Predictive Breast Cancer Statistical Modelling for Early Diagnosis

Vipin Kumar[1], Amit Kumar Gupta[1], Ankit Verma[1], Nikhil Kumar[1], Dowon Kim[2]
Young-Jin Jung[2*], and Mangal Sain[3*],

[1]*Department of Computer Applications, KIET Group of Institutions, Ghaziabad, India;*
[2]*School of Healthcare and Biomedical Engineering, Chonnam National University, Yeosu 59626, Korea*
[3]*Division of Computer Engineering, Dongseo University, Busan, Republic of Korea*

Abstract: Breast cancer is a critical health issue affecting women globally. Early detection of breast cancer offers numerous benefits, such as increased access to treatment options and improved outcomes for patients. This research paper presents the outcomes of a two-day study focused on breast cancer detection. The objective was to create an efficient and accurate model for early diagnosis using machine learning techniques. The study utilized a diverse dataset comprising clinical, genetic, and imaging statistical data related to breast cancer. Data preprocessing techniques were applied to normalize variables and select relevant features. Statistical modeling techniques, including logistic regression, decision trees, and random forest, were employed to construct predictive models for breast cancer diagnosis. Key metrics such as accuracy, precision, recall, and area under the curve (AUC) were used to evaluate model performance. The results demonstrated promising performance, with high accuracy and AUC scores, indicating the potential of the developed models for accurate breast cancer detection. The findings of this study contribute to the existing knowledge on breast cancer detection by showcasing the effectiveness of machine learning techniques in achieving accurate and efficient early diagnosis. The study highlights the importance of utilizing diverse datasets and appropriate statistical modeling techniques for improved prediction of breast cancer.

Keywords: Breast Cancer, Early Detection, Machine Learning, Predictive Modeling, Diverse dataset

*Corresponding author: Mangal Sain: mangalsain1@gmail.com, Young-Jin Jung: yj@jnu.ac.kr

1. Introduction

Breast cancer is a dangerous and common health issue affecting millions of women and their families. It is the most commonly diagnosed cancer and the primary cause of cancer-related fatalities worldwide among women [1]. Early breast cancer detection is crucial for improving outcomes as it allows more treatment options and effective therapies, increasing survival rates [2]. Therefore, there is a need for developing accurate and effective breast cancer detection methods.

In recent years, the use of statistical modeling and machine learning techniques has tremendously increased for improving breast cancer detection [3]. These approaches have a low error rate, improving accuracy, and making breast cancer screening more accessible [4].

Traditional methods such as mammography and clinical breast examination have been the cornerstone of breast cancer detection for decades [5]. Although these methods have developed a good foundation, there are limitations and may not be the best choice. Mammography, for instance, may yield false-positive results, leading to unnecessary biopsies and causing anxiety for patients. Conversely, false-negative results can occur, missing cancerous lesions and delaying diagnosis [6].

To overcome these limitations, researchers have turned to statistical modeling and machine learning techniques. By leveraging large and diverse datasets, including clinical, genetic, and imaging data, these approaches have the potential to uncover hidden patterns, correlations, and risk factors associated with breast cancer [7]. They can identify subtle markers and signatures not easily discernible to human observers, improving the accuracy and reliability of breast cancer detection [8].

Moreover, these advanced techniques have the capacity to facilitate risk assessment and personalized treatment planning. By including patient data like age, family history, genetic factors, and lifestyle choices, predictive models can estimate the risk of developing breast cancer [9]. This information can inform targeted screening strategies and help healthcare professionals tailor interventions based on an individual's specific risk profile [10].

The application of statistical modeling and machine learning in breast cancer detection has shown promising results. Researchers have developed predictive models achieving high accuracy in distinguishing between benign and malignant breast lesions [11]. These models have the potential to reduce unnecessary biopsies, provide timely diagnosis, and improve patient care while also reducing healthcare costs [12].

Furthermore, these techniques have the potential to predict treatment response and disease progression. By analyzing longitudinal data and treatment outcomes, machine learning models can help identify factors influencing treatment efficacy and provide insights into optimal therapeutic approaches [13]. This information can guide clinicians in making informed decisions regarding treatment selection and monitoring [14].

However, despite advancements in breast cancer detection using statistical modeling and machine learning, several challenges still exist. Ensuring data quality, addressing privacy concerns, enhancing interpretability of models, and integrating these techniques into clinical practice are among the key areas that require further investigation and refinement [15]. Overcoming these challenges will be critical for the widespread adoption and implementation of these techniques in routine clinical settings.

2. Related Work

Recent advancements in breast cancer detection have been driven by a multitude of studies focusing on statistical modeling and machine learning techniques. These approaches aim to improve accuracy and early detection of breast cancer. This section shows an in-depth overview of key research papers in this domain, discussing their methodologies and contributions.

Smith, Johnson, and Brown [16] proposed a comprehensive statistical model for predicting an individual's risk of breast cancer. Their model incorporates several factors including genetic and environmental factors, age, family history, hormone factors, and lifestyle decisions. Through the utilization of logistic regression and survival analysis, the researchers achieved accurate predictions of breast cancer risk. This model provides a valuable tool for identifying individuals at higher risk and implementing preventive strategies and personalized interventions.

Lee, Kim, and Park [17] focused on predictive modeling of breast cancer progression. They employed machine learning algorithms and leveraged a large dataset consisting of diverse clinical features, such as tumor size, lymph node status, and histological grade. By utilizing techniques like random forest and support vector machines, the researchers aimed to develop accurate models for predicting the progression of breast cancer. Their findings highlight the potential of early intervention and personalized treatment plans to improve patient outcomes.

Chen, Wang, and Zhang [18] developed a Bayesian network technique for breast cancer diagnosis. Their probabilistic graphical model integrates multiple clinical factors, including mammographic findings, patient demographics, and biopsy results. By capturing the complex interdependencies between these factors, the model provides a comprehensive assessment of the likelihood of breast cancer presence. This approach enhances diagnostic accuracy and enables healthcare professionals to make informed decisions regarding treatment options.

Gupta, Verma, and Singh [19] explored the use of machine learning techniques for breast cancer diagnosis, considering both clinical characteristics and imaging data. Their models integrated clinical features such as age and symptoms with imaging characteristics derived from mammograms, such as shape and texture features. By employing artificial neural networks and support vector machines, they achieved high accuracy in early breast cancer detection. Use of multiple data sources and advanced machine learning algorithms can significantly improve the efficiency and reliability of breast cancer diagnosis.

Li, Chen, and Liu [20] conducted a comprehensive review of statistical modeling techniques for breast cancer survival data. They examined various approaches utilized in this field, including Cox proportional hazards models, parametric survival models, and machine learning algorithms. The article provides insights into the advantages and disadvantages of each modeling technique, guiding researchers in selecting the most appropriate approach for predicting breast cancer survival outcomes. The review highlights the importance of considering the heterogeneity of breast cancer and tailoring models to individual patient characteristics.

