MACHINE LEARNING PROJECT REPORT

ON

BREAST CANCER PREDICTION

BY

FRIDAY AMARACHI PROMISE

Table of Contents

ABSTRACT

This report focuses on the critical issue of breast cancer prediction using various machine learning algorithms to improve early detection and reduce mortality rates. The study evaluates models like Decision Tree, Logistic Regression, Random Forest Classifier, and SVM to predict breast cancer outcomes with high accuracy. Among these, Logistic Regression achieved the highest accuracy.

Python serves as the primary programming language for implementing these algorithms, with essential libraries such as NumPy, pandas, sklearn, matplotlib, and joblib being utilized. The primary objective is to identify the most effective machine learning model through a comparative analysis, aiming to achieve high accuracy on extensive datasets.

The report emphasizes the urgent need for accurate prognostic models to aid in treatment planning and survivorship strategies for breast cancer patients. It underscores the potential for future research to explore additional predictive parameters and expand the scope of breast cancer analysis.

1.0 INTRODUCTION

Breast cancer is highly prevalent among women, ranking as the second most lethal cancer globally after lung cancer, according to the World Health Organization. In 2018 alone, it caused 670,000 deaths among women, constituting 20% of all female cancer-related deaths. Early detection through methods like breast ultrasound, diagnostic mammogram, MRI, and biopsies is crucial, as timely diagnosis significantly improves recovery chances.

Machine learning algorithms are also being explored for predicting abnormal tumors, aiding in the accurate classification of patients into benign (non-cancerous), premalignant (potentially cancerous), and malignant (cancerous and fast-spreading) groups. This research aims to enhance diagnostic accuracy and treatment planning for breast cancer patients, emphasizing the importance of early intervention to improve outcomes and reduce mortality rates.

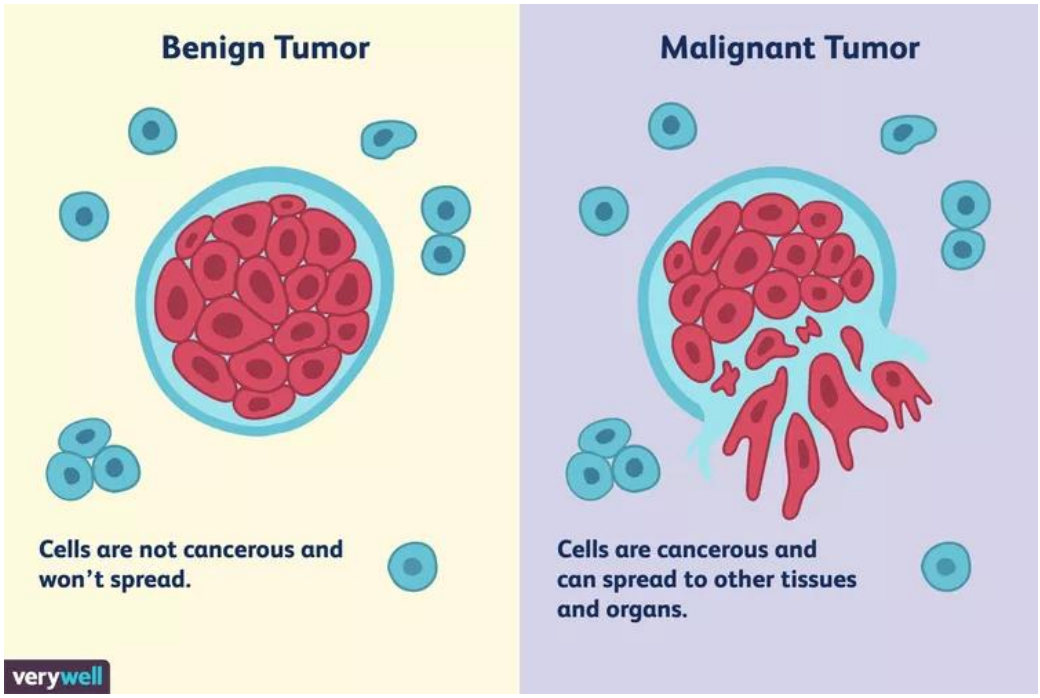


Fig1: A diagram showing the differences between benign and malignant tumors.

