which can be error-prone and time-consuming.

Logistic Regression models have been extensively investigated for breast cancer prediction, demonstrating their potential in accurately classifying benign and malignant. Random Forest is a machine learning method that creates multiple decision trees and combines their results to make predictions. It is great at handling complicated patterns and interactions between features. This method is highly accurate and reliable for predicting whether breast cancer is present, making it very useful for breast cancer classification.[4]

**Methodology:**

**1)dataset :** The Wisconsin breast cancer diagnostic dataset , It includes a total of 569 instances , consisting of 357 benign and 212 malignant cases , Each case is represented by real\_valued features including radius, perimeter, area, compactness, concavity, and concavetexture\_mean,smoothness\_mean,symmetry\_mean,fractal\_dimension\_mean,texture\_se,symmetry\_se,fractal\_dimension\_se,symmetry\_se mean values of cell smoothness\_mean,symmetry\_mean,fractal\_dimension\_mean,texture\_se,symmetry\_se,fractal\_dimension\_se,symmetry\_se points have been identified as informative features for the classification of breast cancer. Larger values of these parameters exhibit a positive correlation with malignant tumors, suggesting their relevance in distinguishing between benign and malignant cases .





**2)Machine Learning Model:**

This research focuses on the application of machine learning algorithms , including Logistic Regression Classification , Decision Tree Classification , Random Forest Classification , KNN , Naive Bayes for breast cancer prediction then identify the top predictors that contribute significantly to the accurate prediction of breast cancer.

The identification of key predictors contributing significantly to breast cancer prediction will aid in the development of more effective and personalized treatment strategies.

Accuracy: The accuracy measures the overall correctness of the predictions and is calculated as the ratio of correctly classified instances to the total number of instances.

Logistic Regression: Accuracy on training data = 0.7296703296703296

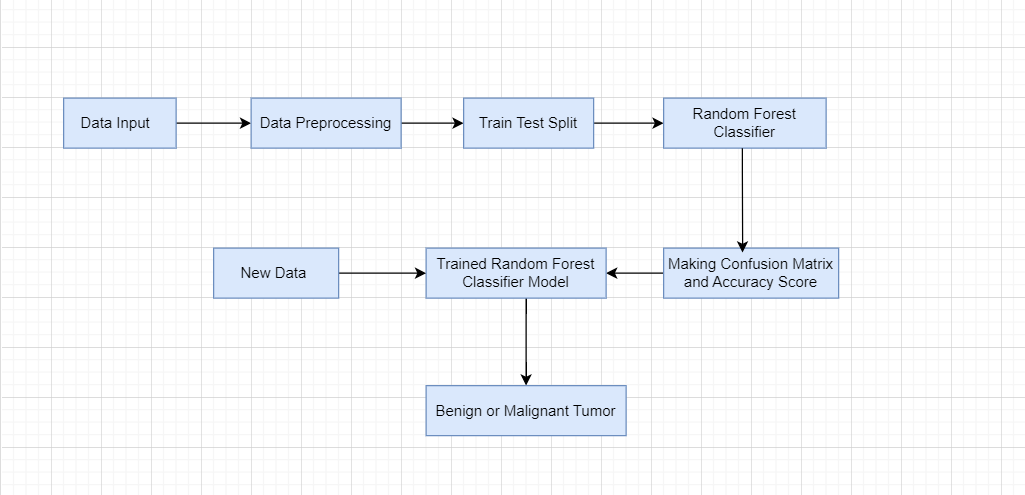
Accuracy on test data = 0.7105263157894737

Random Forest Classification: Accuracy on training data = 1.0

Accuracy on test data = 0.8508771929824561

Random Forest Classifier model achieved an accuracy of approximately 85% on the test dataset, indicating its ability to classify breast masses accurately.