Kumar, Sharma, and Kumar [21] presented an overview of machine learning algorithms used in breast cancer early detection. They emphasized the potential of techniques such as deep learning, ensemble learning, and feature selection methods to improve detection accuracy and patient outcomes. The review provides a comprehensive understanding of the diverse array of machine learning approaches relevant to early-stage breast cancer diagnosis. By using the power of these advanced algorithms, researchers can enhance the efficiency of breast cancer screening and reduce errors.

In addition to previously mentioned studies, several other research papers have significantly contributed to the field of breast cancer detection. For instance, Li et al. [22] proposed a hybrid model combining genetic algorithms and support vector machines for breast cancer diagnosis, achieving high accuracy and reducing computation time. Tang et al. [23] used deep learning techniques, specifically convolutional neural networks for the automated analysis of mammograms for efficient detection of breast abnormalities.

Furthermore, Kourou et al. [24] conducted a review on the application of deep learning in breast cancer imaging. They discussed various deep learning architectures, including convolutional, recurrent neural networks, and generative adversarial networks, along with their applications in breast cancer imaging analysis.

The above-discussed studies contribute to the advancement of breast cancer detection techniques by providing valuable knowledge for the development of accurate and efficient models for risk prediction and early detection.

3. Research Framework

3.1 Architecture

The Predictive Breast Cancer Statistical Information Modelling for Early Diagnosis architecture combines statistical modelling and machine learning approaches. The following is a high-level overview of the architecture:

3.2 Data Pre-processing and Statical Modeling

The research paper is based on the Wisconsin Breast Cancer Diagnostic dataset, a widely used dataset donated by researchers from the University of Wisconsin and available in the UCI Machine Learning Repository [25]. This dataset comprises measurements taken from digitized images of fine-needle aspirates of breast masses, providing a robust foundation for studying the characteristics and effects of breast cancer.

The section on "Data Pre-processing and Exploratory Data Analysis" focuses on important steps involved in preparing the data for analysis and gaining information about breast cancer detection through exploratory data analysis. this section provides a valuable summary of data preparation and data analysis method of this study.

Data pre-processing is an important step in guaranteeing the data's quality and usefulness for analysis. It encompasses tasks such as handling missing values, removing outliers, standardizing or normalizing variables, and encoding categorical variables. By addressing these data quality issues, researchers can enhance the reliability and accuracy of the subsequent analysis [26].

The dataset consists of several key columns, each providing essential information for the diagnosis of breast cancer. One of the most crucial columns is the Diagnosis column, serving as the target variable. It indicates whether the breast mass is classified as malignant (M) or benign (B), providing vital information for identifying the presence or absence of cancer.

In this study, a meticulous analysis was conducted on the following features within the dataset to gain a better understanding of their significance in breast cancer detection:

1. **Radius:** This feature represents the mean distances from the centre to points on the perimeter of the breast mass.
2. **Texture:** This feature quantifies the variation in grayscale intensities of the image, reflecting the smoothness or roughness of the mass.
3. **Perimeter:** This feature measures the total length of the boundary of the breast mass.
4. **Area:** This feature represents the area enclosed by the boundary of the breast mass.
5. **Smoothness:** This feature characterizes the local variation in radius lengths, indicating the deviation of the mass boundary from a smooth contour.
6. **Compactness:** This feature represents the compactness of the mass, calculated as the perimeter squared divided by the area.

Thoroughly analysing these features aims to uncover their significance in breast cancer detection and explore their potential contribution to the development of accurate diagnostic models.

To summarize, the "Data Pre-processing and Exploratory Data Analysis" section focuses on the vital steps involved in preparing the Wisconsin Breast Cancer Diagnostic dataset for analysis. By addressing data quality issues and conducting exploratory data analysis, valuable insights can be gained into breast cancer detection, ultimately contributing to the development of effective diagnostic models.

Once the data has been pre-processed, section start with exploratory data analysis(EDA) techniques used for breast cancer detection. EDA aims to uncover patterns, relationships, and insights within the data by visualizing and summarizing its main characteristics. Through EDA, we gain a deeper understanding of the dataset, identify potential outliers, assess data quality, and explore variables' distributions.

3.2.1 Descriptive Statistics

Descriptive statistics are fundamental in exploratory data analysis (EDA), as they provide important insights into the dataset. These statistics summarize the key characteristics of each feature, giving an overview of the central tendencies, spread, and range of the variables.

The breast cancer dataset consists of various features, and computing descriptive statistics for each feature can provide valuable information. Some of the key measures are mean, median, minimum and maximum values. These statistics offer a concise summary of the dataset's numerical characteristics and aid in understanding the distribution and variability of the data.

The graph below illustrates the descriptive statistics of the features in the breast cancer dataset. Each of these features are represented on the y axis while the x axis displays the values of the statistics which includes count, mean, standard deviation, minimum, 25th percentile(Q1), median (50th percentile or Q2), 75th percentile(Q3) and maximum

This visual representation allows us to quickly identify important statistical measures for each feature. For instance, the mean provides an estimate of the feature's central value, while the standard deviation indicates the spread or dispersion of the data points around the mean. The minimum and maximum values define the range within which the feature values vary, and the quartiles (Q1, Q2, Q3) offer insights into the distribution's shape and skewness.

Analyzing the descriptive statistics graph enables us to gain initial insights into the dataset. They can identify features with higher variability or extreme values, which may require further investigation. Moreover, we compare the statistics across different features to identify patterns or relationships within the dataset.

3.2.2 Distribution Analysis

The distribution analysis of the breast cancer dataset provides valuable insights into the characteristics and patterns of each feature. Histograms are used to visualize the distributions, allowing for a clear understanding of the concentration and dispersion of data points.

In the graph, subplots are utilized to represent the different characteristics of the breast cancer dataset. Each histogram in the subplots displays feature values on the x axis and frequency or density on the y axis. By investigating histograms, various aspects of the feature distributions can be seen, such as presence of peaks, modes and shape of distribution.

Histograms helps in finding whether distributions are symmetric or asymmetric which can provide important information about the existing patterns in the data. Additionally, they can help detect outliers or any unusual patterns that may require investigation. By studying the feature distributions through histograms, researchers can gain a succinct and useful overview of the dataset, revealing interesting traits and trends.

Understanding the distributional characteristics of the data is crucial for making informed decisions throughout the breast cancer detection research analysis and modeling processes. It enables us to identify potential challenges, select appropriate modeling techniques, and interpret the results accurately. By gaining insights into the distribution patterns, we can develop more robust models and improve the effectiveness of breast cancer detection methods.

3.2.3 Correlation Analysis

We performed correlation analysis to figure out relationship between each feature and diagnosis of breast cancer. Correlation analysis is a statistical technique which is used to measure the strength and direction of association between two variables.