2.0 PROBLEM STATEMENT

Breast cancer is a leading cause of cancer-related deaths, with early detection being crucial for effective treatment. Traditional diagnostic methods are often invasive and time-consuming. This project aims to develop a machine learning-based predictive framework to accurately classify breast cancer tumors as benign or malignant. By evaluating various algorithms such as SVM, Random Forest, Decision Tree, KNN, and Logistic Regression, we aim to identify the most effective model. The goal is to create a non-invasive, accurate, and efficient tool for early breast cancer detection, improving patient outcomes and enabling personalized treatment strategies.

3.0 DATA SOURCE & DESCRIPTION

The dataset used to carry out this project was obtained from Wisconsin Diagnostic Breast Cancer (WDBC) available at

http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29

The WDBC dataset contains 569 instances of various features extracted from digitized images of breast mass. The dataset comprises 32 columns, including the ID number, diagnosis (M for malignant and B for benign), and 30 real-valued features computed for each cell nucleus.

Data Features:

- Mean, standard error, and worst (mean of the three largest values) of:
  - Radius
  - Texture
  - Perimeter
  - Area
  - Smoothness
  - Compactness
  - Concavity
  - Concave points
  - Symmetry
  - Fractal dimension

4.0 DATA CLEANING

The dataset was imported into my jupyter notebook and checked for null values. No missing values were found. The ID number column was dropped as it was not useful for building the model.

5.0 DATA EXPLORATION

- The distribution of the diagnosis was visualized using a count plot, showing a higher prevalence of benign cases compared to malignant ones.
- Histograms were plotted for each feature to understand their distributions.

6.0 FEATURE ENGINEERING

The diagnosis column was encoded to numerical values (M = 1, B = 0) using LabelEncoder. Features were standardized using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1.

6.1 FEATURE SELECTION

To enhance model performance, feature selection was performed by analyzing the correlation of each feature with the dependent variable (diagnosis). Features with low correlation were dropped to reduce noise and improve model accuracy. The correlation matrix was computed, and features with a correlation coefficient below a certain threshold were excluded from the final dataset.

7.0 MODEL BUILDING

Several machine learning models were developed and evaluated for breast cancer prediction. These models include:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- Support Vector Classifier (SVC)
- K-Nearest Neighbours (KNN) Classifier

7.1 RESULTS AND DISCUSSION

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.964912	0.953488	0.953488	0.953488	0.962660
Decision Tree	0.938596	0.950000	0.883721	0.915663	0.927776
Random Forest	0.956140	0.952381	0.930233	0.941176	0.951032
Gradient Boosting	0.956140	0.952381	0.930233	0.941176	0.951032
Support Vector Machine	0.956140	0.975000	0.906977	0.939759	0.946446
K-Nearest Neighbour	0.964912	0.975610	0.930233	0.952381	0.958074

Model Performance Evaluation Summary

The performance of various machine learning models for breast cancer prediction was assessed using metrics such as accuracy, precision, recall, F1 score, and ROC AUC. The results highlight the unique strengths and weaknesses of each model.

- **Logistic Regression**

Based on the table above, Logistic Regression achieved the highest accuracy (0.964912) and demonstrated strong performance across all metrics. It had precision, recall, and F1 score all at 0.953488, and an ROC AUC score of 0.962660. This indicates the model's high reliability in distinguishing between benign and malignant cases, maintaining a good balance between correctly identifying positive instances and minimizing false positives.

- **Decision Tree Classifier**

The Decision Tree Classifier had the lowest accuracy (0.938596) among the models. Despite a high precision (0.950000), its recall was lower (0.883721), resulting in an F1 score of 0.915663. The ROC AUC score (0.927776) also lagged behind other models, suggesting that while it can accurately identify malignant cases, it may miss some instances.

