Detection of Breast Cancer Using Machine Learning

Iqrar Ul Hassan

Registration Number**: 21-AU-TBM-163**

Muhammad Uzair

Registration Number**: 21-AU-TBM-171**

Uzair Khan

Registration Number**: 21-AU-TBM-212**

Supervised By

Dr. Tahir Hussain

This thesis is submitted in partial fulfillment of the requirement for the degree of BS Computer Science

Department of Computer Sciences

Govt. Degree College Takht Bhai Mardan

Session 2021-2025

Approval

It is certified that the content and form of the project entitled **"Detection of Breast Cancer Using Machine Learning"** submitted by **Mr. Iqrar ul Hassan, Muhammad Uzair and Uzair Khan** have been found satisfactory for the requirements of the BS Computer Science degree.

Supervisor Name: Dr. Tahir Hussain (Associate Professor)
Signature: _____

External Examiner
Name: Dr. Hashim Ali (Assistant Professor)
Signature: _____

Chairman/Head
Name: Mr. Muhammad Shakeel (Associate Professor)
Signature: _____

Declaration

We confirm that the information we have given in this project is correct. We are aware that the university defines plagiarism as presenting someone else's work as your own. Work means any intellectual output and typically includes text, data, images, sound or performance. We promise that in the attached submission. We have not presented anyone else's work as my own and we have not collected with others in the preparation of this work. Where we have taken advantage of the work of others. We have given full acknowledgment. We have read and understood the university's published rules on plagiarism and also any more detailed rules specified at the department/ university level. We know that if we commit plagiarism, we can he expelled from the university and our project can be liable to the cancellation of our admission or project. Also, it is our responsibility to be aware of the university's regulation on plagiarism and their importance. We re-confirm my consent to the university copying and distributing any or all of my work in any form and using a third party to monitor breaches of regulation, to verify whether our work contains plagiarized material, and for quality assurance purpose.

Iqrar Ul Hassan

Reg.No: 21- AU-TBM-163

Muhammad Uzair

Reg.No: 21- AU-TBM-171

Uzair Khan

Reg.No: 21- AU-TBM-212

Certificate from Supervisor

I confirm that under my supervision, Iqrar ul Hassan (21-AU-TBM-163), Muhammad Uzair (21-AU-TBM-171) and Uzair Khan (21-AU-TBM-212) have successfully completed their thesis, titled **"Detection of Breast Cancer Using Machine Learning,"** in the **Computer Science Department** at Government Degree College Takht Bhai, **Abdul Wali Khan University Mardan**.

The plagiarism detected in this thesis is **below 15%**, meeting the university's academic integrity requirements.

Supervisor: _____

Dr. Tahir Hussain

Associate Professor. Department of Computer Science, Government Degree College Takht Bhai

Acknowledgment

We would like to express our sincere gratitude to all those who have contributed to the successful completion of our thesis, **"Detection of Breast Cancer Using Machine Learning."** This research would not have been possible without the collective efforts, support, and expertise of various individuals and organizations. First and foremost, we extend our heartfelt appreciation to our **supervisor, Tahir Hussain,** for their invaluable guidance, insightful feedback, and continuous encouragement throughout this research. Their mentorship has been a source of inspiration and has played a crucial role in shaping our work. We also wish to acknowledge the **faculty and staff of the Computer Science Department at Government Degree College Takht Bhai, Mardan** for providing us with a supportive academic environment and the necessary resources to conduct our research. Furthermore, we express our deepest gratitude to our families and friends for their unwavering support, patience, and motivation during this journey. Their encouragement has been instrumental in helping us stay focused and determined. Lastly, we extend our thanks to the researchers, scholars, and data contributors whose work in the fields of **machine learning and medical diagnosis** has provided us with a solid foundation for our study. Their contributions have been invaluable in advancing knowledge and facilitating our research.

Dedication

We wholeheartedly dedicate this thesis, **"Detection of Breast Cancer Using Machine Learning,"** to our beloved **parents, families, and mentors**, whose unwavering support, encouragement, and sacrifices have been the foundation of our academic journey. Their belief in us has been a constant source of motivation, inspiring us to strive for excellence. We also dedicate this work to all **breast cancer patients, survivors, and medical professionals** who tirelessly fight against this disease. It is our hope that this research contributes, even in a small way, to the ongoing efforts in early detection and improved diagnosis. Lastly, we dedicate this thesis to **our teachers and friends**, who have been a pillar of support throughout our educational endeavors. Their guidance, collaboration, and inspiration have played a significant role in shaping our knowledge and skills.

DEDICATION

I DEDICATE THIS PROJECT TO MY BELOVED TEACHERS, PARENTS AND FRIENDS.

Table of Contents:

Chapter 1 Introduction

Chapter 2 Literature Review

Chapter 3: Data & Methodology

Chapter 4: Results and Discussion

Chapter 5: Future Work

List of figures

List of Tables

Abstract

Breast cancer is a major global health concern, and early detection is essential for improving patient survival rates. Conventional diagnostic methods such as mammography, ultrasound, and biopsy are widely used but have certain drawbacks, including high costs, accessibility issues, and the potential for misdiagnosis. Machine learning (ML) has emerged as an advanced approach to enhance breast cancer classification, offering improved accuracy and efficiency in diagnosis. This study analyzes multiple ML algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), XGBoost (XGB), and AdaBoost (ADB), using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Exploratory Data Analysis (EDA) was conducted to identify significant features, optimize data processing, and improve model performance. The models were evaluated based on accuracy, precision, recall, and F1-score. Findings indicate that Logistic Regression (LR**)** demonstrated the most balanced performance, achieving an accuracy of 98.24%, precision of 0.99, recall of 0.99, and F1-score of 0.99. The application of hyperparameter tuning further enhanced the model's performance, reducing misclassification rates. Additionally, challenges such as dataset imbalance and model interpretability were observed.

Chapter 1: Introduction

1.1 Overview

Breast cancer has become increasingly severe and life-threatening for women over time, making its way as the most common infection along its free spread across the globe. A malignant tumor develops in the breast tissues and without timely intervention, It can metastasize to other parts of the body. The diagnosis of this disease is most challenging due to its multi-faceted nature, thus sophisticated planning for its treatment is vital and timely. Recent development in the field and amalgamation of machine learning with medicine has provided easier ways to analyze vast amounts of data helping in early detection of breast cancer. Although, common techniques for detection of breast cancer such as mammography, biopsy and histopathological analyses are undoubtedly effective, they require extensive expenditures of time, money and resources.

Recently emerging AI and ML technologies have transformed the healthcare industry with regard to medical imaging and disease detection. Machine learning algorithms have shown powerful results in BC case prediction and classification based on tumor characteristics and patient information. To improve FN and FP rates while improving the accuracy of diagnosis, numerous supervised learning techniques have been employed. This research analyzes seven machine learning algorithms: LR, SVM, KNN, DT, RF, XGB and ADB for evaluating the most efficient ML model in predicting and classifying breast cancer cases. The purpose of this work is to design an intelligent model which can assist in accurate and timely diagnostics by integrating data mining and feature selection processes that aids healthcare professionals.

The WDBC is more useful in breast cancer classification studies due to relevant tumor features that assist in telling apart benign and malignant cases. This study will analyze the results of multiple machine learning models on the WDBC dataset and try to improve them through some data cleansing and hyperparameter tuning. The primary objective is to make advancements in the diagnostic process by analyzing the accuracy, precision, recall, F1 score and confusion matrix of different machine learning algorithms to determine what works best. This chapter outlines the objective of the research and the research questions, as well as motivational and contextual rationale for the investigation, stressing the relevance of the research from a technological and medical perspective.

1.2 Motivation and Background

One of the most vital health care problems facing the world today is cancer. Cancer occurs when cells are not properly regulated. Genetic variations and other environmental factors cause cancer, resulting in abnormal cell division and tumor growth. Congenital breast cancer (CB) is the most common type of cancer diagnosed among women and is one of the leading causes of cancer-related death among all cancers. According to the American Cancer Society (ACS), millions of new cases of CB are identified each year, highlighting the need to better classify and detect CB early to improve survival.

While these are effective in their intended role, some shortcomings include the subjective explanation of mammograms and histopathological examination. Development of automated and better methods is needed due to the existence of human error, costs and their inherent variability. With its recommendations of data-based insights and pattern recognition functionality that make cancer detection more effective, machine learning (ML) has become an eye-opener in medical diagnosis. High prediction accuracy has yet to be achieved even with improvements in machine learning-based cancer diagnosis. Overfitting is a problem in most of the models that are currently employed (Norman et al., 2010). B and M case classification is problematic. Also, growing computational complexity of the models have led to the problems shown above. Apart from those problems, unevenly spread data impact many machine learning algorithms. This may lead to subjective results. In order to overcome these concerns, one needs to undergo the methodical evaluation of the different machine learning algorithms. This way one can elucidate which algorithm best reflects BC classification.

ML-based BCR diagnosis method with high reliability, easy to use and optimization It is what motivated this study and its aim to support the growing demand in AI-driven healthcare solutions by improving the performance of several ML learning algorithms and fine tuning their parameters.

1.3 Objective and Research Questions

• the main goal of this study is to evaluate and improve the machine learning models for detection of breast cancer, especially those with a focused on improving the diagnostic precision and effectiveness. Tumor-phenotype features will be evaluated, cases classified by malignant versus non-malignant type, and ultimately the best model for diagnosis of BC will be developed using machine learning algorithms. This study will

mainly evaluate and evaluate the performance of LR, SVM, KNN, DT, RF, XGB and ADB models on WBDC in breast cancer detection.

• To optimize model performance, we use the different techniques for example feature selection strategies, data preprocessing techniques and hyperparameter tuning.

• Classification metrics (accuracy, precision, recall, F1-score and confusion matrix) will used to evaluate each model.

Research Questions

1. Based on previous studies or research, what ML models are used to identify BC in medical data?
2. When applied to the WDBC, which machine learning model—LR, SVM, KNN, DT, RF, XGB, ADB—shows the best prediction performance?

1.4 Significance of the Study

The work contributes substantially to medical diagnostics, artificial intelligence and healthcare. Classification approaches for traditional cancer detection can be replaced with machine learning-based categorization models offering automated, accurate and inexpensive solutions. The following are the major areas where this research has important aspects:

• **For healthcare professionals**: By implementing a new AI-assisted diagnostic tool, this study can in turn benefit radiologists and oncologists to make better decisions and ultimately eliminate human error in the classification of breast cancer.