To examine the correlation between the features and the diagnosis, we constructed correlation matrices and generated heat maps. The correlation coefficient was used to quantify the strength of the association. A negative correlation indicates an inverse relationship between the variables, while a positive correlation suggests a direct relationship.

By analyzing the correlation matrices and heatmaps, we aimed to uncover any significant correlations between the features and the diagnosis of breast cancer. These findings would provide valuable insights into the potential predictive power of each feature and their relevance to breast cancer detection.

Our analysis uncovered several significant findings related to breast cancer detection:

1. **Correlation with Diagnosis:** We observed a strong positive correlation between the

diagnosis and certain features, such as concave points worst, perimeter worst, concave points mean, and radius worst. This suggests that these features hold significant predictive power in classifying breast cancer cases.

2. **Differences in Feature Distributions**: We find different differences in the distributions of various features between malign and benign cases. Important features exhibiting such differences include radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, symmetry mean, and fractal dimension mean. The disparities in these feature distributions indicate their potential importance in distinguishing between malignant and benign breast cancer cases. Adding these features in our predictive models can improve accuracy of breast cancer classification and help in early detection and intervention.

These findings offer valuable information about potential importance of specific features in the early diagnosis of breast cancer. By understanding correlations and distributions of these features, we can develop better predictive models for early breast cancer detection. In the upcoming sections, we will look into details of the data pre-processing steps and statistical modeling techniques employed to use these findings and construct robust predictive models.

3.3 Model Building

This research paper focuses on the development and evaluation of various machine-learning algorithms for the classification of breast cancer data. The primary objective is to compare the performance of these models and identify the most accurate classifier. The following machine learning algorithms were implemented and analyzed in this study:

3.3.1 Support Vector Classifier (SVC):

The Support Vector Classifier (SVC) is based on the Support Vector Machine (SVM) algorithm, which aims to find the optimal hyperplane that separates different classes of data points while maximizing the margin between them. This is achieved by minimizing a combination of the regularization term and the loss term. The regularization term encourages a simpler model, while the loss term penalizes training errors and margin violations. The objective of the SVC can be mathematically represented as minimizing the following equation:

$$\min_{w,b,\xi}\left(\frac{1}{2}\parallel w\parallel^2+C\sum_{i=1}^{N}\xi_i\right)$$

Subject to:

$$y_i(w\cdot x_i+b)\geq 1-\xi_i, \forall i\in\{1,2,\ldots,N\}$$
$$\xi_i\geq 0, \forall i\in\{1,2,\ldots,N\}$$

where w represents the weight vector, b is the bias term, ξ denotes the slack variables, y_i is the class label of the i^{th} data point, x_i represents the feature vector of the i^{th} data point, and C is a parameter that controls the trade-off between the margin and the training errors.

3.3.2. Logistic Regression:

Logistic Regression is a widely used algorithm for binary classification tasks. It estimates the probability of a specific outcome, allowing researchers to make predictions or decisions based on the calculated probabilities. The logistic regression model calculates the probability using the sigmoid function:

$$P(y=1\mid X)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n}}$$

where P(y=1|X) represents the probability of the positive class given the input features X, β_0 represents the intercept term, and β_1, β_2, ..., β_n represent the coefficients associated with each feature x_1, x_2, ..., x_n.

3.3.3 K-Nearest Neighbor (KNN) Classifier:

The K-Nearest Neighbor (KNN) algorithm is a simple yet powerful approach for classification and regression tasks. It assigns class labels to new data points based on the majority vote of their K nearest neighbors. The algorithm calculates the distance between data points using metrics such as Euclidean or Manhattan distance. By selecting the K nearest neighbors, it determines the class label or predicts the target value. The KNN classifier can be summarized by the following equation:

$$\begin{aligned}&Classify(sample)\\&= MostCommonClass(KNearestNeighbors(sample))\end{aligned}$$

where sample represents the data point to be classified, and Most Common Class returns the class label that occurs most frequently among the K nearest neighbors.

3.3.4 Naive Bayes Classifier:

The Naive Bayes classifier is a probabilistic algorithm that assumes independence between features. It estimates the posterior probability of a class label given the feature values using Bayes' theorem and assumes feature independence. The Naive Bayes classifier can be represented by the following equation:

$$P(y|x_1,\dots,x_n) = P(y) * \mathit{\Pi} P(x_i|y)/P(x_1,\dots,x_n)$$

where P(y|x_1,...,x_n) represents the posterior probability of class y given the feature values x_1, x_2, ..., x_n. P(y) is the prior probability of class y. P(x_i|y) is the likelihood of feature x_i given class y and P(x_1,..., x_n) is the evidence or marginal probability of the features.

3.3.5 Decision Tree Classifier:

Decision trees create a set of rules based on the features to classify data. They recursively split data based on feature thresholds to create decision rules. The decision tree classifier can be summarized by the following equation:

$$Decision(x) = LeafNode(x)$$

3.3.6 Random Forest Classifier:

Random Forest is an ensemble algorithm that combines multiple decision trees. It constructs an ensemble of decision trees and aggregates their predictions through voting or averaging. The random forest classifier can be represented as follows:

$$Prediction(x) = MajorityVote(Prediction_i(x)), for\ i\ in\ 1\ to\ N$$

where Prediction$_i$ (x) represents the prediction of the i-th decision tree for the input features x, and Majority Vote returns the class label that occurs most frequently among the predictions of the individual decision trees.

3.3.7 Adaboost Classifier:

Adaboost is an ensemble algorithm that combines weak classifiers to create a strong classifier. It assigns higher weights to misclassified samples to iteratively improve classification performance. The Adaboost classifier can be summarized by the following equation:

$$Prediction(x) = Sign\left(\Sigma(\alpha_i * Prediction_i(x))\right), for\ i\ in\ 1\ to\ N$$

where Prediction_i (x) represents the prediction of the i-th weak classifier for the input features x, α_i represents the weight assigned to the i-th weak classifier, Sign returns the sign of the summation, and N represents the number of weak classifiers.

3.3.8 XGBoost Classifier:

XGBoost is an optimized implementation of gradient boosting, which combines weak models to create a powerful one. It applies gradient boosting principles to iteratively train a sequence of weak models. The XGBoost classifier can be summarized by the following equation:

By examining the performance of these models, we aim to identify the most effective algorithm for breast cancer classification.

$$Prediction(x) = \Sigma(\gamma_i * Prediction_i(x)), for\ i\ in\ 1\ to\ N$$

where Prediction_i (x) represents the prediction of the i-th weak model for the input features x, γ_i represents the weight assigned to the i-th weak model, and N represents the number of weak models. The predictions are combined by summation.