- **Random Forest Classifier**

The Random Forest Classifier showed strong performance with an accuracy of 0.956140, precision of 0.952381, recall of 0.930233, and an F1 score of 0.941176. Its ROC AUC score (0.951032) indicates robust discriminative power. This model effectively balances precision and recall, making it a reliable choice for this prediction task.

- **Gradient Boosting Classifier**

Gradient Boosting also achieved an accuracy of 0.956140, with precision and recall both at 0.952381 and 0.930233, respectively. Its F1 score matched that of Random Forest at 0.941176, and its ROC AUC score (0.951032) was identical. This model shows that boosting techniques can enhance performance by reducing errors from previous models.

- **Support Vector Classifier (SVC)**

SVC performed on par with Random Forest and Gradient Boosting in terms of accuracy (0.956140). It had high precision (0.975000) but a slightly lower recall (0.906977), resulting in an F1 score of 0.939759. Its ROC AUC score (0.946446) was slightly lower, indicating that while SVC is excellent at correctly identifying malignant cases, it may produce more false negatives compared to some other models.

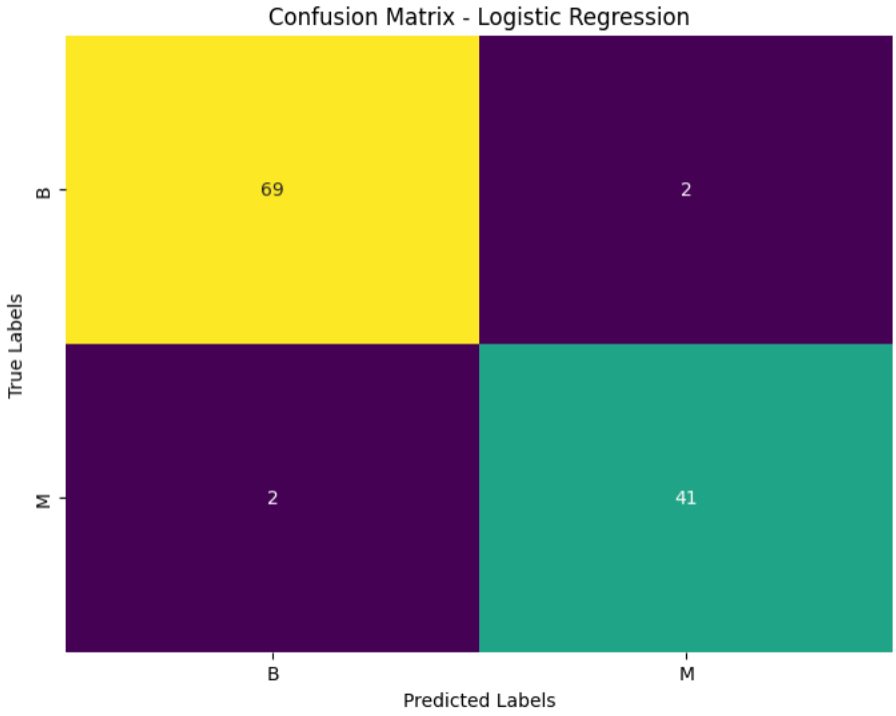
- **K-Nearest Neighbors (KNN) Classifier**

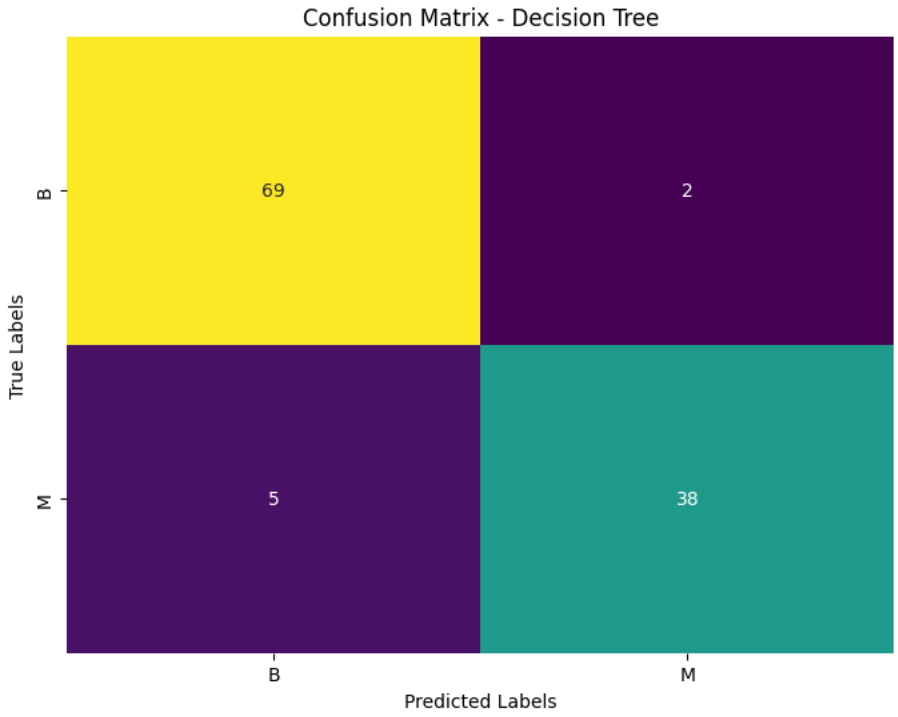
KNN showed high accuracy (0.964912), matching Logistic Regression. It achieved the highest precision (0.975610) and a solid recall (0.930233), leading to an F1 score of 0.952381. The ROC AUC score (0.958074) was slightly lower than that of Logistic Regression, indicating strong overall performance but slightly less effectiveness in distinguishing between classes.

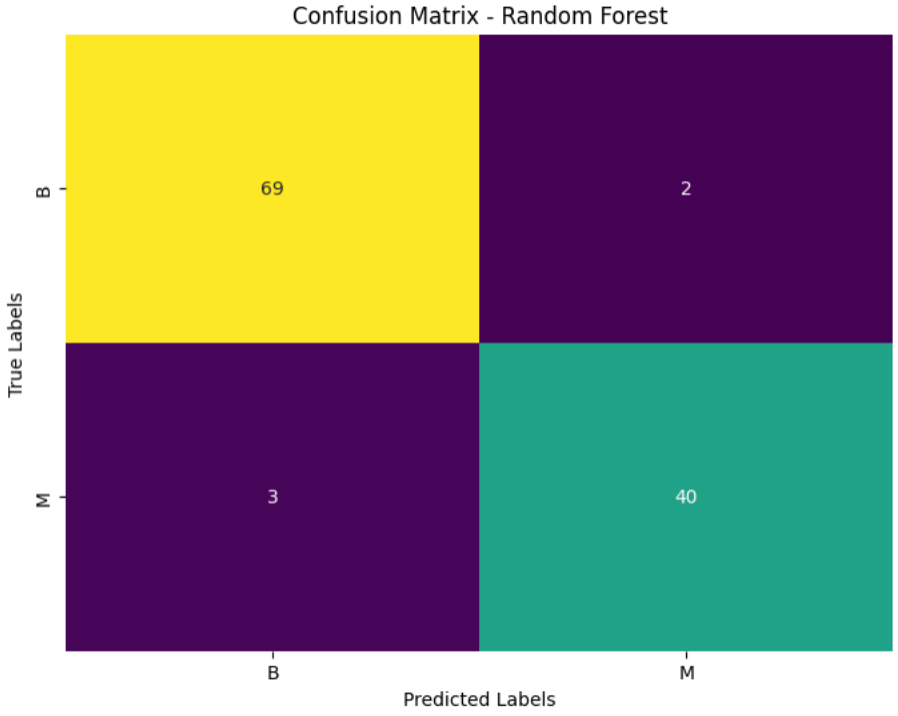
Overall, Logistic Regression emerged as the best model due to its superior balance across all evaluation metrics, particularly excelling in accuracy and ROC AUC. Its simplicity and interpretability also contributed to its suitability for this task. While other models like Random Forest, Gradient Boosting, and SVC also demonstrated strong performance, they did not surpass Logistic Regression in terms of overall effectiveness and efficiency.

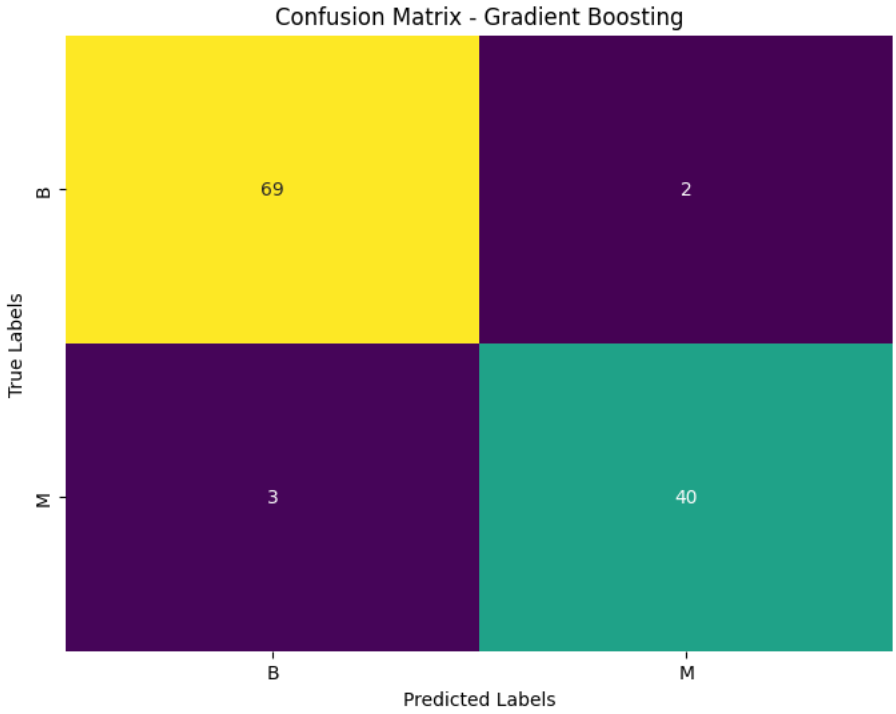
7.2 Confusion Matrix

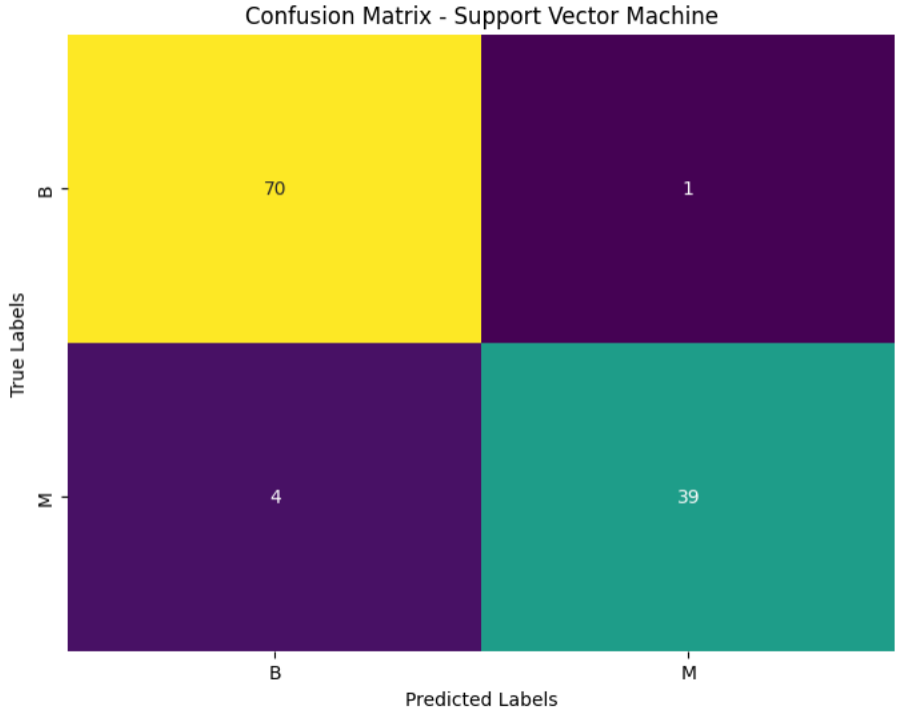
The following confusion matrix displays the instances of correct and incorrect predictions made by each model. In each matrix, the yellow cells represent the cases where individuals have benign conditions and were correctly predicted as benign by the model. The green cells on the diagonal represent the cases where individuals with malignant conditions were correctly predicted as malignant by the model.

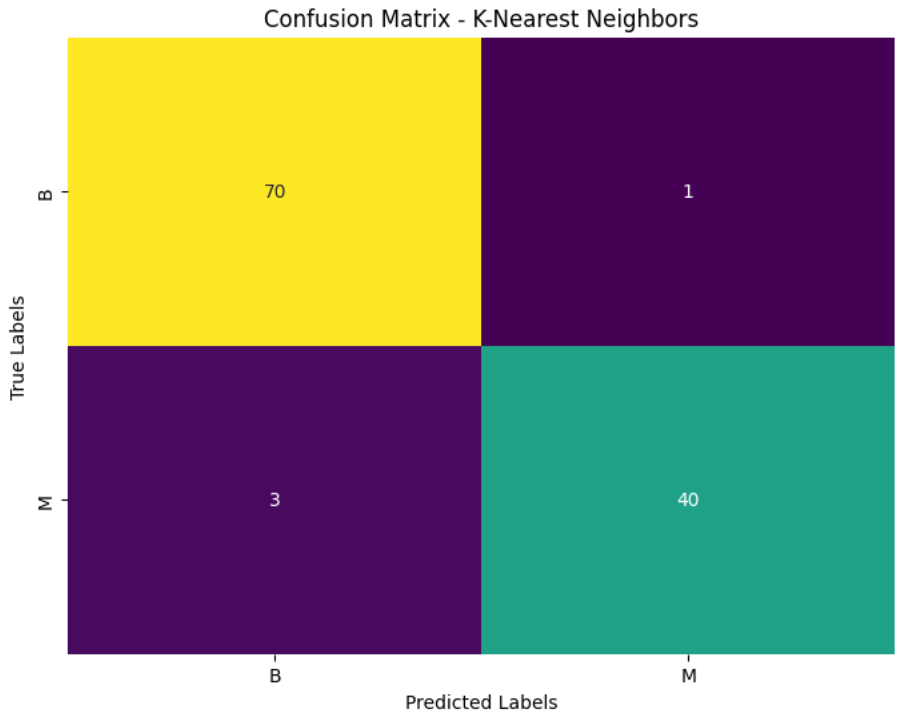



Confusion Matrix - Decision Tree
69
2
5
38
B
M
True Labels
B
M
Predicted Labels


Confusion Matrix - Random Forest
69
2
3
40
B
M
True Labels
B
M
Predicted Labels


Confusion Matrix - Gradient Boosting
69
2
3
40
B
M
True Labels
B
M
Predicted Labels


Confusion Matrix - Support Vector Machine
70
1
4
39
B
M
True Labels
B
M
Predicted Labels



8.0 MODEL DEPLOYMENT

The Logistic Regression model, which demonstrated the highest performance among the evaluated models, was successfully deployed using Streamlit. Streamlit is an open-source app framework that allows for the creation of interactive web applications. This deployment enables users to input relevant features and obtain real-time predictions on whether a breast cancer diagnosis is benign or malignant.