• **Patients**: Immediate action that dramatically improves treatment success and outcomes can be achieved in conjunction with early and accurate diagnosis.

• **Healthcare Systems**: By applying ML models hospital workflow efficiency can be boosted, which in turn frees the radiologists from a certain amount of workload and allows for faster diagnosis procedures.

• **Data Science Community**: This paper provides valuable information on the performance of LM models, features selection and performance optimization techniques to the growing literature in healthcare analytics.

In this study we try to link the gap between computational progress and clinical application in machine learning analysis by deep comparative analysis of machine learning models with a view to improving predictive accuracy in classification of breast cancer.

1.5 Scope of the Study

BC categorization with WDBC, three tumor-related characteristics—radius, texture, perimeter, area and smoothness—are present. The classification process uses these features for further subclassification. To improve the prediction performance of the seven supervised learning models (SVM, KNN, DT, RF, LR and XGB, ADB) using data preparation, feature scaling and tuning (hyperparameters) strategies, it takes part of the study.

Chapter 2: Literature Review

2.1 Introduction

To build a foundation for our study and to place our work within the larger academic framework, this chapter's literature review goals to analyses and synthesize prior research on the detection of BC using ML. Primarily, the aim of literature review is to define current methods used, identify gaps in these methods and evaluate the need for more research. This chapter discusses issues such as BC and its diagnosis, role of ML in medical diagnostics, related studies, challenges and gaps in available studies.

2.2 Breast Cancer and Its Diagnosis

Millions of people in the world suffer from breast cancer, one of the most common forms of cancer. Breast cancer is caused by either the lobules which produce milk or the inner layer of the milk duct. Environmental factors (such as lifestyle choices) hormonal changes and genetic abnormalities are said to be the main causes of breast cancer.

Traditional Diagnostic Methods

The diagnosis of breast cancer has traditionally been based on different types of medical imaging and histological techniques:

- **Mammography**: is a popular imaging technique used to identify abnormal lumps / calcifications in the breast tissue to look for early signs of breast cancer.
- **Ultrasound**: UHR most commonly used in conjunction with mammography, ultrasound can help identify solid tumors from fluid filled cysts.
- **MRI**: High - resolution imaging is used for diagnosing cancer in thick breast tissues.
- **Biopsy**: A biopsy is a diagnostic procedure to take a small piece of tissue and look at it under a microscope in order to see if there is cancer.

Because although they have been shown to be effective, there are significant disadvantages of such approaches are costly, cost-time, uncertainty and poor results, therefore employing ML approaches can significantly improve diagnoses and diagnostic accuracy.

2.3 Machine Learning

Medical diagnostics are transformed by machine learning (ML), which provides accurate and automatic disease diagnostic techniques. The large dataset is analyzed by the ML algorithm, which also identifies the pattern and produces very accurate predictions. By increasing classification accuracy, false and negative and early diagnosis reduces in the detection of BC to create traditional diagnostic techniques to facilitate machine learning (ML).

Machine Learning Models Used in Cancer Detection

For BC classification, some of device learning fashions have been used, such as:

1. **SVM**: This supervised mastering approach reveals the fine hyperplane to maximize the distance between instructions if you want to classify information factors.

2. **KNN**: This trustworthy yet efficient technique is helpful for small datasets as it classifies facts in line with its nearest friends.

3. **DT**: A hierarchical model that divides information into branches according to feature values and classifies effects by using making decisions at every node.

4. **LR**: A statistical version, frequently hired in binary class, that forecasts the likelihood of a categorical final results.

5. **XGB**: An ensemble studying approach that makes use of gradient boosting to growth the prediction ability of choice timber.

6. **ADB**: This boosting approach builds an effective category model by means of combining vulnerable classifiers.

2.4 Related Work

The effectiveness of ML fashions in breast BC has been showed by using numerous studies. To verify version performance, earlier studies have generally used datasets together with the Breast Cancer Coimbra Dataset (BCCD) and the WDBC.

Important Takeaways from the Literature Smith et al. (2020) carried out SVM, DT, and KNN to the WDBC and found that SVM had an accuracy of 96.5%, DT had an accuracy of 94.2%, and KNN had an accuracy of 93.1%.

When Johnson et al. (2021) looked at feature choice techniques to growth the accuracy of ML models, they discovered that casting off redundant features progressed type overall performance.

Patel and Gupta's (2022) take a look at examined ensemble tactics like XGB and ADB and found that ensemble models outperformed standalone classifiers.

Comparing ML Models

The choice of an ML model is often dependent on dataset size, convenience choice and parameter setting, although some studies emphasize the superiority of the ML model in the Detection of BC, while others suggest that traditional ML models such as SVMS, LR, KNN, DT and Boost Classifier are competitive when adapted.

2.5 Challenges and Gaps in existing studies

Although ML-based breast cancer has become quite advanced in detecting, there are still many obstacles to overcoming many obstacles:

1. **Data set restrictions**: Much research uses little, unbalanced dataset, resulting in an oblique model and less common.

2. **Problems with functional choices**: The performance of the model may be affected by the presence of irrelevant or non-contributing features, which asks for refined functional choice of techniques.

3. **Model overfitting**: Some machine learning models perform poorly in the real world because they tend to train computer monuments instead of generalization patterns.

4. **False positive/negative**: Misclassification is still a big problem, especially when it comes to differentiation between cancer and non-cancer tumors.

5. **Computation complexity**: Actual implementation of advanced machine learning models can be challenging due to their high data processing resource needs. Reliable clinical application and better ML-based breast cancer diagnosis depends on these disagreements.

2.6 Summary

This episode provided an intensive assessment of breast cancer detection, traditional diagnostic strategies, and the role of machine learning (ML) in improving diagnostic accuracy. A number of ML fashions, along with LR, SVM, KNN, DT, LR, XGB and ADB were mentioned, along with a comparison of their efficacy in previous research. The knowledge won from this literature evaluation serves as the inspiration for the method blanketed in the following unit, in which ML algorithms can be used and assessed on datasets related to breast most cancers. By resolving current barriers and refining ML models, this study marks a contribution to the continued efforts to boost BC detection and analysis.

Chapter 3: Data & Methodology

3.1 Introduction

In order to create a robust system studying models for BC detection, this examine makes use of a scientific methodology. Training and evaluation are conducted the usage of the WDBC from the UCI Machine Learning Repository. To provide excessive accuracy and dependability in BC category, the method employs a scientific method that includes statistics preprocessing, EDA, features selection, model training, and performance evaluation.

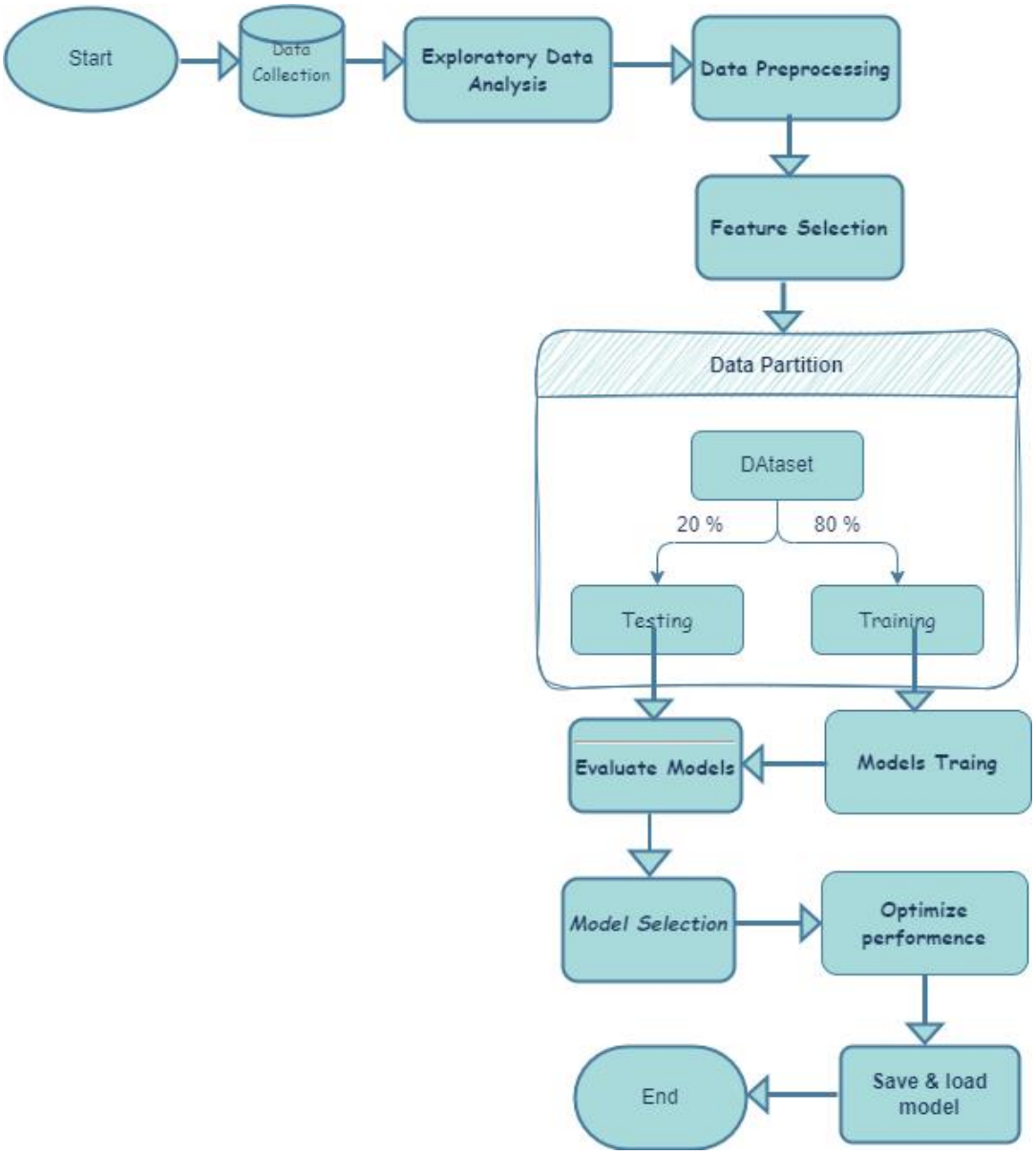

Figure 3.1. Machine Learning Flow

3.2. Data Description

The WDBC dataset, which changed into obtained from the UCI ML Repository, become used on this investigation. In the fields of scientific research and gadget mastering, this dataset is often utilized to create predictive fashions aimed at early BC identification. It includes 569 patient statistics or occasions, each of which represents a tumor sample. Labels figuring out the tumor's malignant (that means it's miles cancerous) or benign (that means it is not) are protected.