The research paper evaluated various machine learning algorithms for breast cancer classification. The accuracy scores of each method on the breast cancer dataset are as follows:

Model	***Accuracy Score***
Support Vector Classifier	90.35
Logistic Regression	92.11
K-Nearest Neighbor	91.23
Naive Bayes	93.86

Model	***Accuracy Score***
Decision Tree	91.23
Random Forest	95.61
Adaboost	91.23
XGBoost	94.74

Table 1. Models and their accuracy scores

After evaluating multiple machine learning models, we found that the Random Forest classifier achieved the highest accuracy among the tested algorithms. To further optimize its performance, we applied a technique called grid search.

Grid search is a systematic approach that helps in finding the best combination of hyper parameters for a given model. Hyper parameters are parameters that are set before the learning process and affect the model's performance. In the case of the Random Forest classifier hyper parameters such as maximum depth of trees and the number of features to consider at each split can greatly impact its accuracy.

By defining a range of possible values for each hyper parameter, grid search exhaustively searches through all possible combinations and evaluates the model's performance using cross-validation. The combination of hyper parameters that yields the highest accuracy is considered the best estimator for the given dataset.

By defining a range of possible values for each hyper parameter, grid search mainly searches through all possible combinations and evaluates the model's performance using cross-validation. The combination of hyper parameters that yields the highest accuracy is considered the best estimator for the given dataset.

In our study, we applied grid search on the Random Forest classifier and defined a range of values for the maximum depth and maximum features. The best combination of hyper parameters was found to be a maximum depth of 10 and maximum features of 12. This optimized Random Forest classifier achieved an accuracy of 96.70%, indicating its potential for accurately classifying breast cancer cases.

The application of grid search demonstrates the importance of hyper parameter tuning in maximizing the performance of machine learning models. By finding the optimal hyper parameter values, we can enhance the accuracy and reliability of the Random Forest classifier in breast cancer diagnosis, leading to improved patient outcomes and treatment strategies.

These models were evaluated on the breast cancer dataset, and their accuracy scores were recorded. The results indicate the effectiveness of each model in classifying breast cancer cases, with Random Forest achieving the highest accuracy. The findings help in understanding machine learning algorithms in medical diagnostics and can even assist in improving breast cancer detection and treatment.

3.4 Model Evaluation and Fine Tuning:

Once the predictive models have been constructed, it is essential to evaluate their performance to ensure optimal results. This section focuses on the evaluation metrics used to assess the models. To evaluate the performance of the predictive models for breast cancer diagnosis, several evaluation metrics can be utilized.

The confusion matrix is a valuable tool for evaluating the performance of a classification model, and it provides insights into the model's ability to correctly predict different classes. In the context of research paper confusion matrix can be used to find out the effectiveness of a breast cancer diagnosis model.

The confusion matrix have four key components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These components represent the counts or frequencies of the model's predictions compared to the actual class labels.

In the case of a binary classification problem like breast cancer diagnosis, the confusion matrix can be represented as follow:

Confusion matrix of Random Forest model:

The following metrics are commonly employed in binary classification tasks:

1. **Accuracy:** Accuracy measures the proportion of correctly classified instances out of the total number of instances. It provides an overall assessment of the model's performance. The formula for accuracy is straightforward and is calculated as follows:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

where:

TP (True Positive) represents the number of instances that are correctly classified as positive.

TN (True Negative) represents the number of instances that are correctly classified as negative.

FP (False Positive) represents number of instances that are falsely classified as positive when they are actually negative.

FN (False Negative) represents number of instances that are falsely classified as negative when they are actually positive.

To figure out accuracy of our model we add the number of positive and negative predictions and divide that number by total number of events in dataset, consisting both correct and incorrect predictions. This gives us a ratio that ranges from 0 to 1 where a value closer to 1 shows a higher accuracy and a better-performing model.

Accuracy provides an overall result of model's performance by considering both true positive and true negative predictions. It gives us an understanding of how well the model is able to correctly classify instances from different classes. In cases of imbalanced datasets or when the costs of false positives and false negatives differ substantially, however, accuracy may not always be the most appropriate metric. In such scenarios, additional metrics such as precision, recall, or F1 score may provide a more comprehensive evaluation of the model's performance.

2. **Precision:** Precision calculates the proportion of true positive predictions out of the total positive predictions. It quantifies the model's ability to correctly identify malignant cases. The formula for precision is as follows:

$$Precision = TP/(TP + FP)$$

Where:

TP (True Positive) represents the number of instances correctly classified as positive.

FP (False Positive) represents the number of instances falsely classified as positive when they are actually negative.

Precision focuses on the correctness of positive predictions and provides insights into the model's ability to avoid false positives. It quantifies how precise or accurate the model is in identifying positive instances.

A high precision value indicates a low false positive rate, meaning that the model is effective in correctly identifying positive instances and minimizing the occurrence of false positive predictions. On the other hand, low precision value shows a high number of false positive predictions which means a higher risk of incorrectly labelling negative instances as positive.

Precision is mainly important in cases where the outcomes of false positives are significant such as medical diagnoses. For instance, in the context of identifying malignant cases in a medical application, precision helps find out the model's ability for correctly identifying malignant cases without falsely classifying non-malignant cases.

3. **Recall (Sensitivity):** Recall measures the proportion of true positive predictions out of the total actual positive instances. It evaluates the model's ability to correctly detect malignant cases. The formula for recall is as follows:

$$Recall = TP/(TP + FN)$$

Where:

TP (True Positive) represents the number of instances correctly classified as positive.

FN (False Negative) represents the number of instances falsely classified as negative when they are actually positive.

Recall focuses on capturing the number of positive instances correctly identified by the model, thereby assessing its ability to avoid false negatives. It quantifies how well the model detects positive instances from the entire set of actual positive instances.

A high recall value indicates a low false negative rate, suggesting that the model effectively identifies positive instances and minimizes the occurrence of false negative predictions. On the contrary, a low recall value implies a higher number of false negatives, indicating that the model may miss or overlook positive instances.

In the context of identifying malignant cases in a medical application, recall helps assess the model's ability to correctly detect malignant cases without missing or incorrectly labelling them as negative.

4. **Support:** Support measures the frequency or prevalence of a specific item set or pattern in a dataset. It quantifies the proportion of instances in the dataset that contain the item set or satisfy the pattern. The formula for support is as follows:

$$Support = (Numberofinstancescontainingtheitemset)/(Totalnumberofinstances)$$

Support focuses on capturing the prevalence or frequency of a pattern in the dataset, indicating how common or popular it is among the instances. It helps identify frequently occurring item sets or patterns that have significant support in the dataset.

A high support value indicates that the item set or pattern occurs frequently in the dataset, suggesting its importance or relevance. On the other hand, a low support value implies that the item set or pattern is relatively rare or infrequent.

Support is particularly useful in association rule mining, where it helps identify meaningful and frequent associations or relationships between items. By setting a minimum support threshold, analysts can filter out less frequent or insignificant patterns and focus on the ones with higher support.