**Proposed Model Work Flow :**



**Existing System**

The existing system for multiple disease prediction utilizes various machine learning algorithms to predict diseases such as diabetes, heart disease, and Parkinson's disease. These algorithms include Naive Bayes, Decision Trees, Random Forest, Support Vector Machines (SVM), and Logistic Regression, each capitalizing on different data features to make accurate predictions. The system is user-friendly, thanks to the Streamlite library, which facilitates model deployment for healthcare professionals. Among these models, SVM has achieved significant success with 76% accuracy in diabetes prediction and 71% in Parkinson’s disease prediction, highlighting its effectiveness in medical diagnostics. Other algorithms, such as Logistic Regression and Decision Trees, also contribute by predicting binary outcomes and offering interpretable models. The Random Forest algorithm, through ensemble learning, further enhances accuracy. While the system has demonstrated the potential of machine learning in disease prediction, there is ongoing work to improve its performance by refining the algorithms and improving data representation.[1]

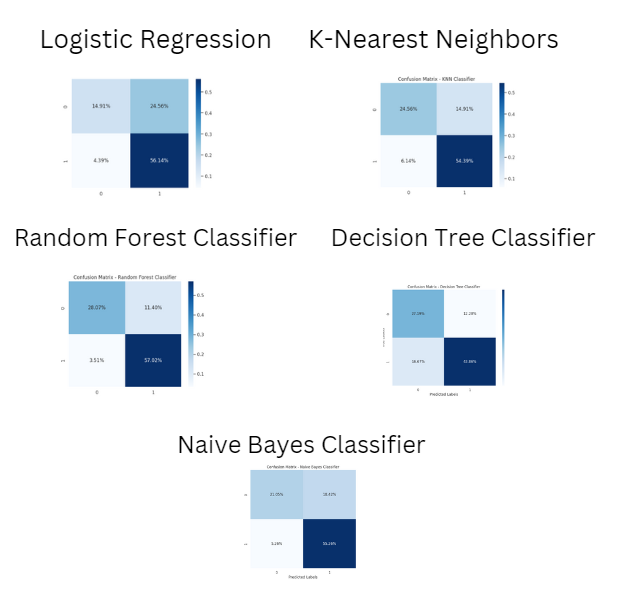
Despite its strengths, the existing system faces several disadvantages. One of the major issues is the potential for inaccurate predictions and incorrect diagnoses due to biased or insufficient trainingdata, which can particularly affect underrepresented communities. Additionally, machine learning models, especially when overtrained on specific datasets, may exhibit poor generalization to new data, reducing their reliability. Another limitation is the black-box nature of many machine learning algorithms, making it difficult to interpret how predictions are made, which is especially problematic in medical applications where transparency is crucial. Furthermore, for rare diseases, there may not be enough data to train the model effectively, limiting its prediction accuracy. The implementation of such systems in healthcare is also expensive and time-consuming, requiring significant investment in hardware, software, and staff training.[2]

**Proposed System**

Our proposed system focuses on predicting two diseases: heart disease and breast cancer, using four machine learning models: Logistic Regression, Gaussian Naive Bayes, Decision Tree Classifier, and Random Forest Classifier from the scikit-learn library. These models were evaluated using key metrics like accuracy, precision, recall, and F1 score. Among them, Logistic Regression achieved the highest accuracy of 71.59%, followed closely by Random Forest at 71.08%. While the Decision Tree showed higher recall at 63.77%, Random Forest had the best overall balance with an F1 score of 63.53%. This comparative approach enables us to identify the strengths and weaknesses of each model, ensuring better prediction accuracy and performance.

Advantages of the Proposed System

Our system is more effective than others due to its higher accuracy and precision (Logistic Regression with 99.86% precision), better model generalization with Random Forest, and superior recall for detecting true positives. It combines interpretability, flexibility, and scalability, making it well-suited for real-world healthcare applications. Additionally, it can be easily extended to other diseases, providing a versatile solution for disease prediction. Another significant advantage of our system is its reduced computational complexity, particularly with Logistic Regression, which performs well with relatively low processing power, making it ideal for integration in resource-limited healthcare settings. The system also offers faster training times for models like Naive Bayes, allowing quicker iterations and updates to the models as new data becomes available. Furthermore, the interpretability of Decision Trees and Random Forest makes the system transparent, allowing medical professionals to understand and trust the predictions. This is crucial in healthcare, where explainability is necessary for clinical decision-making. Finally, our system is robust against overfitting due to the use of ensemble learning in Random Forest, ensuring it can generalize well on unseen data and deliver consistent results across various datasets.