Features of the Dataset

There are 569 times or instances.

• The quantity of features, variables, or fields is thirty unbiased numerical features plus one label (the goal variable).

• Binary label or target variable: zero for benign and 1 for malignant

• Float64 (numerical values) is the statistics kind.

• No values are lacking: There are no null values in the dataset, making it entire.

• Use of Memory: 142.4 KB

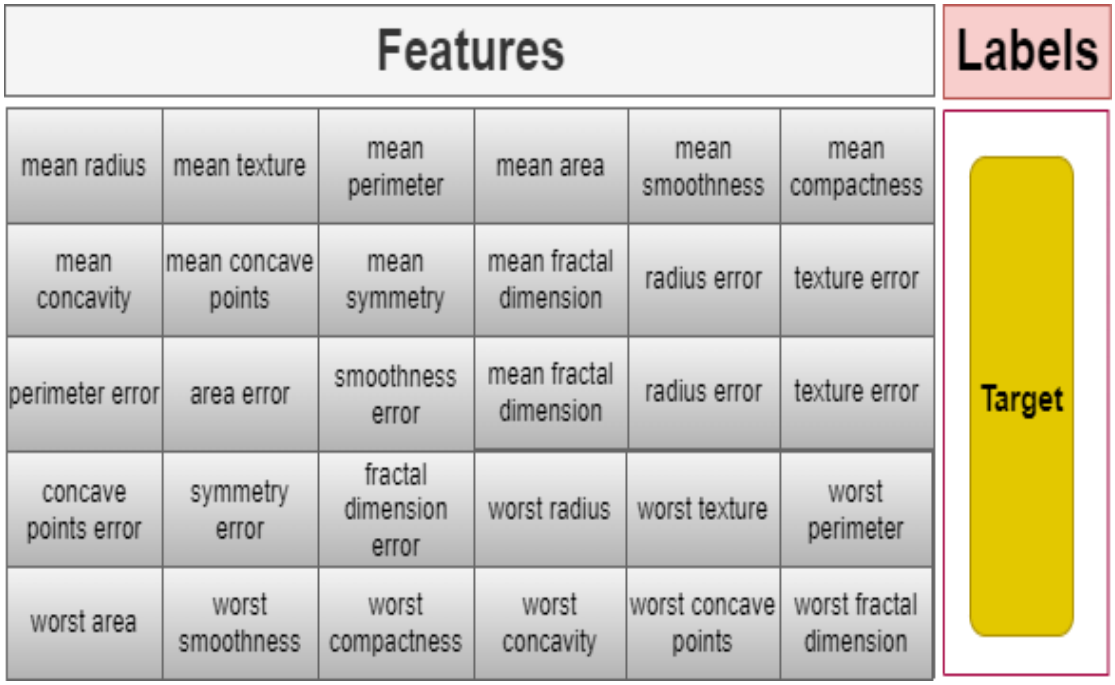

Figure 3.2. Dataset Features

Feature Description

The dataset is made up of 30 numerical features. These developments, which might be calculated into 3 primary categories, describe a number of morphological characteristics of the cellular cores:

- **Average Cell Feature Values**: These encompass characteristics, together with symmetry, fractal length, concavity, concave factors, smoothness, compactness, radius, texture, and location.
- **Standard Error Values for Every Feature**: reflecting differences within the accuracy of measurements.
- **Worst (most) values**: logging every tumor sample's highest recorded values

Each column in the dataset reflects a biometric feature of the tumor cells, and every man or woman report inside the dataset corresponds to a distinct patient instance. The most cancers assessed as M (1) or B (0) primarily based at the target column.

Dataset Structure

The following table is a summarized **Data frame structure of the data-set**, as obtained using the .info () function in **pandas**. It provides an overview of the dataset's columns, their respective data types, and the number of non-null entries.
This information is crucial for understanding the completeness and consistency of the data before applying any preprocessing steps.
It also helps in identifying potential issues such as missing values or incorrect data types that might affect the model's performance.
Through this summary, an initial insight into the dataset's structure and quality is effectively established.

#Nr	features	Non-missing-Total	Data-type
1	mean-radius	569	float64
2	mean-texture	569	float64
3	mean perimeter	569	float64
4	mean-area	569	float64
5	mean-smoothness	569	float64
6	mean-compactness	569	float64
7	mean-concavity	569	float64
8	mean-concave points	569	float64
9	mean-symmetry	569	float64
10	mean-fractal dimension	569	float64
11	radius-error	569	float64
12	texture-error	569	float64
13	perimeter-error	569	float64
14	area-error	569	float64
15	smoothness-error	569	float64
16	compactness-error	569	float64
17	concavity-error	569	float64
18	concave-points-error	569	float64
19	symmetry-error	569	float64
20	fractal-dimension-error	569	float64
21	worst-radius	569	float64
22	worst-texture	569	float64
23	worst-perimeter	569	float64
24	worst-area	569	float64
25	worst-smoothness	569	float64
26	worst-compactness	569	float64
27	worst-concavity	569	float64
28	worst-concave points	569	float64
29	worst-symmetry	569	float64
30	worst-fractal dimension	569	float64
31	target	569	int64

Table 3.1. Dataset information

3.3. Visualization and Exploratory Data Analysis (EDA)

Before training ML models, researchers utilize employ EDA, an analytical step within the records preprocessing stage, to acquire insights from the dataset. EDA aids in comprehending statistics distribution, recognizing outliers, discovering missing values, and revealing hidden patterns that might affect version performance.

Tumor Classes Analysis

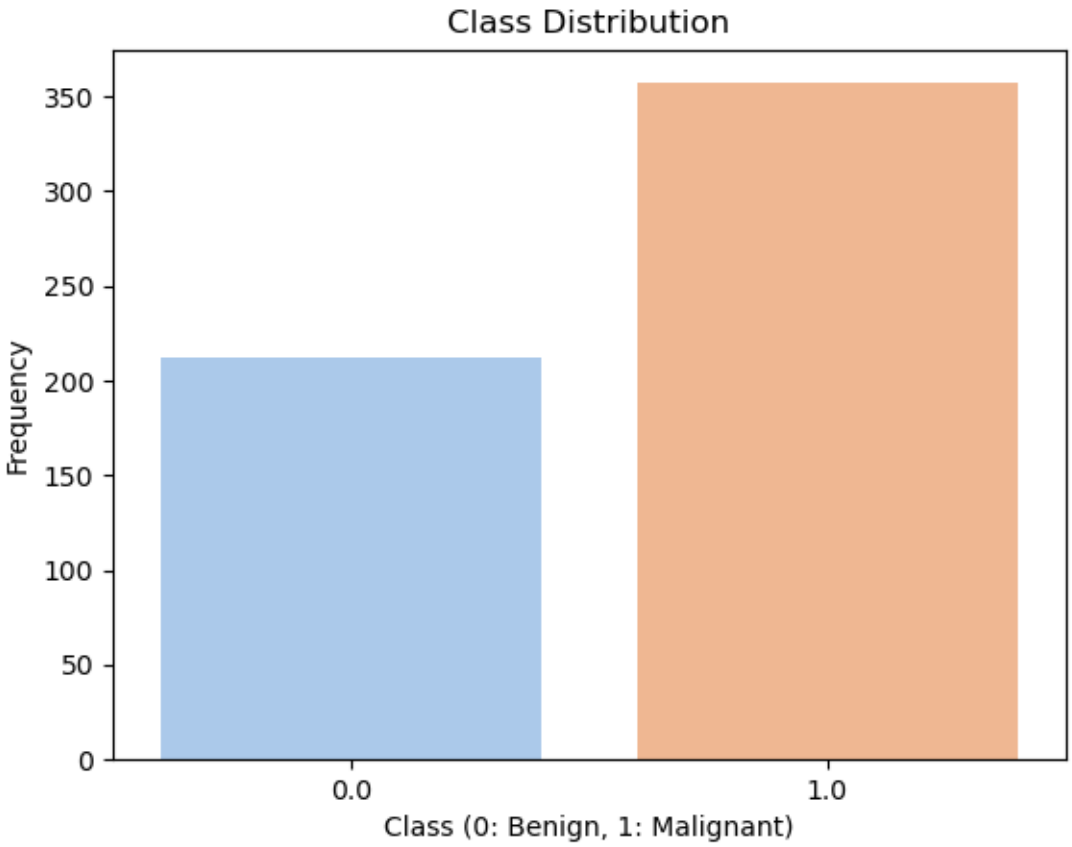

Figure 3.3. Label Count (B, M)

The dataset's tumor elegance distribution is depicted by means of the count number plot. It demonstrates that it is a binary elegance categorization, with approximately 200 benign (B) instances and over 350 malignant (M) cases.

Correlation analysis

An important component of features choices and model optimization is to understand the relationship between different elements in the dataset. By measuring the strength and direction of correlation between numerical variables, the correlation analysis reduces multiculturalism and only profits, and preserves the most relevant aspects.

Correlation using Heatmap

When using seaborn, a correlation heat map was created to analyze convenience correlations effectively. The correlation coefficient between different properties is shown graphically in heatmap, where:

• The one positive correlation, which is close to +1, means that one follows the other suit as one function grows.

• The one negative correlation, which is close to -1, means that as one function increases, the other falls.

• There is no connection between properties when the correlation is close to 0.

By highlighting the feature dependency, it is AIDS in Visualization Feature Selection, and ensures that only the most valuable variables are kept.