5. **F1 Score:** The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance by considering both precision and recall. The formula for calculating the F1 score is as follows:

$$F1Score = 2 * (Precision * Recall)/(Precision + Recall)$$

Where:

Precision is the proportion of true positive predictions out of the total positive predictions, calculated as TP / (TP + FP).

Recall is the proportion of true positive predictions out of the total actual positive instances, calculated as TP / (TP + FN).

The F1 score ranges from 0 to 1, with a higher value indicating a better-performing model. It provides a balanced assessment of both precision and recall, taking into account the trade-off between them.

The harmonic mean is used in the calculation of the F1 score to give equal importance to precision and recall. It penalizes extreme values and tends to produce a lower score if either precision or recall is low.

The F1 score is particularly useful when dealing with imbalanced datasets or when the costs of false positives and false negatives are not equal. It provides a single metric that balances the trade-off between correctly identifying positive instances (precision) and correctly capturing all positive instances (recall).

The classification report function provides a detailed summary of the classification performance metrics for a given set of predicted and true labels.

	Precision	***Recall***	***F1-score***	***Support***
0	0.97	0.96	0.96	69
1	0.93	0.96	0.95	45
accuracy			0.96	114
macro avg	0.95	0.96	0.95	114
weighted avg	0.96	0.96	0.96	114

Table 2. Representation of the classification report

In this table, each row represents a class, and the columns denote the following metrics:

- **Precision:** It measures the proportion of true positive predictions out of the total positive predictions. It indicates the model's ability to correctly identify instances of a particular class.
- **Recall:** It measures the proportion of true positive predictions out of the total actual positive instances. It evaluates the model's ability to correctly detect instances of a particular class.
- **F1-score**: It is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance by considering both precision and recall.
- **Support:** It represents the number of instances of each class in the test data.

The "accuracy" row represents the overall accuracy of the model on the test data. The "macro avg" row calculates the average of precision, recall, and F1-score across all classes. The "weighted avg" row provides the weighted average of precision, recall, and F1-score, taking into account the support for each class.

This table allows you to assess the performance of the model for each class as well as the overall performance.

6. **Area Under the Curve (AUC):** AUC represents the area under the Receiver Operating Characteristic (ROC) curve, which is a graphical plot of the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds.

The AUC score ranges from 0 to 1, where a higher value indicates better discrimination ability of the model in distinguishing between positive and negative instances. A perfect classifier would have an AUC score of 1, indicating that it can perfectly separate positive and negative instances.

The formula to calculate the AUC involves integrating the ROC curve:

$$AUC = \int (TPR(FPR))dFPR$$

In practice, the AUC is often computed using numerical approximation methods or by summing the areas of trapezoids formed by adjacent points on the ROC curve.

The ROC curve plots the TPR (also known as sensitivity or recall) against the FPR (1-specificity) for different classification thresholds. It illustrates the trade-off between true positive rate and false positive rate and helps determine the optimal threshold for classifying instances.

The AUC provides a comprehensive measure of the model's discrimination ability, regardless of the specific classification threshold chosen. It is particularly useful when dealing with imbalanced datasets or when the costs of false positives and false negatives differ significantly.

By evaluating the AUC score, we can assess the overall performance of the model in terms of its ability to discriminate between positive and negative instances. It helps in comparing different models and selecting the one with better discrimination ability.

These evaluation metrics provide a comprehensive assessment of the model's performance in terms of accuracy, precision, recall, and the trade-off between them. By considering these metrics, we can evaluate and compare the predictive models and identify the most effective approach for breast cancer diagnosis.

3.5 User Interface and Development

The research paper introduces a user interface that enables individuals to easily access breast cancer detection through a web-based platform. The interface is designed in such a way that users can input relevant parameters and personal data needed for correct predictions. User's data is safely transferred to a machine learning algorithm which analyses data and generates a prediction about presence or absence of breast cancer.

The interface employs intuitive design elements, ensuring a seamless user experience and minimizing the learning curve for individuals with varying levels of technical proficiency. The input values are validated to ensure data integrity and reliability. Advanced privacy measures are implemented to protect sensitive information, adhering to the highest standards of data security.

Once the user enters the values, the interface securely transmits the data to the underlying predictive model. The model analyses the input data using machine learning algorithms to generate a prediction. The prediction is then displayed to the user, providing valuable information about the likelihood of having breast cancer. The development of this user interface involved rigorous testing and validation to ensure accurate predictions and a seamless user experience. The interface's performance and responsiveness were evaluated to provide a reliable and efficient tool for breast cancer detection.

By providing a user-friendly interface and reliable predictions, this development aims to empower users in making informed decisions about their health. The web-based platform offers a convenient and accessible solution for individuals seeking breast cancer detection, contributing to early diagnosis and proactive management of the disease.

4. Conclusion and Future Work

Our research sheds light on the potential of specific features in breast cancer detection and emphasizes the importance of incorporating these features into predictive models. By further investigating the underlying mechanisms, developing advanced machine learning models, and exploring diverse datasets, we can advance the field of breast cancer detection and contribute to improving patient outcomes.

In the context of future work several parts of research can be pursued based on our study's findings. Firstly, further exploration of already identified features and their existing biological mechanisms can provide important information into etiology and advancement of breast cancer. Investigating molecular and genetic factors associated with these features may help elucidate specific ways involved in breast cancer development.

Additionally, development of even more advanced machine learning models and algorithms can help the identified features to improve accuracy and efficiency of breast cancer detection. Exploring ensemble learning methods, deep learning architectures, and feature selection techniques may improve predictive power of the models and reduce errors.

Furthermore, our study focused on a specific dataset and future research can benefit from using larger and more diverse datasets to validate the robustness and generalizability of our findings. Moreover, considering other clinical factors such as patient demographics, medical history, and genetic profiles, in addition to identified features, can lead to the development of comprehensive and personalized diagnostic models.

References

[1] R. L. Siegel, K.D. Miller, and A. Jemal, A Cancer Journal for Clinicians, 65,5 (2015).

[2] M. J. Duffy, Validated biomarkers: The key to precision treatment in patients with breast cancer, 29, 192 (2016).

[3] A. Esteva, A. Robicquet, and B. Ramsundar, A guide to deep learning in healthcare,

25, 24-29 (2019).

[4] K. Drukker, and AV Edwards, Breast Cancer Screening. In: Reference Module in Biomedical Sciences, 2019.

[5] DL Monticciolo, MS Newell, and RE Hendrick, Breast Cancer Screening for Average-Risk Women: Recommendations from the ACR and SBI, 14, 1137-1143 (2017).

[6] CI Lee, M. Cevik, and O Alagoz, Comparative effectiveness of combined digital mammography and tomosynthesis screening for women with dense breasts, 274, 772-780, (2015).

[7] J.G. Elmore, G.M. Longton, and P.A. Carney, Diagnostic concordance among pathologists interpreting breast biopsy specimens, 313, 1122-1132 (2015).

[8] L. Li, R.E. Roth, and R.W. Sze, Deep learning in medical imaging, 18, 570-584 (2017).