**Result :** ****

**Conclusion :**

In this study, we compared three machine learning methods—Logistic Regression, Decision Tree,Naïve Bayes Classification , K-Nearest Neighbors and Random Forest—for predicting breast cancer using the Wisconsin breast cancer diagnostic dataset. Random Forest was the most accurate, using multiple decision trees to enhance predictions. We identified key features important for breast cancer prediction, such as smoothness\_mean,symmetry\_mean,fractal\_dimension\_mean,texture\_se,symmetry\_se,fractal\_dimension\_se,symmetry\_sswhich can help healthcare professionals improve diagnostic decisions.

**Reference:**

[1]

Parshant and D. Rathee, “Multiple Disease Prediction Using Machine Learning.” Accessed: Oct. 10, 2024. [Online]. Available: <https://www.irejournals.com/formatedpaper/1704650.pdf>

[2]

B. B. S. Kumar, Ashutosh Sharma, Ashish Verma, Ranjeet Maurya, and D. Yusuf, “Machine Learning for the Multiple Disease Prediction System,” *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, vol. 10, no. 3, pp. 673–684, Jun. 2024, doi: https://doi.org/10.32628/cseit24103217.

[3]

Predictive modeling for breast cancer classifcation in the context of Bangladeshi patients by use of machine learning approach with explainable AI Taminul Islam1 , Md.Alif Sheakh2 , Mst. SaziaTahosin2 , Most. Hasna Hena2 , ShopnilAkash3 , YousefA. Bin Jardan4 , Gezahign FentahunWondmie5\*, Hiba‑Allah Nafdi6 & Mohammed Bourhia7

[4]

**Breast Cancer Prediction Based on Machine Learning**

[Yuanzhou Wei](https://www.scirp.org/journal/articles?searchcode=Yuanzhou++Wei&searchfield=authors&page=1)1, [Dan Zhang](https://www.scirp.org/journal/articles?searchcode=Dan++Zhang&searchfield=authors&page=1)2, [Meiyan Gao](https://www.scirp.org/journal/articles?searchcode=Meiyan++Gao&searchfield=authors&page=1)1, [Yuanhao Tian](https://www.scirp.org/journal/articles?searchcode=Yuanhao++Tian&searchfield=authors&page=1)3, [Ya He](https://www.scirp.org/journal/articles?searchcode=Ya++He&searchfield=authors&page=1)4, [Bolin Huang](https://www.scirp.org/journal/articles?searchcode=Bolin++Huang&searchfield=authors&page=1)5, [Changyang Zheng](https://www.scirp.org/journal/articles?searchcode=Changyang++Zheng&searchfield=authors&page=1)6  
1[College of Engineering and Computing, Florida International University, Miami, Florida, USA](https://www.scirp.org/journal/articles?searchcode=College+of+Engineering+and+Computing%2c+Florida+International+University%2c+Miami%2c+Florida%2c+USA&searchfield=affs&page=1).  
2[School of Information Science and Engineering, Shandong University, Jinan, China](https://www.scirp.org/journal/articles?searchcode=School+of+Information+Science+and+Engineering%2c+Shandong+University%2c+Jinan%2c+China&searchfield=affs&page=1).  
3[Steven J. Green School of International & Public Affairs, Florida International University, Miami, Florida, USA](https://www.scirp.org/journal/articles?searchcode=Steven+J.+Green+School+of+International+%26+Public+Affairs%2c+Florida+International+University%2c+Miami%2c+Florida%2c+USA&searchfield=affs&page=1).  
4[School of Economics, Capital University of Economics and Business, Beijing, China](https://www.scirp.org/journal/articles?searchcode=School+of+Economics%2c+Capital+University+of+Economics+and+Business%2c+Beijing%2c+China&searchfield=affs&page=1).  
5[Viterbi School of Engineering, University of Southern California, Los Angeles, California, USA](https://www.scirp.org/journal/articles?searchcode=Viterbi+School+of+Engineering%2c+University+of+Southern+California%2c+Los+Angeles%2c+California%2c+USA&searchfield=affs&page=1).  
6[School of Engineering, Brown University, Providence, Rhode Island, USA](https://www.scirp.org/journal/articles?searchcode=School+of+Engineering%2c+Brown+University%2c+Providence%2c+Rhode+Island%2c+USA&searchfield=affs&page=1).