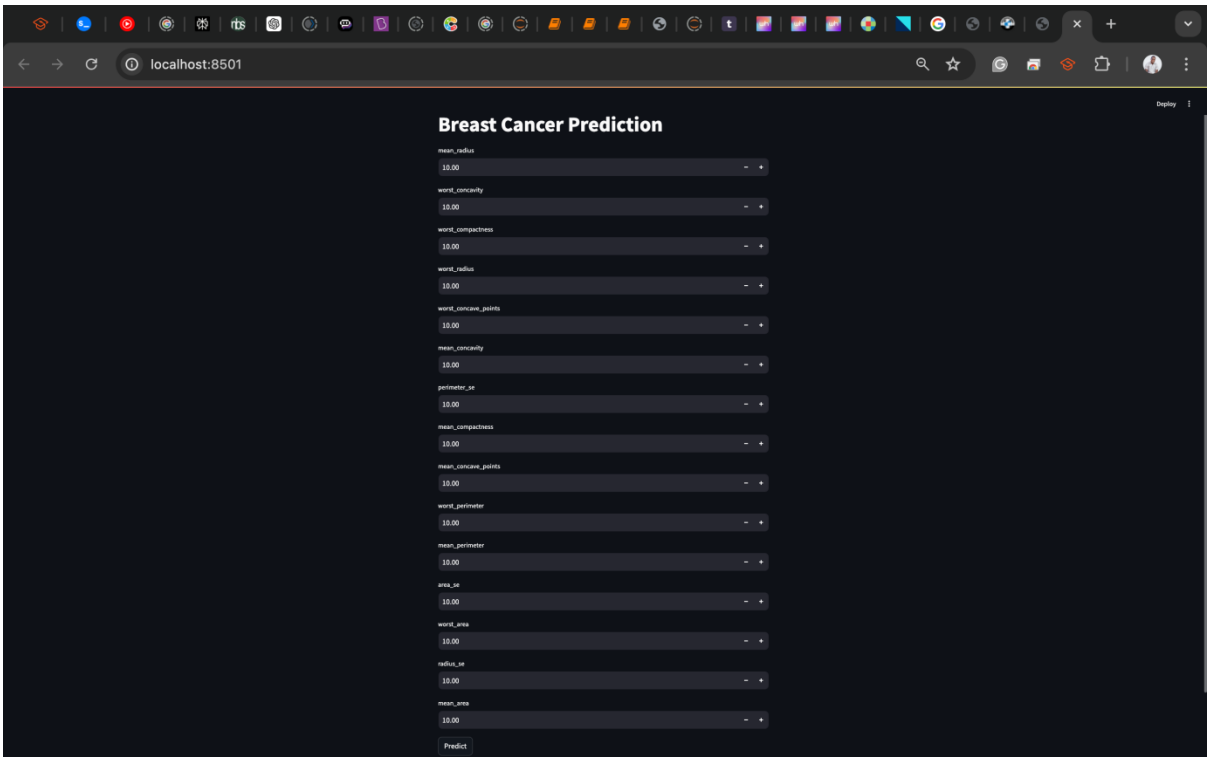


Fig 2: Model deployment

9.0 CONCLUSION

This study utilized various machine learning algorithms to develop predictive models for breast cancer diagnosis, using the Wisconsin Diagnostic Breast Cancer dataset. Through extensive evaluation, Logistic Regression emerged as the optimal model, achieving the highest accuracy of 96.49%. It demonstrated robust performance across all metrics, including precision, recall, F1 score, and ROC AUC, indicating its effectiveness in distinguishing between benign and malignant cases. The deployment of the Logistic Regression model through Streamlit provides a practical tool for real-time breast cancer diagnosis, enhancing early detection efforts and improving patient outcomes. Future research could explore additional features and advanced techniques to further refine predictive accuracy and broaden the applicability of machine learning in oncology.