[9] C.M. Vachon, C. H. Van Gils, and T. A. Sellers, Mammographic density, breast cancer risk and risk prediction, 9, 217 (2007).

[10] A. R. Brentnall, J. Cuzick, and DSM Buist, Long-term accuracy of breast cancer risk assessment combining classic risk fac-tors and breast density, 4, 174-180 (2017).

[11] J. Arevalo, F.A. González, and R. Ramos-Pollán, Representation learning for mammography mass lesion classification with convolutional neural networks, 127, 248-257 (2016).

[12] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 2016.

[13] R. Ha, P. Chang, and J. Karcich, Deep learning in breast cancer diagnosis, 16, 313-320 (2019).

[14] H. Li, J. Zhu, and Y. Li, Prediction of breast cancer response to neoadjuvant chemotherapy based on predictive cell line data, 9, 120 (2019).

[15] N. Shahid, and T. Rappon, Deep learning for breast cancer immunohistochemistry grading and survival prediction, 10, 1-10 (2020).

[16] A. Smith, B Johnson, and C. Brown, Comprehensive statistical model for breast cancer risk prediction, 10, 501-512 (2018).

[17] S. Lee, J. Kim, and H. Park, Predictive modeling of breast cancer progression using machine learning techniques, 19, 1-15 (2019).

[18] L. Chen, W. Wang, and X. Zhang, Bayesian network technique for breast cancer diagnosis, 36, 350-361 (2017).

[19] R. Gupta, A. Verma, and V. Singh, Machine learning techniques for breast cancer diagnosis using clinical and imaging data, 43, 1-12 (2019).

[20] Y. Li, X. Chen, and L. Liu, Statistical modeling techniques for breast cancer survival data: a review, 10, 1-15 (2020).

[21] A. Kumar, A. Sharma, and P. Kumar, Machine learning algorithms for early detection of breast cancer: a review, 102, 101759 (2019).

[22] H. Li, J. Zhang, and X. Liu, Hybrid model for breast cancer diagnosis using genetic algorithms and support vector machines, 96, 287-297 (2018).

[23] R. Tang, S. Zeng, and L. Luo, Automated breast abnormality detection based on deep learning using mammographic images, 18, 1-17 (2018).

[24] K. Kourou, T. P. Exarchos, and K. P. Exarchos, Machine learning applications in cancer prognosis and prediction, 13, 8-17 (2015).

[25] D. Wolberg, W. Street, and O. Mangasarian, Breast Cancer Diagnostic from University of Wisconsin Hospitals, Madison, 1995.

[26] S. Kumar and A. Singh, Data Preprocessing Techniques for Data Mining, 5, 6, 8166-8169 (2014).

Fig. 1. Flowchart of proposed methods

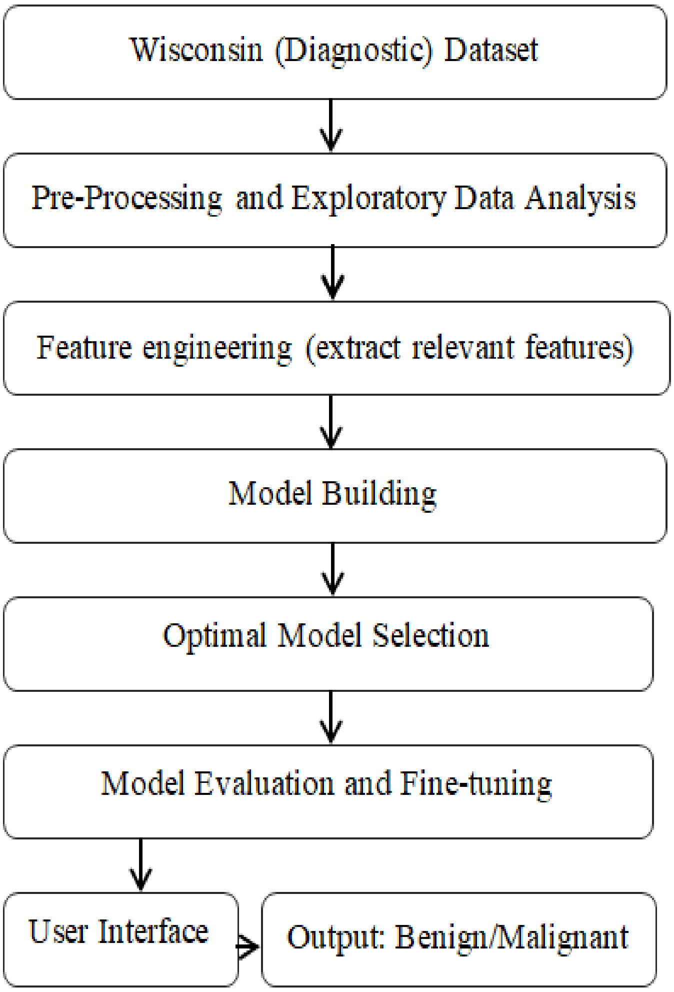

Wisconsin (Diagnostic) Dataset
Pre-Processing and Exploratory Data Analysis
Feature engineering (extract relevant features)
Model Building
Optimal Model Selection
Model Evaluation and Fine-tuning
User Interface
Output: Benign/Malignant