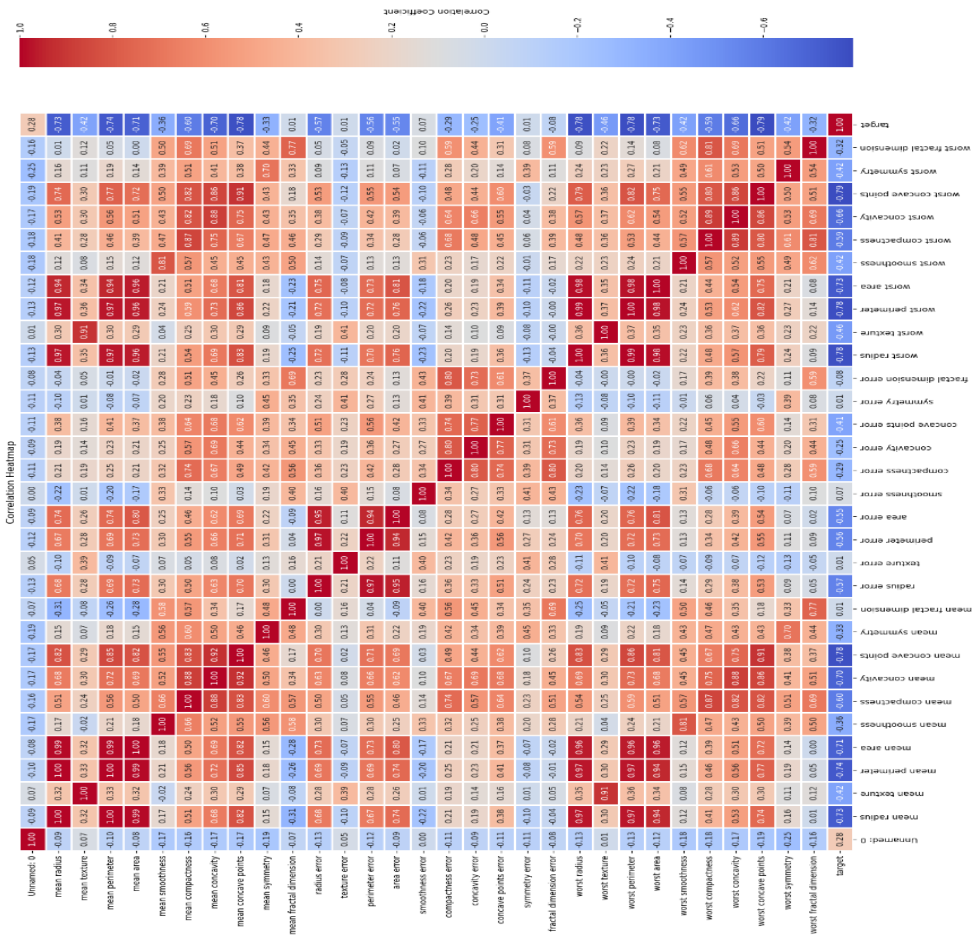

Figure 3.4. Heatmap (Correlation)

3.4. Feature Selection

A necessary preprocessing step in system studying, feature choice goals to boom generalization abilities, less computational complexity and improve model performance. The method includes identifying and maintaining the fine aspects even as getting rid of those which are superfluous or vain. In addition to improving version accuracy, accurate characteristic selection additionally allows to avoid issues like overfitting and underfitting, that may have a massive have an effect on prediction overall performance.

Importance of Feature Selection

To generate predictions, system getting to know models depend upon enter capabilities. Not every feature, although, makes a big contribution to the mastering technique. Certain features should bring about useless mastering with the aid of introducing noise, redundance, or needless complexity. The following are the principal blessings of feature selection:

• Increasing Model Accuracy: By doing away with superfluous characteristics, the version is not able to examine from noise, which improves forecast accuracy.

• Preventing Overfitting: Overfitting can occur when models with an immoderate variety of functions memorize training statistics rather than generalizing to clean facts. This problem may be lessened by using lowering characteristic dimensionality.

• Preventing Underfitting: While casting off useless functions is a good concept, doing so too much can bring about underfitting, that's while the version would not have enough facts to recognize patterns.

• Reducing Computational Complexity: A version with fewer traits is extra effective since it calls for much less memory and trains extra quickly.

CFS approach

The most relevant functions are systematically placed in this section, while those who are meaningless and poorly associated are removed using correlation -based construction strategies. This process includes the following movements:

Calculation of Strength between features: To check the link between the input Features X and the target variable Y, a correlation matrix is created. The correlation

coefficient of ρ for the piercing is often used to find out where focused and vigilant linear relationships between the variables.

- ✓ Eliminate strongly correlated features: Features are meaningless and eliminated with high correlation coefficients (ρ> 0.9) of 90% or more. Similar information contributes with highly correlated properties, resulting in multicollinearity, which can lead to interpretation and model performance.

- ✓ Removal of features with weak correlations: Facilities are eliminated with a correlation coefficient of less than 10% (ρ <0.1) with target variables. These properties have no future power and can cause noise, interfering with the model's ability to identify the pattern in question.

- ✓ Selection of the best features:
    - Only the most independent and educational features are spectacular once, and unnecessary properties are eliminated.
    - This guarantees a plant collection that is well balanced and a model makes a valuable contribution to both training and generalization.

Effect of convenience choice on model performance

The proposed approach improves the efficiency and strength of the model by systematically choosing relevant functions. The main benefits of this strategy are:

• **Better generalization**: Instead of learning exercise data from the heart, the model acquires knowledge from important patterns.

• **Better interpretation**: It is easy to think and understand through model decisions when the data set is small.

• **Increased training efficiency**: The time complications and resources required for ML model training are reduced when the functions are reduced.

Selected Features		
mean radius	mean texture	mean smoothness
mean compactness	mean concavity	mean symmetry
radius error	compactness error	concavity error
concave points error	worst smoothness	worst compactness
worst concavity	worst symmetry	worst fractal dimension

Table 3.2 Dataset Selected Features

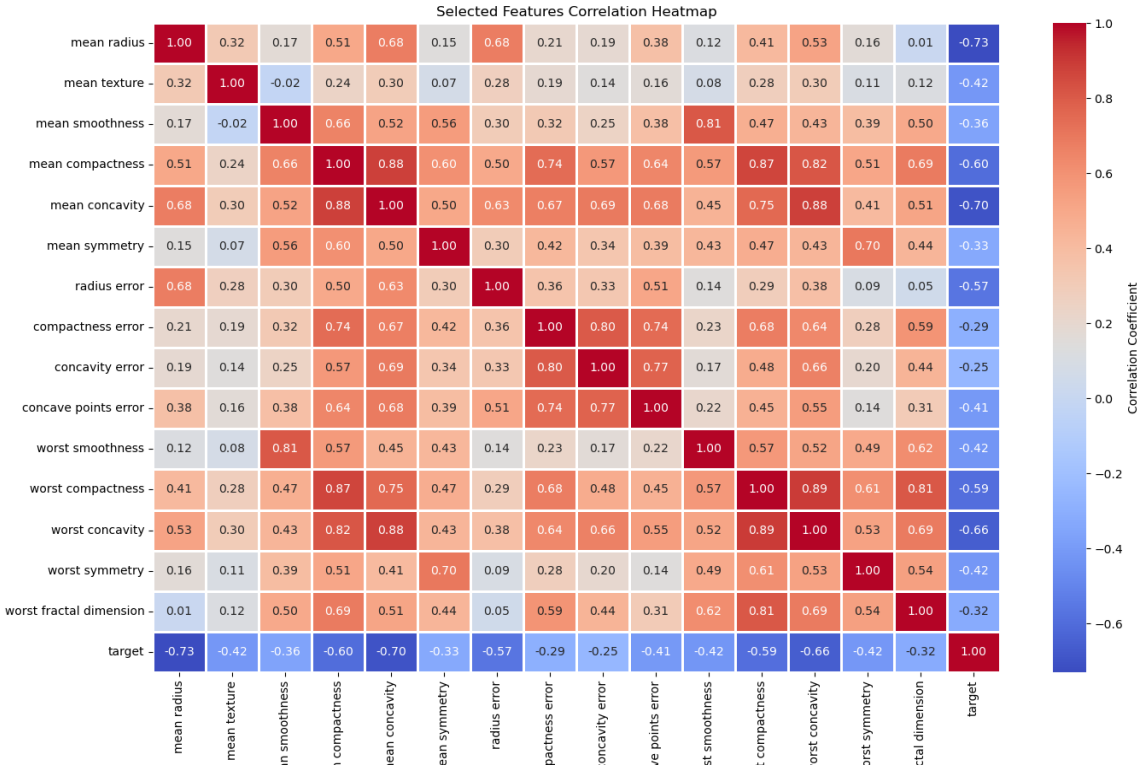

Figure 3.5. Selected Feature Heatmap (Correlation)

3.5. Dataset Partition

The machine learning is one important step is the data set partitions, one of the most important preparatory features, which involves dividing the dataset into two subdivisions Dataset: for a test and for a training. Usually, 20% of the data for test procedures are maintained, while the remaining 80% are used for training.

• **Training kit (80%):** ML model is trained using a training kit, which allows you to identify trends, correlations and patterns in available data. Based on this information, the ML model changes its internal limitations and parameters to produce exact predictions.

• **Test set (20%):** The performance of the model is tested on uncontrolled data using this superior. Measurement of important performance indicators matrix including accuracy, Recall, Precision, F1 score and CM, helps to assess how well the models generalize for new inputs.

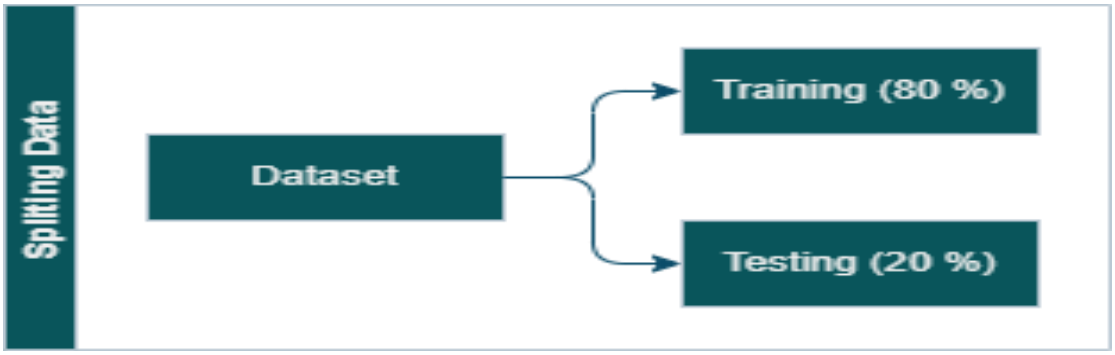

Figure 3.6. Dataset Split (Train, Test)

3.6. Standardization (Feature Scaling)

An important preparation step is standardization, which changes the functions of STD (σ) of 1 and an average (μ) of 0. This ensures that each feature contributes equally to the model, uneven convenience avoids bias by scales.

The formula is:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

We change training data by using standards from sklearn. Preprocessing, and we repeat changes in test data. This model improves stability and convergence, especially for machine learning techniques that are more sensitive to the order of magnitude, support vector machines, nervous networks and other models.

3.7. Model Selection

One of the best methods for deciding on the best version based on its effectiveness, efficiency, and potential for generalization is version choice. To confirm which model quality suits the supplied dataset and issue place, it entails comparing several fashions using exclusive performance indicators.

Training Multiple Models for Selection

One of the most vital and number one jobs in system mastering is model training. In this step, we teach numerous fashions the use of the education dataset, compare their performances, and choose the top-rated version for deployment.

• **LR**: This linear device studying model is by and large employed for bi-class.

• **DT**: This tree-primarily based device gaining knowledge of version divides facts in step with the significance of capabilities.

• **RF**: This ensemble getting to know approach uses decision trees to lessen overfitting, growth accuracy, and manage complex records.

• **SVM**: This device getting to know model determines the high-quality hyperplane for categorization.

● **KNN**: is a distance-based system studying model that uses nearby factors to categorize statistics.

• **XGB**: This gradient boosting approach is performance and speed optimized.

• **ADB**: An organization approach that makes use of boosting to reinforce susceptible freshmen.

3.8. Evaluation of Machine Learning Models

In machine learning, evaluating models is an essential step in determining their efficacy, reliability, and ability to generalize to new data. Accurate forecasts and most excellent overall performance in actual-world settings are guaranteed by a properly-organized evaluation method. Robust assessment methods are vital for confirming if the version can accurately distinguish between benign and malignant times within the prognosis of BC perspective. Because medical diagnoses convey such excessive stakes, an excellent model have to reduce fake positives (misclassifying B patients as M) and false negatives (no longer detecting malignant instances), each of that could have unfavorable outcomes on patient care and remedy.

Several assessment measures are used to assure thorough overall performance assessment; every one gives a unique viewpoint on the predictive electricity of the version.

These include:

- **Accuracy**: Indicates the percentage of effectively expected cases amongst all samples supplied. Although accuracy is a broadly used metric, it may not be the maximum beneficial signal whilst there are imbalanced classes, wherein one class has a high remember while the other has a low one.

- **Precision**: Shows the percentage of instances that had been appropriately expected to be malignant out of all cases that were anticipated to be such. In order to save you pointless clinical approaches, fewer fake positives are predicted to have excessive precision.

- **Recall**: Evaluates the model's capacity to discover every real instance of most cancers. Less fake negatives are guaranteed by means of high do not forget, which means that fewer cancer instances remain undiscovered.

- **F1-Score**: Provides a balance between don't forget and precision, giving it a greater applicable statistic in cases in which the dataset is unbalanced. It is in particular useful in clinical settings where FP and FN must be saved to a minimum.

- **CM**: A more thorough understanding of model mistakes is obtainable through the tabular show of real and anticipated classifications. It presents a clearer view of model overall performance thru the evaluation of FP, FN, TP, and TN.

• **ROC**: A visible depiction of the change-off among FP price and TP charge (don't forget). It is hired to determine the version's ability to distinguish across classes at numerous thresholds.

3.9. Hyperparameter Tuning

In system mastering version optimization, parameter tuning is an important step or episode that ambitions to enhance performance by choosing the premiere set of parameters. Hyperparameters are predetermined mixtures of parameters that drive the learning procedure and without delay affect a model's accuracy, generalization, and performance. They are distinct from the version default parameters and are learnt from the schooling facts. The model will now not underfit (be too easy, failing to seize patterns) or overfit (be too complicated, acting well on training data however poorly on unknown facts) if it's far nicely tuned.

Several hyperparameter mixtures have been systematically searched using GridSearchCV that allows you to gain the great version overall performance. By the use of this technique, the great set of hyperparameters that produce the excellent accuracy and most balanced precision-keep in mind values are selected. The optimised fashions confirmed better category overall performance via reducing FP and FN and growing basic reliability by hyperparameter satisfactory-tuning.

This manner turned into critical in verifying that the ML fashions correctly differentiated between benign and malignant BC instances, which improved their suitability for packages related to medical diagnosis. In the stop, deciding on the proper hyperparameters improves generalization and enables the model to feature properly with unknown enter.

3.10. Saving and loading a Model

The last step model for a machine learning model training is loading and savings, which allows the trained model to save for later and is used in practical applications. This process saves time and calculation resources by confirming that the model does not need to withdraw each time.

1. Sparing model: The model is sorted using a pickle library and stored as a (.pkl) file after training. This makes it possible to reuse without re -use the model.

2. Model load: The program uploads the program stored when a user interacts with the interface receives input data (e.g. symptoms or medical reports) and predict real -time.

3.11. User Interface for Machine Learning Model Deployment

Provision is the next step after users distributed an ML model on a user interface (UI) to interact with models and activate predictions. By using HTML, CSS and Kolbe, trained models have been integrated with an online interface to create a straight but practical application that makes sense for non-technical users.

Steps to Deploy a Model on a UI

Train & Save the Model: Use the Pickle library to train and save the ML model.

1. Use Flask to create a backend.
   o A backend server that loads the model and handles user inputs is made using Flask, a lightweight Python web framework.
   o the model inference is conducted and requests are handled by a Flask route (@app. route).

2. Use HTML and CSS to create the frontend, or user interface.
   o HTML: Provides the web page's structure, including the buttons, input forms, and result display.
   o CSS: Improves the user experience by styling the user interface.
   o JavaScript: May be included to improve interactivity.

3. Link the Flask Frontend and Backend.
   To serve the HTML page, use render template () in Flask.
   o Use POST requests to transfer user inputs from the frontend to the backend.

   o After processing the input and running the ML model, the backend generates predictions that are shown on the user interface.

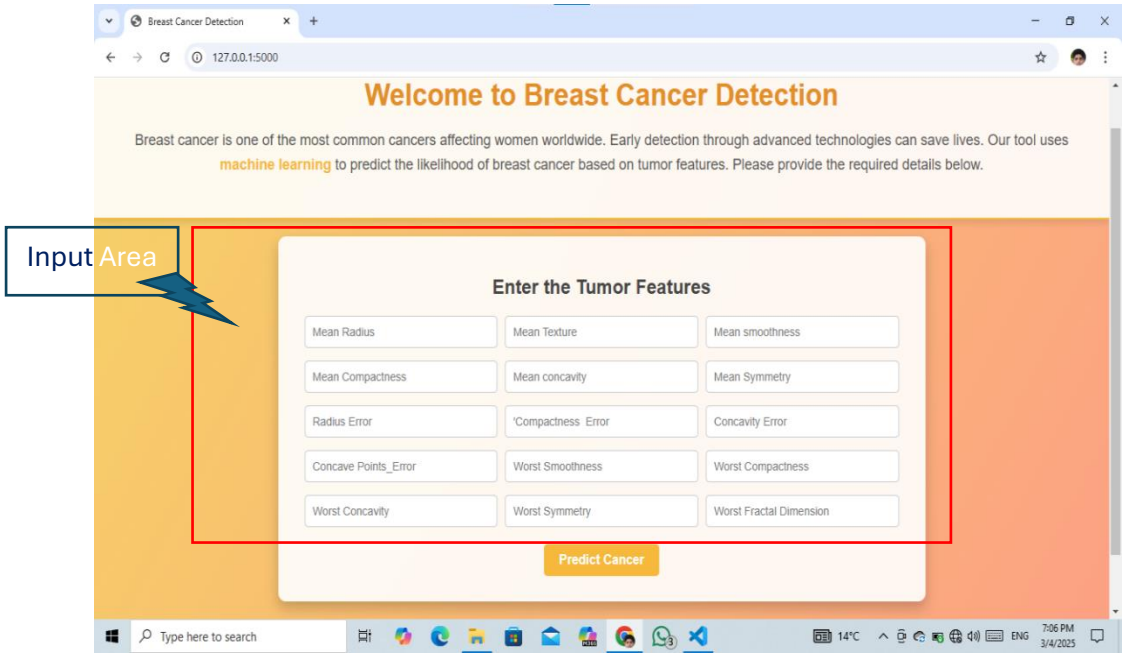

Figure 3.7. User Interface (UI)

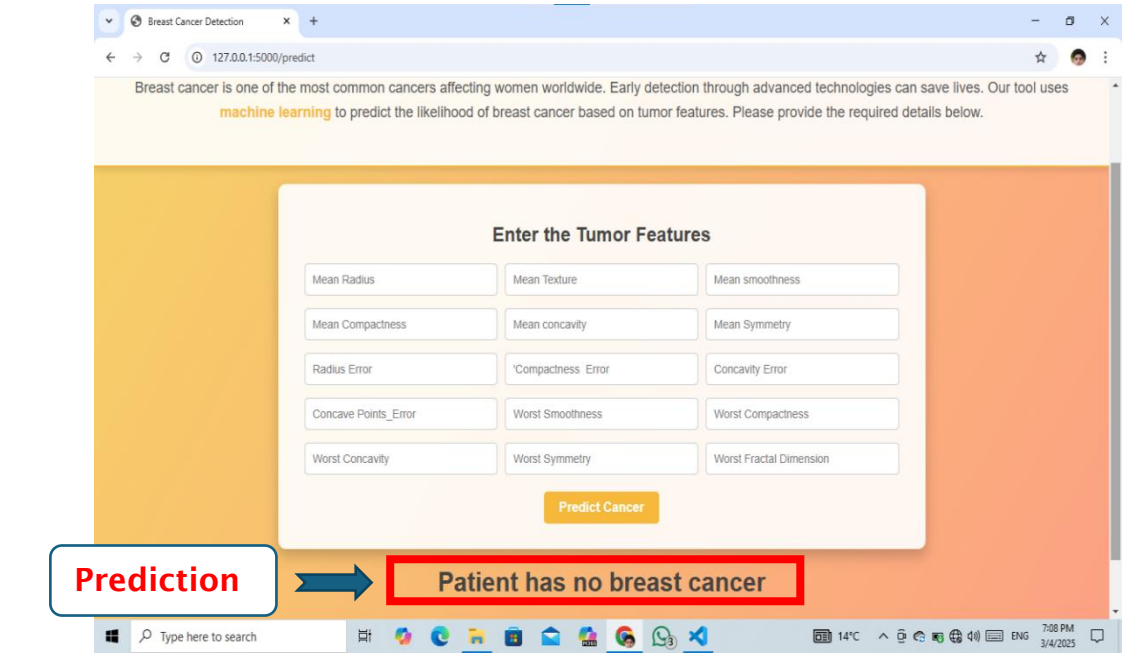

Figure 3.8. User Interface (UI) with prediction

Chapter 4: Results and Discussions

4.1. Introduction

Multiple machine learning (ML) algorithms were implemented in this study to enhance the detection and classification of BC within the developed system. Various performance criteria are used to evaluate the effect of the model, and a comparison is studied to determine which model is best. This chapter also includes model optimization and adjustment of hypermeters, as well as comparing previous research results. To find out how ML BC affects the classification, data set properties, proclamation of methods, model performance and evaluation strategies are all cautious.

4.2. Dataset summary

WDBC is one of the two popular BC datasets used in this study. These databases, which are openly available, include the necessary properties required for diagnosis of breast cancer.

Wisconsin Diagnostic Breast Cancer

Five hundred sixty-nine times or instances with thirty numerical capabilities—along with details about the residences of cell nuclei—are protected within the WDBC. Two instructions had been diagnosed from the dataset:

- Two hundred and twelve cases of malignant (M)
- Three hundred and fifty-seven cases of benign (B)

4.3. Model Performance Comparison

The overall performance of seven device studying models—LR, SVM, KNN, DT, RF, XGB, and ADB—is classed within the record. Accuracy, precision, bear in mind, F1-score, and CM were used to decide how precise these fashions were when they have been skilled on 80% of the facts and tested on 20%.

Performance Metrics

An evaluation of many fashions employing critical evaluation metrics, along with accuracy, F1-score, don't forget, precision, and CM heatmaps, is shown within the following images. These metrics provide us with a thorough insight of ways nicely the version works, specially whilst coping with unbalanced datasets.

As a properly-adjusted degree of accuracy and keep in mind, the F1 rating makes certain that both FP and FN are considered. When it comes to evaluating category performance, a better F1 rating shows a really perfect alternate-off between these standards. When it comes to scientific diagnostics, like BC detection, where reducing fake negatives is important, remember measures the model's capacity to correctly categorize fantastic instances. On the alternative hand, precision gauges how correct fantastic predictions are, making sure the model does not generate too many fake positives.

Furthermore, the confusion matrix heatmap visualization presents a smooth-to-apprehend depiction of version classification effects. Accurate predictions are shown via a huge concentration of values along the diagonal, whereas fewer misclassifications are indicated through minimal off-diagonal values. The excellent method for BC identification can be discovered by way of comparing those visible insights across many models, ensuring ideal classification performance with few mistakes.

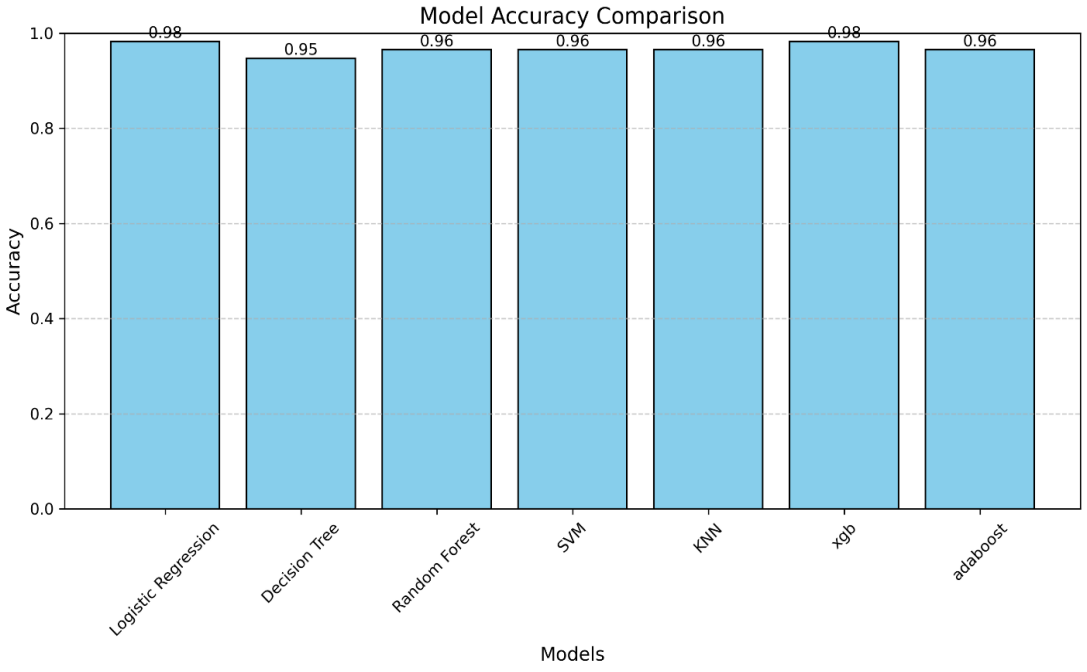

Figure 4.1. Model performance Comparison (Accuracy)

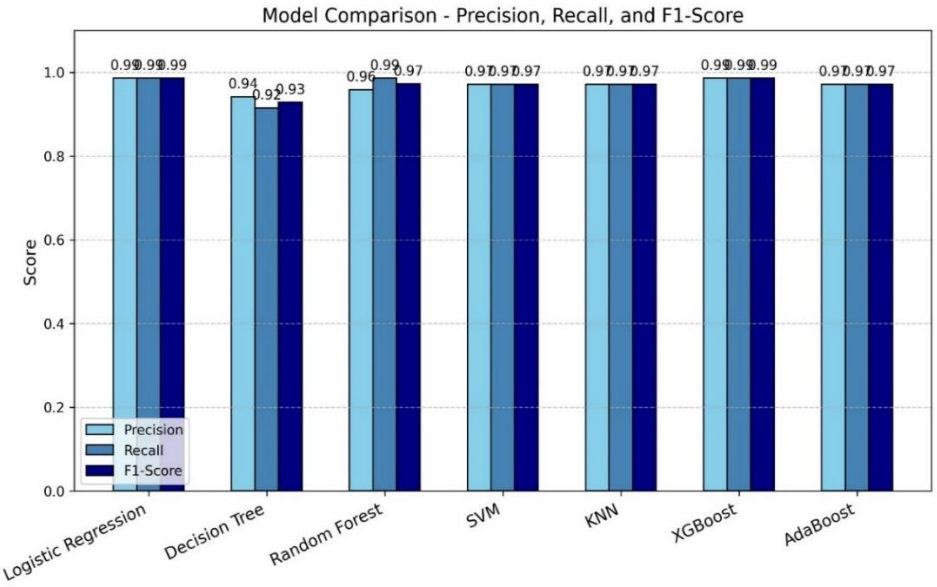

Figure 4.2. Model performance Comparison (F1, Recall, Precision)

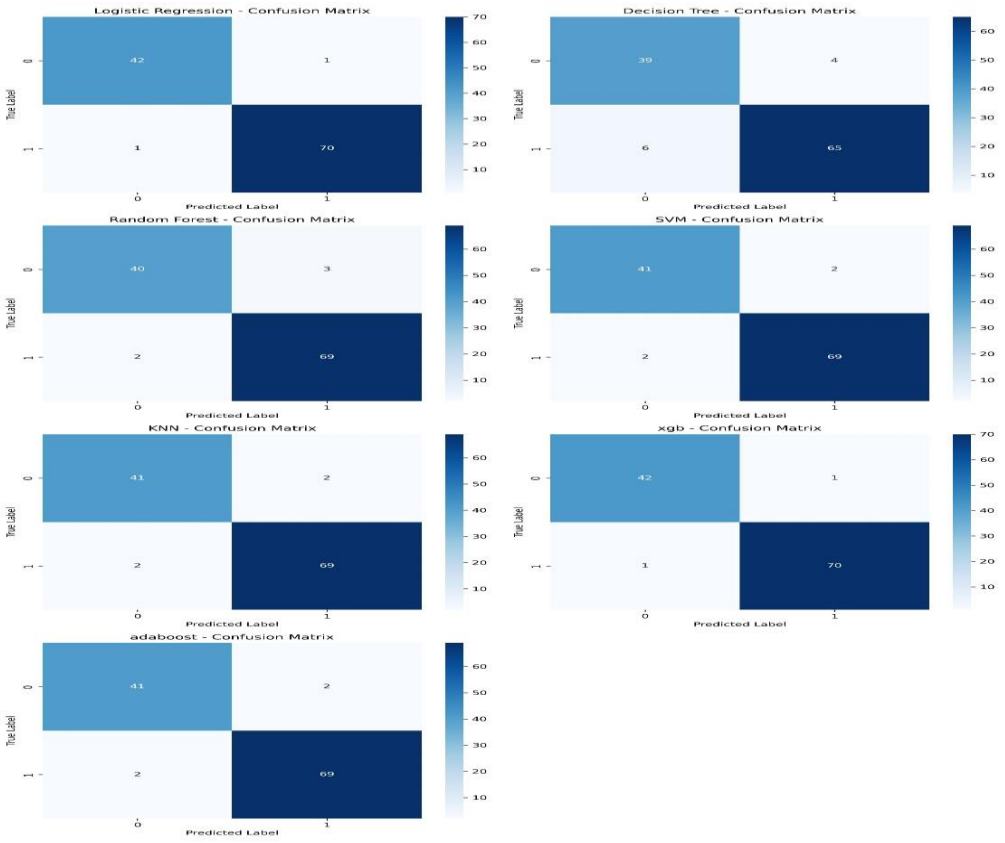

Figure 4.3. Model performance Comparison by Heatmap (Classification Report)

According to the results, **XGB** and **LR** outperform other models in terms of overall accuracy and other metrics.

Model Selected (LR)

Depending on accuracy, accuracy, recalling and F1 score, logistic regression (LR) and XGBOST (XGB) improved other models during evaluation. We chose the logistics region due to binary classification tasks, lecturers and its efficiency in low data costs, although XGB is a powerful dress learning technique with great future forces. Because it produces potential results, logistic regression is suitable for medical diagnosis jobs where the understanding of prediction is necessary, such as identification of breast cancer. In addition, unlike complex attire models such as XGBOOST, LR avoids overfit and works well with small datasets. This is the best option for our study due to simplicity, efficiency and implementation systems.

LR

A statistical model referred to as LR is carried out to bi-classification issues. It makes use of the sigmoid function to forecast the likelihood that an instance will belong to a specific magnificence. The version maps the output between 0 and 1 after estimating a linear connection between the impartial variables and the target variable.

Mathematical Representation

Logistic Regression makes use of the sigmoid activation characteristic:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

4.4. Hyperparameter Tuning and Optimization

We used GridserachCV for **Hyperparameter or parameter tuning** to improve the performance and normalization of the Logistics Region (LR) model. It was an important goal to detect the ideal combination of parameters that maximize accuracy when preserving generalization. We accommodate many important **Hyper parameter**, such as Solver ('Liblinear', 'Saga') to identify adaptation methods, to guarantee Max_iter, C (reversed by regularization power) to regulate model complexity and fined type for regularization (L1, L2). We demanded to improve the model performance by looking at

the future indication of the model for the model for breast cancer classification, reducing overfitting and systematically through different parameters' combinations.

```
Best Parameters for Logistic Regression: {'C': 1, 'max_iter': 100, 'penalty': 'l2', 'solver': 'liblinear'}
```

Figure 4.4 Best parameter for LR (Hyper Parameter)

Discussion of results

4.5. Selected Model Performance Analysis (LR)

Due to its strong performance in binary classification problems, we have chosen the LR model. Following the choice of the model, we used **Hyper parameter** adaptation to maximize performance and find the ideal ratio of F1 points, recall, accuracy and accuracy. The most important assessment indicators, including a cm, ROC curve and classification report, were used to conduct a comprehensive examination of the model. The ability of breast cancer detection models while reducing the FP and UN were validated by this survey.

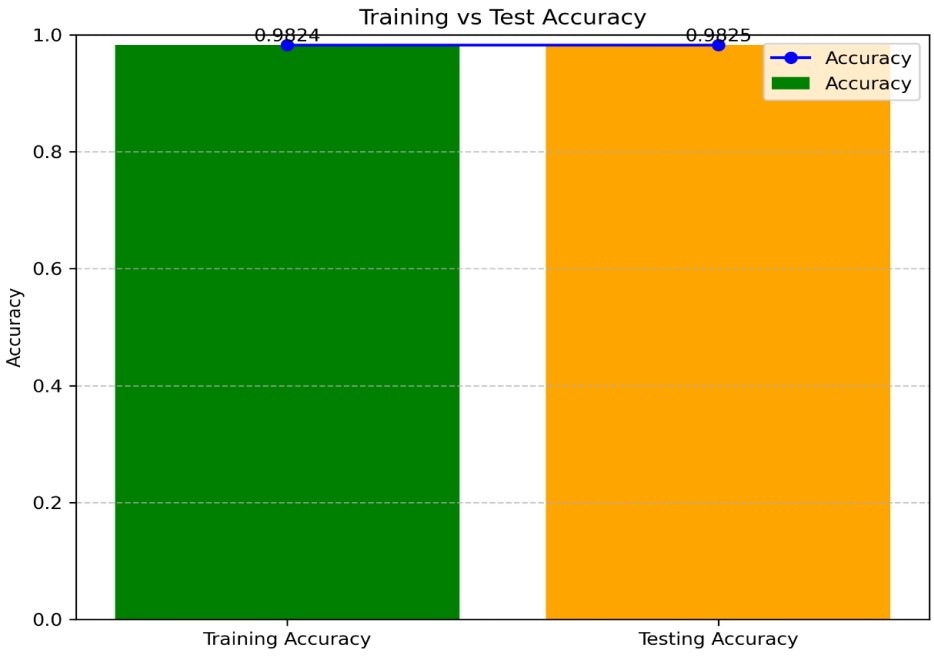

Figure 4.5. LR Accuracy Comparison (Training and Testing)

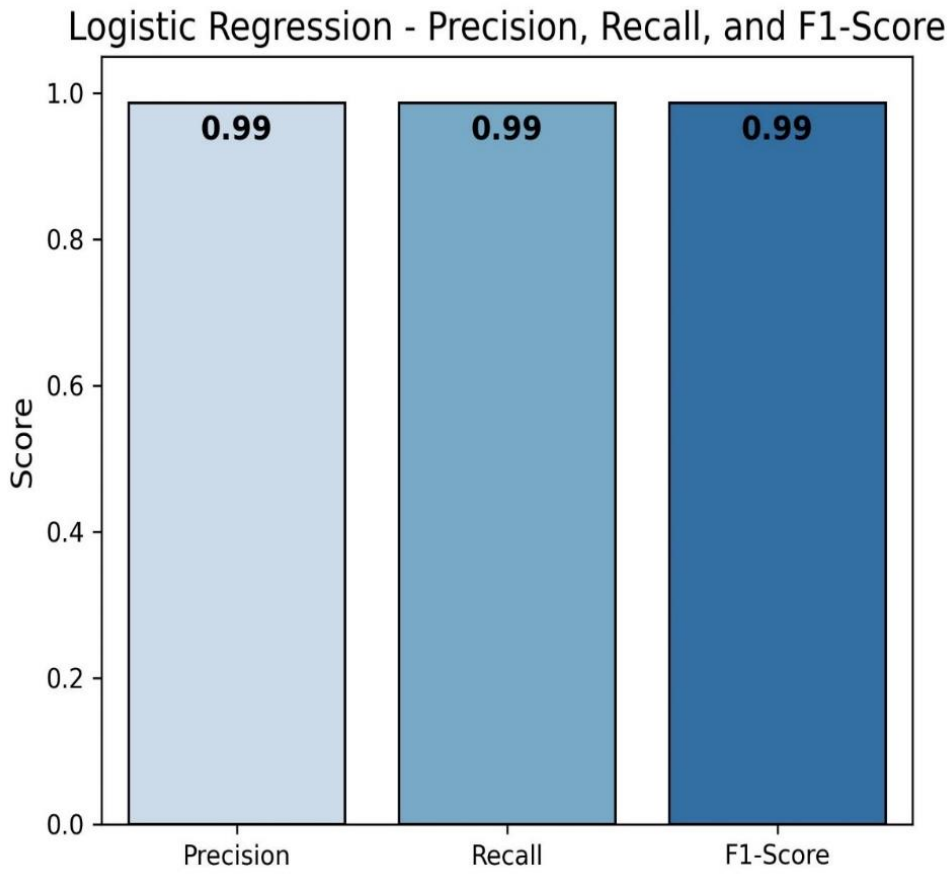

Figure 4.6. LR Precision, Recall & F1-Score

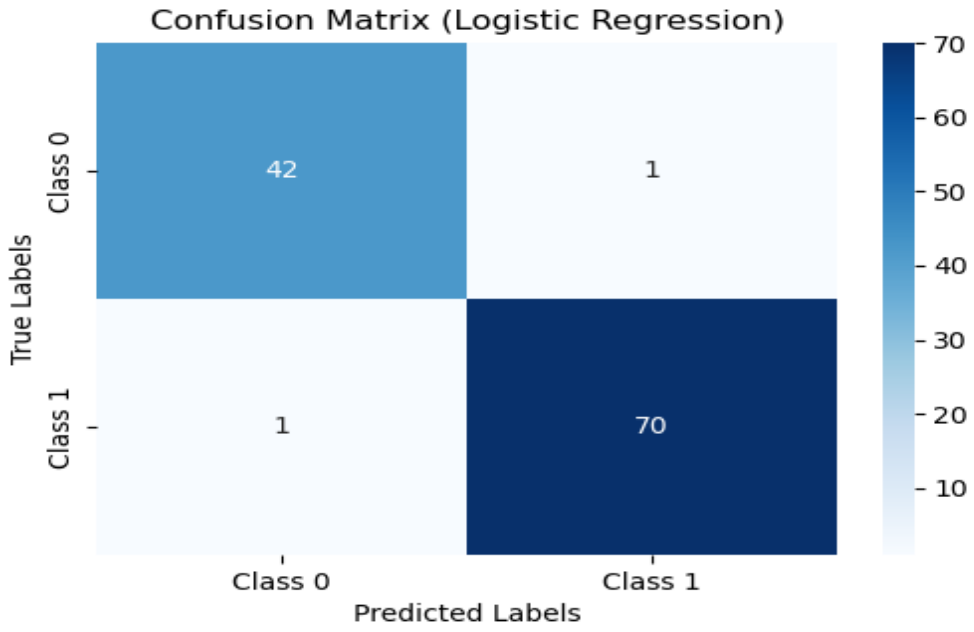

Figure 4.7. LR Classification-Report (Heatmap)

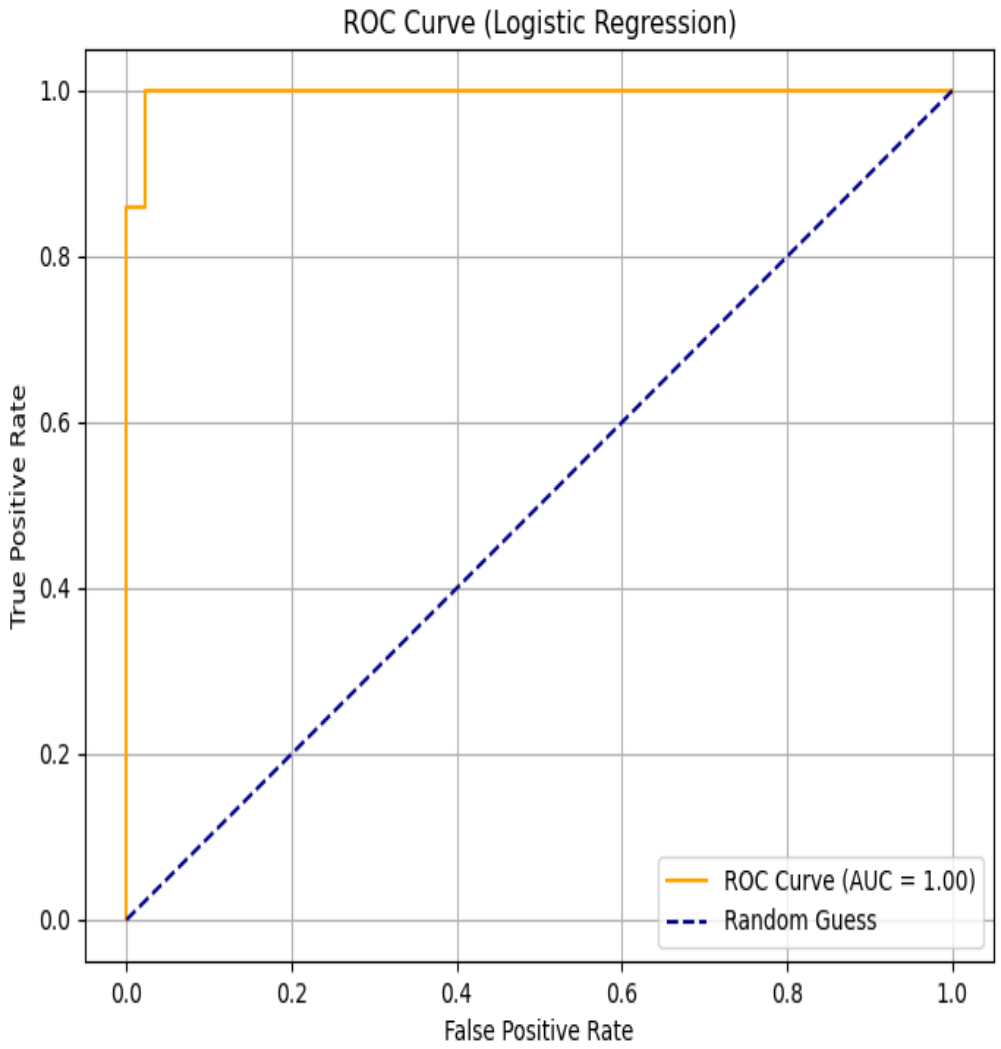

Figure 4.8. LR Roc curve

Comparison with Previous Studies

- Prior studies using **WDBC** dataset report accuracy values ranging from **92% to 96%.**
- Our study improves upon existing models by implementing **optimized LR (98.24%),** showcasing a significant performance gain.
- The results confirm that **LR algorithm outperform than other ML classifiers** in BC detection.

4.6 Summary

With an accuracy of 98.24%, the selected model, logistic regression, completed pretty properly in BC identification. The version's precision, consider, and F1 score are all 99% indicating an amazing balance among the two, that is vital for scientific diagnoses.

With best one FP and one FN, the CM demonstrates that the model had accurate dependability and appropriately identified maximum instances. Additionally, the version's correct ability to distinguish between positive and poor conditions is validated via the ROC curve of 1.00. These findings guide the perception that logistic regression with its high accuracy and interpretability is the first-rate alternative for implementation.

Chapter 5: Future Work

5.1. Introduction

Improvement in ML to detect BC has shown significant promise to reduce the degree of misbalance and increase early diagnosis. Although the classic machine learning model has been used effectively and adapted in this research, the prognosis requires improvement in accuracy, solving calculation problems and more improvement to investigate new approaches. Future research subjects included in this chapter include intensive teaching applications, real -time perfections, model lecturers, integration with medical imaging and moral problems.

5.2. Deep Learning for Breast Cancer Detection

While this research is focused on traditional ML models, DL techniques, especially CNN and RNN, which provide assurances to detect breast cancer. Future research can detect:

• **CNN for medical imaging**: DL models can analyze mammograms, histopathological slides and ultrasound images that can improve classification accuracy.

• **Learning Transfer**: Advance -trained DL models can be fined for BC detection, want to be much marked dataset.

• **Hybrid DL approach**: A combination of CNN with other architecture, such as transformers, can increase the function extraction and classification efficiency.

5.3. Integration of Machine Learning with medical imaging

Medical imaging is an important aspect of BC diagnosis. Future research should focus on integrating the ML model with Advanced Computer Vision System. Potential directions include:

• **Automatic image sharing**: U-Net or Mask R-CNN, Yolo to highlight tumor areas and classify deviations.

• **Multimodal learning**: A combination of image data with clinical reports and genetic markers for better clinical accuracy.

• **AI-Assisted Radiology**: Use machine learning in radiology workflows to make other opinion for oncologists.

5.4. Improving Model Interpretability and Explainability

Model Explainability and transparency need to be elevated if device learning-based totally diagnostic gear are to grow to be greater widely utilized in scientific settings. Future studies have to investigate the subsequent techniques for decoding version predictions:

- SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations).
- To decide which biomarkers, have the maximum have an impact on categorization, use feature importance analysis.
- XAI (Explainable AI) techniques to enhance the interpretability of models for clinical practitioners.

5.5. Real-Time Deployment and Cloud-Based Solutions

- Scalable and powerful deployment techniques are vital for the usage of device getting to know models in real healthcare programs. Future upgrades have to consist of:
- Cloud-Based Deployment: For scalable and dependable inference, use cloud-primarily based systems like Microsoft Azure, Google Cloud, and AWS.
- Using gadget mastering on clinical equipment or mobile packages for on-web site diagnostics is referred to as "Edge Computing for Faster Diagnoses".
- Models are optimized to be used with live-streaming information from medical imaging systems in actual-time processing.

5.6. Ethical Considerations and Bias Reduction

AI-pushed scientific systems need to adhere to ethical hints and decrease biases to make certain fair and reliable results. Future paintings need to consciousness on:

- Bias Detection and Mitigation: Addressing demographic and dataset-associated biases to improve model fairness.
- Diverse Data Collection: Expanding datasets to include instances from special populations for higher generalization.

- Compliance with Medical Regulations: Ensuring AI models meet HIPAA, GDPR, and FDA tips for moral AI use in healthcare.

5.7. Conclusion

Even although breast cancer diagnosis has greatly progressed way to machine learning, there may be nevertheless greater room for improvement. Real-time deployment, enhanced interpretability, DL integration, and multimodal scientific statistics will open the door for greater realistic AI-driven diagnostic solutions. Reliability and equity in medical packages may be further ensured through addressing moral issues and bias discount. Future studies can provide breast cancer detection structures that are more unique, without problems available, and scalable by increasing these regions.

References.

1. Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In *2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT)* (pp. 1-4). IEEE.
2. Osareh, A., & Shadgar, B. (2010, April). Machine learning techniques to diagnose breast cancer. In *2010 5th international symposium on health informatics and bioinformatics* (pp. 114-120).
3. Tahmooresi, M., Afshar, A., Rad, B. B., Nowshath, K. B., & Bamiah, M. A. (2018). Early detection of breast cancer using machine learning techniques. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(3-2), 21-27.IEEE.
4. Gayathri, B. M., Sumathi, C. P., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms-a survey. *International Journal of Distributed and Parallel Systems*, *4*(3), 105.
5. Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, *1*(5), 290.
6. Nallamala, S. H., Mishra, P., & Koneru, S. V. (2019). Breast cancer detection using machine learning way. *Int J Recent Technol Eng*, *8*(2-3), 1402-1405.
7. Bhise, S., Gadekar, S., Gaur, A. S., Bepari, S., & Deepmala Kale, D. S. A. (2021). Breast cancer detection using machine learning techniques. *Int. J. Eng. Res. Technol*, *10*(7), 2278-0181.
8. Elsadig, M. A., Altigani, A., & Elshoush, H. T. (2023). Breast cancer detection using machine learning approaches: a comparative study. *International Journal of Electrical & Computer Engineering (2088-8708)*, *13*(1).
9. Yadav, R. K., Singh, P., & Kashtriya, P. (2023). Diagnosis of breast cancer using machine learning techniques-a survey. *Procedia Computer Science*, *218*, 1434-1443.
10. Jia, X., Sun, X., & Zhang, X. (2022). Breast cancer identification using machine learning. *Mathematical Problems in Engineering*, *2022*(1), 8122895.
11. Masood, H. (2021). Breast cancer detection using machine learning algorithm. *International Research Journal of Engineering and Technology (IRJET)*, *8*(02), 738-747.

12. Alarabeyyat, A., & Alhanahnah, M. (2016, August). Breast cancer detection using k-nearest neighbor machine learning algorithm. In *2016 9th international conference on developments in eSystems engineering (DeSE)* (pp. 35-39). IEEE.
13. Mangukiya, M., Vaghani, A., & Savani, M. (2022). Breast cancer detection with machine learning. *International Journal for Research in Applied Science and Engineering Technology*, *10*(2), 141-145.
14. Kour, S., Kumar, R., & Gupta, M. (2021, October). Study on detection of breast cancer using Machine Learning. In *2021 International Conference in Advances in Power, Signal, and Information Technology (APSIT)* (pp. 1-9). IEEE.
15. Vaka, A. R., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning. *Ict Express*, *6*(4), 320-324.
16. Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, *2*(2), 13.
17. Ghorbian, M., & Ghorbian, S. (2023). Usefulness of machine learning and deep learning approaches in screening and early detection of breast cancer. *Heliyon*, *9*(12).
18. Kumar, M., Singhal, S., Shekhar, S., Sharma, B., & Srivastava, G. (2022). Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning. *Sustainability*, *14*(21), 13998.
19. Harinishree, M. S., Aditya, C. R., & Sachin, D. N. (2021, April). Detection of breast cancer using machine learning algorithms–a survey. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1598-1601). IEEE.
20. Yedjou, C. G., Tchounwou, S. S., Aló, R. A., Elhag, R., Mochona, B., & Latinwo, L. (2021). Application of machine learning algorithms in breast cancer diagnosis and classification. *International journal of science academic research*, *2*(1), 3081.