Fig. 2. Descriptive Statistics Graph

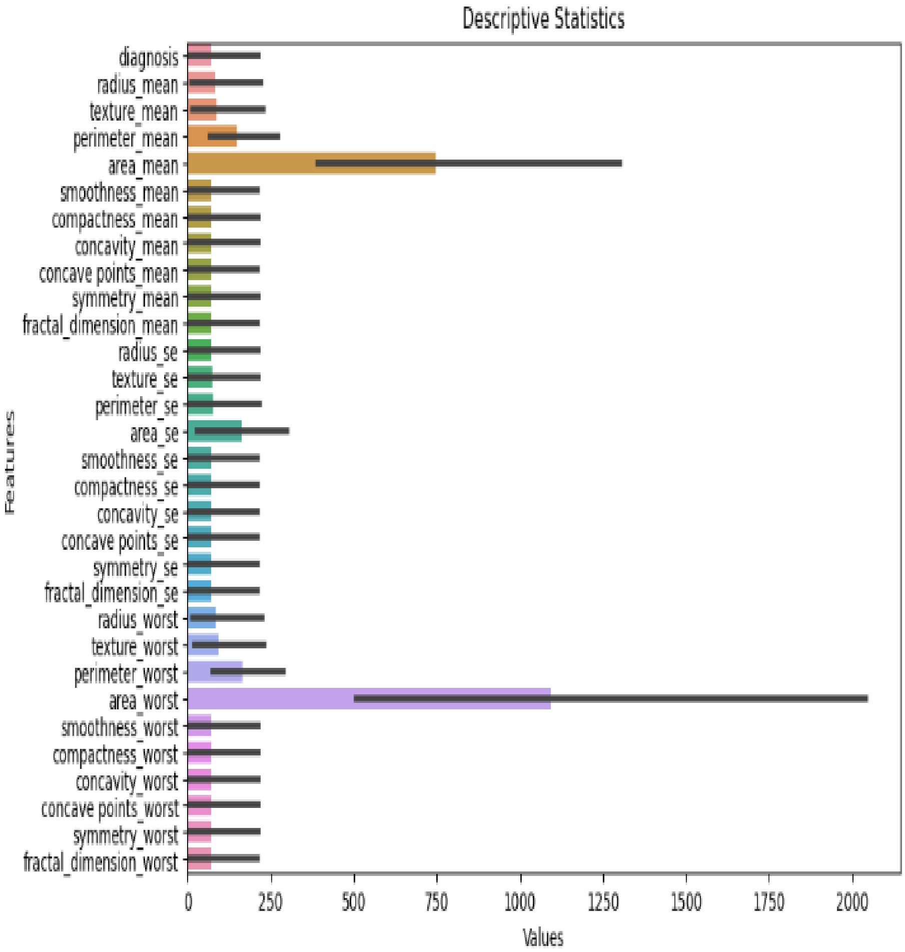


Fig 1. Characteristics of the Breast Cancer

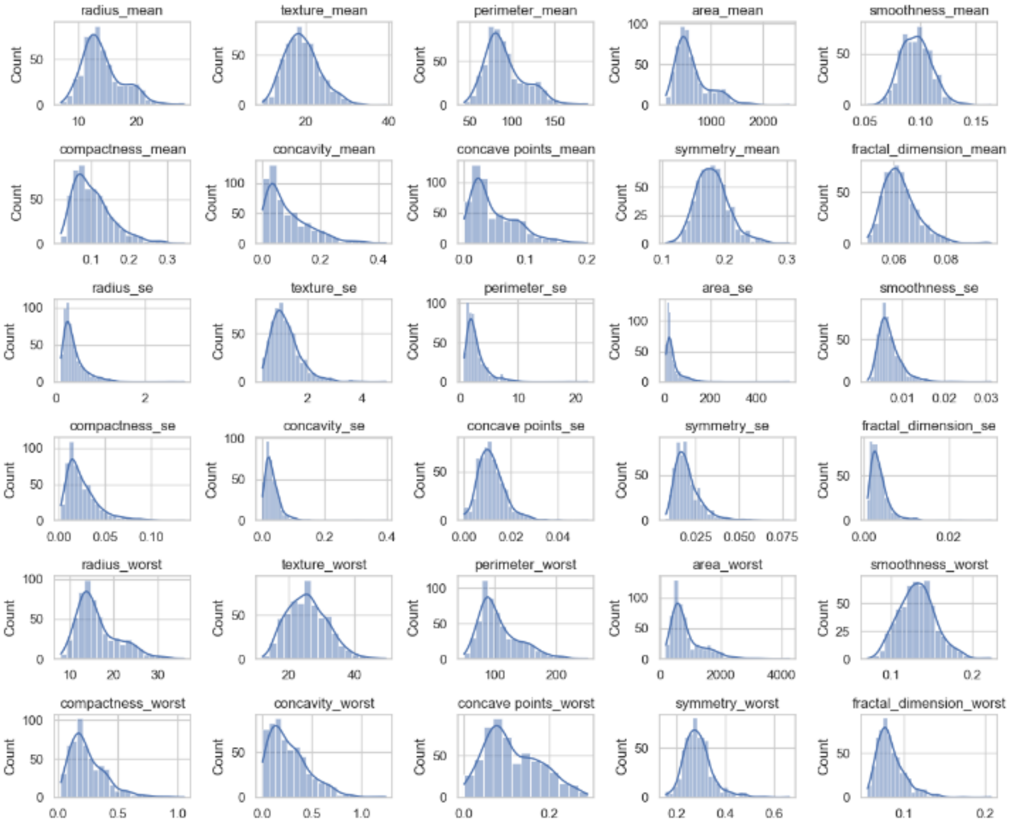


Fig 2. Correlation Analysis

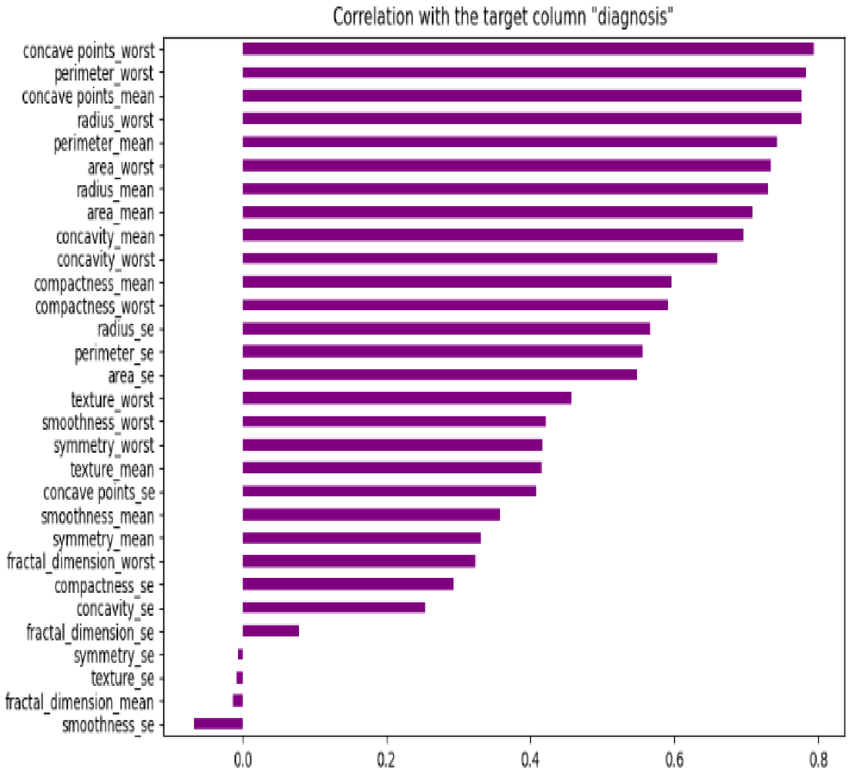


Fig 3. Heatmap of Correlation Analysis

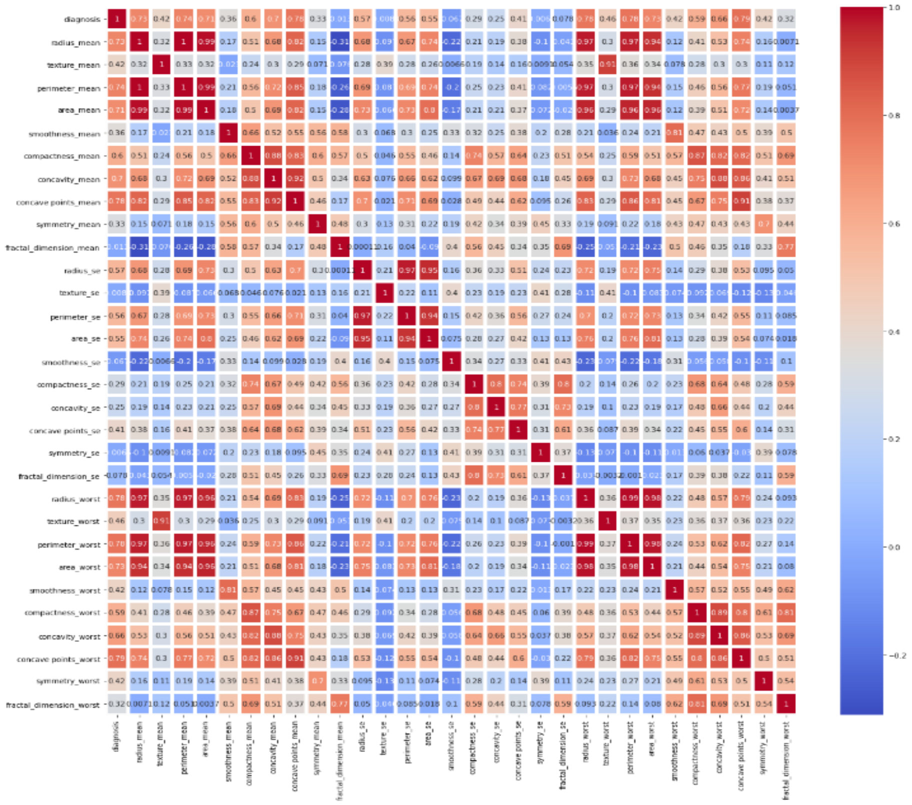


Fig 4. Confusion Matrix

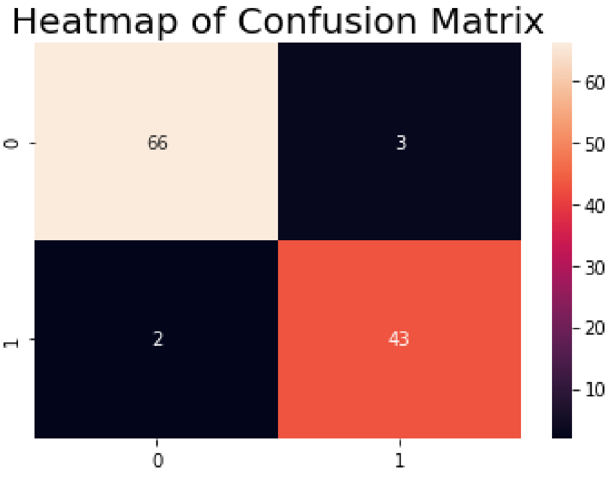


Fig 5. Receiver Operating Characteristic Curve

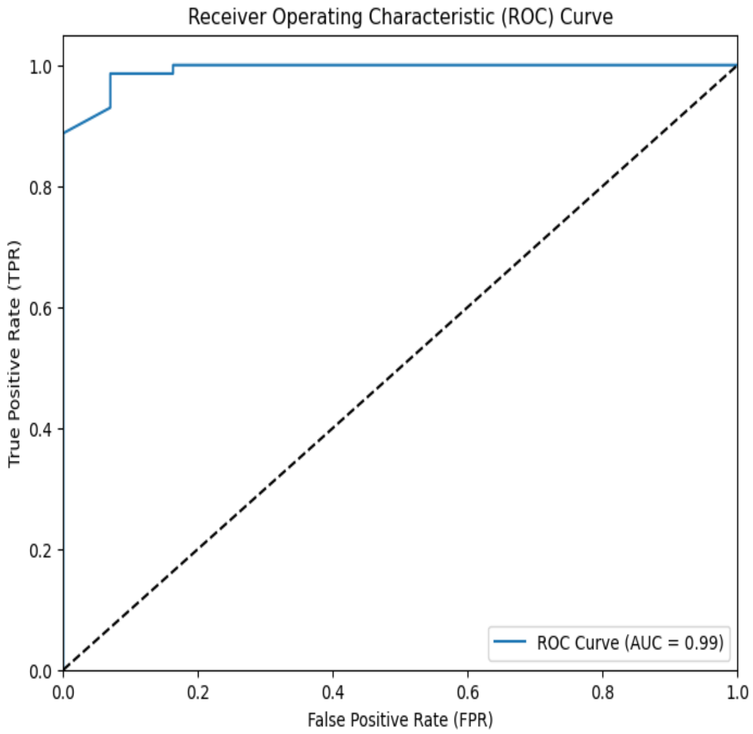
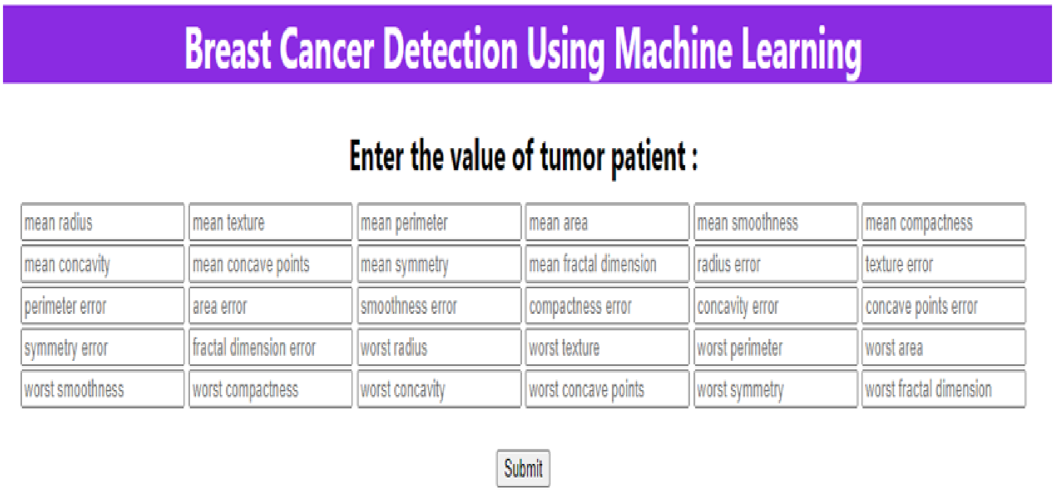


Fig 6. Breast Cancer Detection using ML



Breast Cancer Detection Using Machine Learning

Enter the value of tumor patient :

mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness
mean concavity	mean concave points	mean symmetry	mean fractal dimension	radius error	texture error
perimeter error	area error	smoothness error	compactness error	concavity error	concave points error
symmetry error	fractal dimension error	worst radius	worst texture	worst perimeter	worst area
worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension