ENSET

Université HASSAN II DE Casablanca

Ecole Normale Supérieure de l'Enseignement Technique (ENSET)

Machine Learning

Master : Systèmes Distribués et IntelligenceArtificielle

Project : BREAST CANCER DETECTION

Réalisé par : TAMOUCHE Kaoutar
ZAIM Ibrahim

Encadré par : Pr. HAMIDA Soufiane

Année Universitaire : 2023/2024

Figures List :

Content :

1 CHAPTER1: INTRODUCTION TO BREAST CANCER

1.1 Definition of Breast Cancer

Breast cancer arises from the transformation of normal breast tissue into malignant tissue, where cells grow and divide without the usual control mechanisms. This can result in the development of a tumor that may be felt as a lump or detected through imaging. Tumors in the breast can be categorized into two main types: benign and malignant.

Benign Tumors

Benign tumors are non-cancerous growths in the breast. These tumors are typically not life-threatening and do not spread to other parts of the body. However, they can increase the risk of breast cancer and may require surgery to prevent further complications.

Malignant Tumors

Malignant tumors are cancerous and can invade surrounding tissues or spread (metastasize) to other areas of the body. Breast cancer is primarily known for these malignant tumors, which pose significant health risks and require prompt and often aggressive treatment.

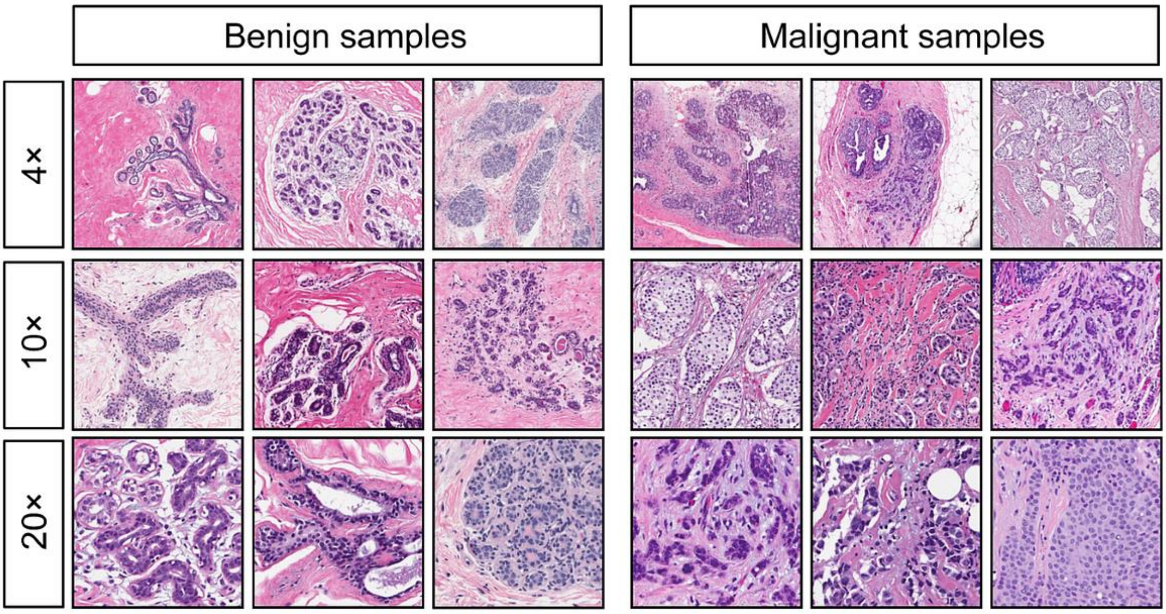


Figure 1: Tumour samples.

1.2 Importance of Early Detection

Catching breast cancer early is crucial for effectively managing the disease. It greatly affects the treatment options available, how long patients survive, and their quality of life after treatment. Detecting breast cancer early improves the outcomes for patients.

Benefits of Early Detection

- Better Survival Chances: Statistics show that finding breast cancer early, when it's still only in the breast, greatly increases the chances of surviving for five years or more, much more so than if it's found later.
- More Treatment Choices: Detecting cancer early usually means patients have more and better treatment options, including less severe surgeries and the possibility of keeping their breasts.
- Lower Healthcare Costs: It's cheaper to treat breast cancer if it's found early because later stages of cancer are more expensive to manage.
- Better Quality of Life: People diagnosed in the early stages of breast cancer generally feel better overall after their treatment, with less impact on their physical and emotional health.

Challenges to Early Detection

- Accessibility: Not everyone has equal access to early detection methods, often due to differences in social and economic status.
- Awareness: More education is needed about the importance of regular breast cancer screenings.
- Compliance: It's important to make sure that people actually go for screenings as recommended.

Relevance to Our Project

Our project seeks to improve the early detection of breast cancer by using machine learning to analyze samples of breast tissue. By creating a model that can accurately identify if breast tumors are benign (non-cancerous) or malignant (cancerous), we aim to support and enhance existing screening efforts.

1.3 Role of Machine Learning in Early Detection

Machine learning is a type of technology that allows computers to learn from examples and experiences without being directly programmed. In the field of breast cancer, machine learning is increasingly important because it helps doctors detect the disease earlier than ever before. Here’s how machine learning is making a difference:

- Improving Accuracy: Machine learning models can analyze large amounts of medical data, like mammograms or ultrasound images, quickly and with high accuracy. They can detect subtle patterns that human eyes might miss, which helps in identifying breast cancer early.
- Consistency in Diagnosis: Unlike humans, who might get tired or overlook details, a machine learning system always performs consistently. This consistency is crucial for reliable breast cancer screenings.
- Handling Complex Data: Breast cancer diagnosis can involve complex data from various tests and screenings. Machine learning can handle this complexity by

integrating and analyzing all the information to make accurate predictions or recommendations.

- Speed: Machine learning can process data and make decisions much faster than human clinicians. This speed means that more screenings can be conducted in less time, potentially leading to earlier diagnoses for more women.
- Personalized Assessments: Machine learning algorithms can also help tailor screening and treatment plans to individual patients based on their unique data. This personalized approach helps in catching cancer early in patients who might develop symptoms differently.

Supporting Early Detection Efforts

Our project applies machine learning to develop a tool that helps clinicians identify whether a tumor is benign or malignant. By using machine learning to analyze tissue samples, our tool aims to provide quick, accurate assessments. This is facilitated by our workflow that integrates feature extraction, selection, and classifier training to ensure the model's effectiveness. This can make screenings more efficient and ultimately help catch breast cancer at an earlier stage, when it is easier to treat.

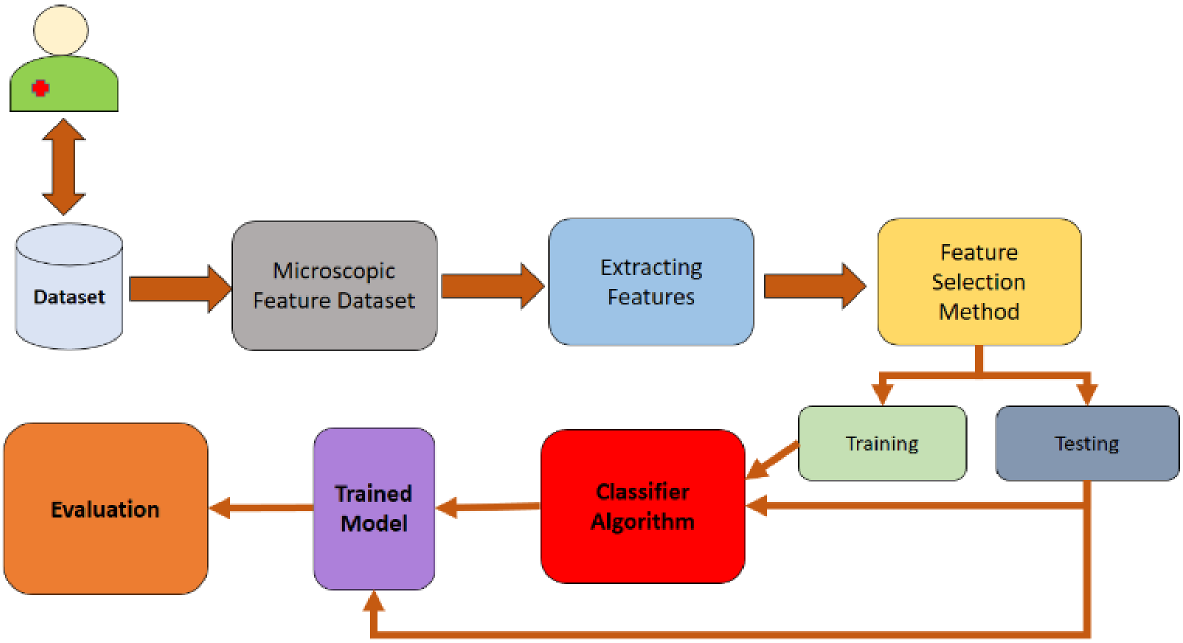


Figure 2: Workflow for our machine learning model.

1.4 Project Goals and Scope

Project Goals

The primary goal of our project is to enhance the early detection of breast cancer through the use of machine learning. By developing a sophisticated model that can accurately classify breast tumors as benign or malignant, we aim to support clinicians in making quicker and more accurate diagnoses. The specific objectives of the project are:

- To Develop an Accurate Model: Create a machine learning model that can reliably distinguish between benign and malignant breast tumors using data from the Breast Cancer Wisconsin (Diagnostic) dataset.
- To Improve Diagnostic Speed: Reduce the time it takes to diagnose breast cancer by automating part of the analysis process, allowing for quicker decision-making in clinical settings.
- To Increase Diagnostic Consistency: Provide a tool that offers consistent assessments, helping to minimize the variability that can occur with human analysis.
- To Facilitate Early Detection: By improving accuracy and speed, the project aims to detect cancer at an earlier stage, which is crucial for successful treatment and improved patient outcomes.

Project Scope

The scope of this project includes the following components:

- Data Preparation and Analysis: We will preprocess and analyze the Breast Cancer Wisconsin (Diagnostic) dataset, focusing on extracting meaningful features that contribute to accurate tumor classification.
- Model Development and Validation: Multiple machine learning models will be developed and evaluated to find the most effective approach. This includes training, testing, and refining the models to achieve the best performance.
- Tool Development: The final model will be integrated into a user-friendly web-based tool developed using Streamlit. This tool will allow clinicians and researchers to input data and receive immediate predictions.
- Deployment and Testing: The tool will be deployed to a cloud platform, ensuring it is accessible to users with varying levels of technical expertise. Comprehensive testing will be conducted to ensure the tool's reliability and usability in real-world settings.

Limitations

While the project aims to address key challenges in breast cancer detection, there are limitations to consider:

- Data Limitations: The performance of the machine learning model is highly dependent on the quality and variety of data available in the Breast Cancer Wisconsin (Diagnostic) dataset.
- Scope of Technology: The model and tool are intended as aids for clinical decision-making, not as replacements for human expertise. It is crucial that they are used in conjunction with professional medical judgment.
- Accessibility and Adoption: The actual deployment and widespread adoption of the developed tool may face challenges related to technological access and resistance to new methods in some clinical environments.

With clear objectives set and the scope defined, the next section will delve deeper into the dataset that forms the backbone of our machine learning project, explaining its composition, features, and the preprocessing steps undertaken to prepare it for effective model building.

2 CHAPTER 2 : DATASET OVERVIEW AND PREPROCESSING

2.1 Data Source

We are using the Breast Cancer Wisconsin (Diagnostic) dataset for our project. This dataset was created by Dr. William H. Wolberg and his team at the University of Wisconsin Hospitals in Madison. It includes data collected from 569 patients using a procedure called fine needle aspirate (FNA), where cells are extracted from the breast to examine under a microscope. From these samples, the dataset calculates 30 different features related to the cells' appearance, such as their size and shape. These features help to identify whether a tumor is benign (not cancerous) or malignant (cancerous). The dataset is known for being reliable and is widely used in medical research to help improve breast cancer detection. It's available for free from the UCI Machine Learning Repository, making it accessible for educational and research purposes.

2.2 Feature Description

The Breast Cancer Wisconsin (Diagnostic) dataset, commonly known as the WDBC dataset, is a popular dataset used in the field of machine learning for predicting breast cancer based on attributes derived from digitized images of a fine needle aspirate (FNA) of a breast mass. Here's a feature description for this dataset:

ID Number: Unique identifier for each sample.

Diagnosis: The classification of the observed cells. Possible values are:

- M (malignant)
- B (benign)

Features 3-32 are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image, divided into three groups: mean, standard error, and "worst" or largest (mean of the three largest values), resulting in ten real-valued features for each group:

Radius (mean of distances from center to points on the perimeter): Mean of the distances from the center of the nucleus to its perimeter.

Texture (standard deviation of gray-scale values): Measure of the variation in gray values in the cell nucleus.

Perimeter

Area

Smoothness (local variation in radius lengths): A measure of the smoothness of the contour of the nucleus.

Compactness (perimeter^2 / area - 1.0): Describes how closely the area and perimeter are related.

Concavity (severity of concave portions of the contour): Measures the severity of concave portions of the boundary of the nucleus.

Concave points (number of concave portions of the contour)

Symmetry

Fractal dimension ("coastline approximation" - 1): Measures the complexity of the shape of the nucleus.

Each of the above measurements is followed by three more features that provide:

Mean: The average of these measurements for each image.

Standard Error: The standard error for these measurements.

Worst: The worst or largest mean found among three largest measurements of these features for each image.

This structured presentation of features helps in creating robust predictive models for diagnosing breast cancer from FNA samples.

2.3 Data Pre-processing

In the crucial phase of preparing our data for analysis, we executed several steps to ensure that the machine learning models we intended to build would function optimally:

Dataset Importation: We initiated the process by loading the dataset using pandas, a powerful Python data analysis tool. This step is akin to gathering all necessary ingredients before starting to cook a complex dish.

Dataset Cleanup: Our first task was to clean the data. We noticed an unnecessary column named 'Unnamed: 32', filled with missing values, which we promptly removed. This helped streamline our dataset by eliminating irrelevant data that could potentially confuse our models.

Handling Missing Data: We performed a thorough search for any missing data across the dataset. In machine learning, missing values can skew results, so it's critical to identify and address them before proceeding.

Dropping the Identifier Column: The 'id' column, which serves as a unique identifier for each record, was removed since it does not contribute to the model's ability to learn and differentiate between cancer types.

Class Distribution Check: We examined the distribution of diagnoses to understand the balance between benign and malignant cases, ensuring our model would be trained on a representative sample of data.

Encoding the Target Variable: The 'diagnosis' column, which contains the labels 'M' for malignant and 'B' for benign, was encoded into a numeric format. Machine learning algorithms require numerical input, so this encoding converts categorical labels into a binary format that the model can interpret.

Feature Scaling: We used StandardScaler, a method that normalizes the range of our features. This is a vital step since features on different scales can disproportionately influence the model.

Data Splitting: Finally, we split our data into two parts: a 'training' set to teach our model and a 'test' set to evaluate how well it learned. A common split ratio is 70% for training and 30% for testing, which we adhered to.

By the end of these preprocessing steps, we ensured that our dataset was clean, relevant features were appropriately scaled, and the data was ready to be fed into machine learning algorithms for model building.

2.4 Visualization Techniques

In the context of breast cancer detection, visualizing the dataset can serve multiple important functions:

- Feature Distribution: Visualizing the distribution of features allows us to see how the data points are spread across different variables. This is particularly important in medical datasets like the Breast Cancer Wisconsin (Diagnostic) dataset, where understanding the variance and behavior of biological features can provide insights into their relevance for classification tasks.
- Identifying Relationships: Scatter plots and correlation matrices help identify relationships between features. In breast cancer datasets, correlations can indicate how different cellular characteristics may be linked to the presence of benign or malignant tumors, guiding the feature selection for machine learning models.
- Illustrating Class Distribution: Visualizations that show the balance between different classes (benign vs. malignant) are essential. They help in understanding the dataset's skewness, which is critical for developing models that are sensitive to less represented classes, thereby improving their diagnostic efficacy.

2.4.1 Histograms

Histograms are used to show the frequency distribution of individual features, helping identify patterns or skewness in data.

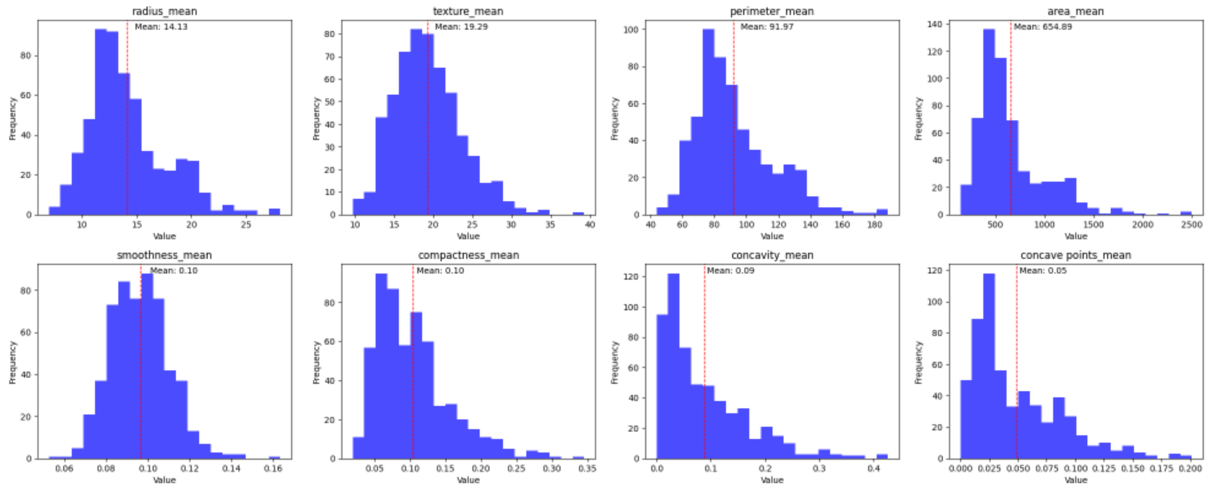


Figure 3: Histograms part 1.

The radius_mean histogram exhibits a moderately right-skewed distribution, indicating the presence of tumors with larger radii. The area_mean histogram is also notably right-skewed, suggesting a subset of tumors with larger areas that may indicate more advanced disease states. Texture_mean and perimeter_mean appear more normally distributed, signifying consistent patterns of these features across the cell nuclei samples. The histograms for smoothness_mean, compactness_mean, concavity_mean, and concave points_mean display roughly normal distributions with slight right skewness, providing insights into the physical characteristics of tumors at a cellular level.

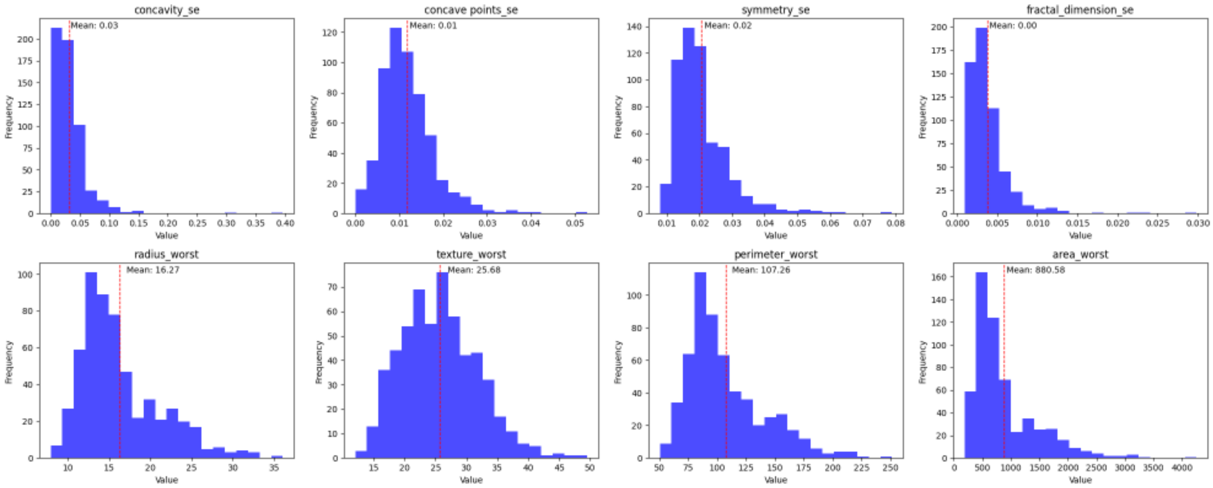


Figure 4 : Histograms part 2.

In Figure 2, we examine the precision of measurements through the standard error histograms for concavity, concave points, symmetry, and fractal dimension. The concavity_se and concave points_se histograms show a significant right skew, indicating variability in these measurements across samples. Symmetry_se and fractal dimension_se have a more uniform distribution, reflecting consistent measurement precision. The 'worst' values, representing the mean of the three largest values for each feature, are shown for radius, texture, perimeter, and area. These histograms—radius_worst, texture_worst, and perimeter_worst—demonstrate normal distributions with right skewness, while area_worst shows a pronounced skew, highlighting the potential for more aggressive tumor behavior.

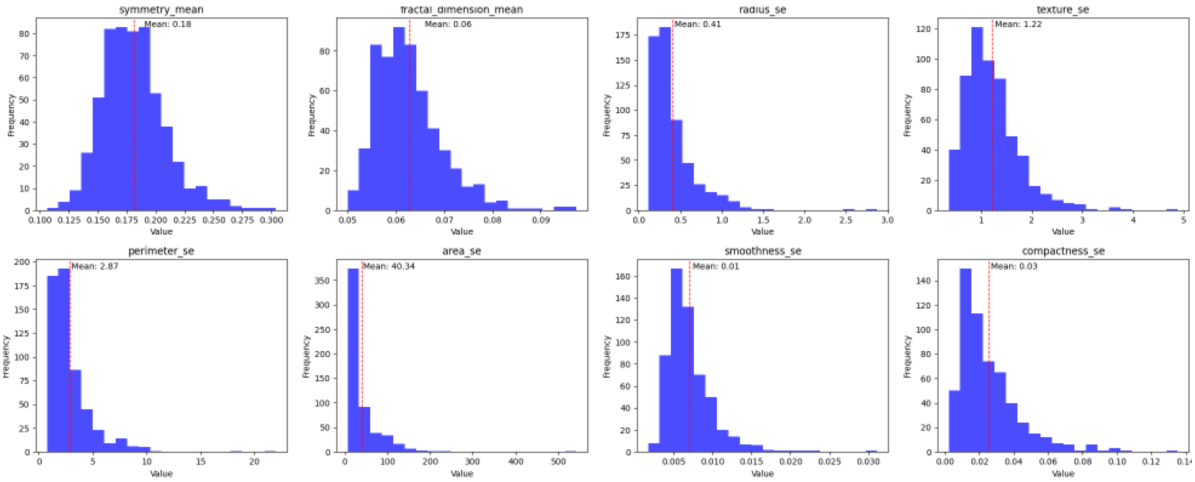


Figure 5: Histograms part 3.

The mean values for symmetry and fractal dimension are explored in Figure 3, with symmetry_mean showing a slightly right-skewed normal distribution, whereas fractal dimension_mean exhibits a more pronounced right skew. The standard error histograms for radius_se and texture_se reveal substantial right skewness, denoting a greater variability. Conversely, perimeter_se and area_se reflect a similar distribution pattern, essential for understanding the measurement precision of tumor shape and size. Finally, the smoothness_se and compactness_se histograms indicate a less varied distribution, suggesting these features have relatively consistent measurement errors across the dataset.

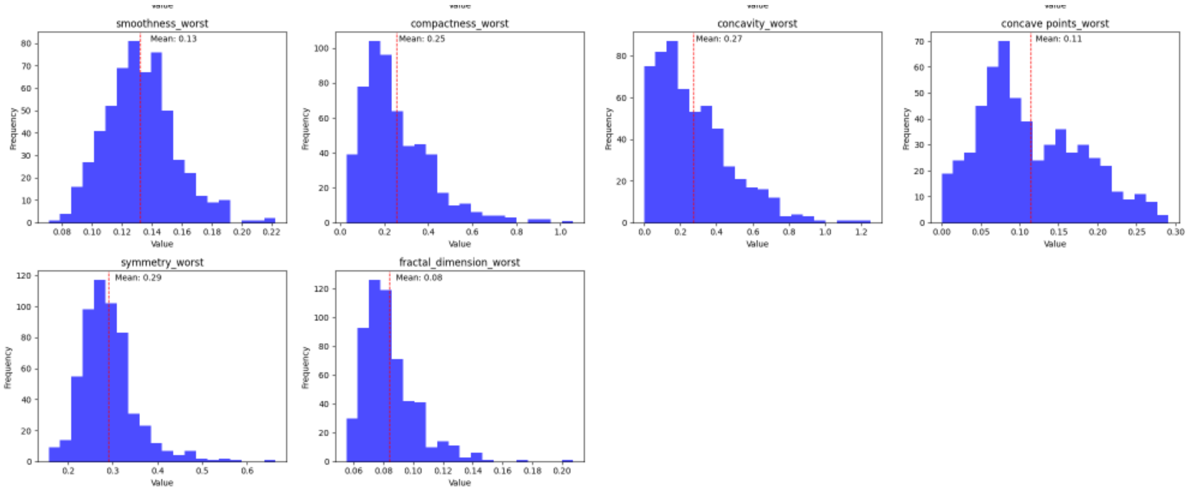


Figure 6: Histograms part 4.

Figure 4 showcases the distribution of the 'worst' values for smoothness, compactness, concavity, and concave points, as well as for symmetry and fractal dimension. These represent the most extreme observed values and help us understand the severity of cellular characteristics. For example, the smoothness_worst histogram, depicting a normal distribution with a slight right skew, could be indicative of more aggressive tumor types. Compactness_worst, concavity_worst, and concave points_worst all show significant right skewness, pointing to pronounced physical deformities in some tumors, which are critical markers for malignancy. The variability in the extreme values of symmetry_worst and fractal dimension_worst might also correspond to the aggressiveness of tumors.

The patterns revealed by these histograms are critical for our machine learning project. The observed skewness and variability will inform our feature engineering and data preprocessing strategies, ensuring that our machine learning models are robust and capable of making accurate predictions.

2.4.2 Box Plots

Box plots offer a concise visualization of the distribution of data, clearly delineating the central tendency, dispersion, and outliers.

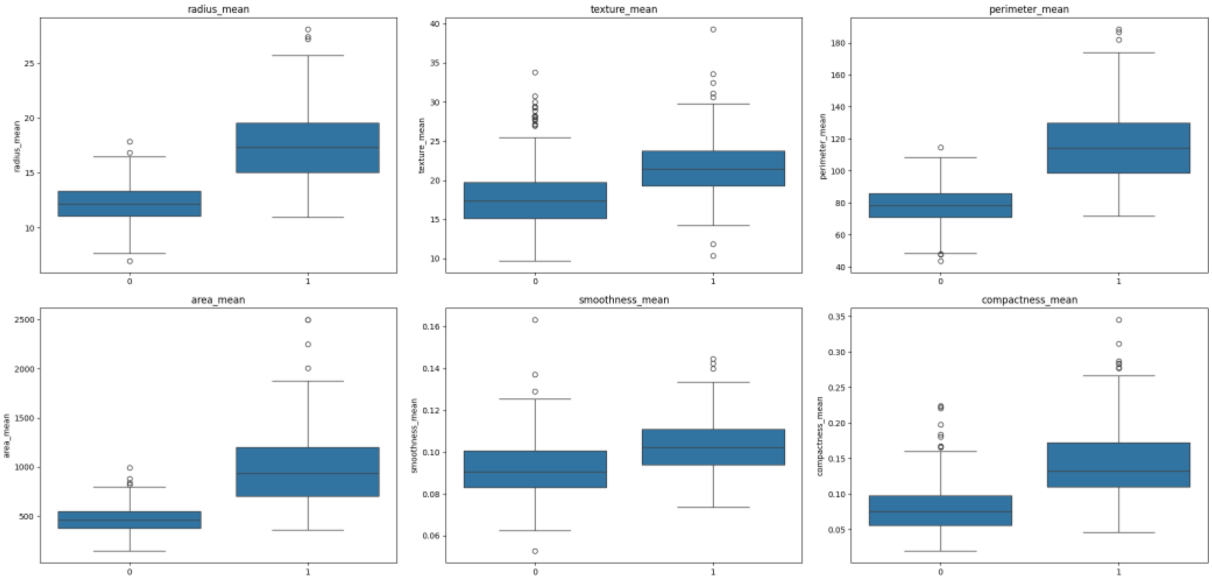


Figure 7: Box Plots -Part 1.

Box plots for radius_mean and area_mean show that malignant tumors tend to have higher medians and a broader distribution, which may reflect more advanced disease states. The plots for texture_mean and perimeter_mean also depict a wider spread in malignant cases, suggesting these features' potential association with tumor malignancy. While smoothness_mean indicates a slight median increase in malignancy, it is compactness_mean that notably presents a more pronounced median elevation and several outliers in malignant cases, which could serve as a potential indicator of malignancy due to increased cell density.

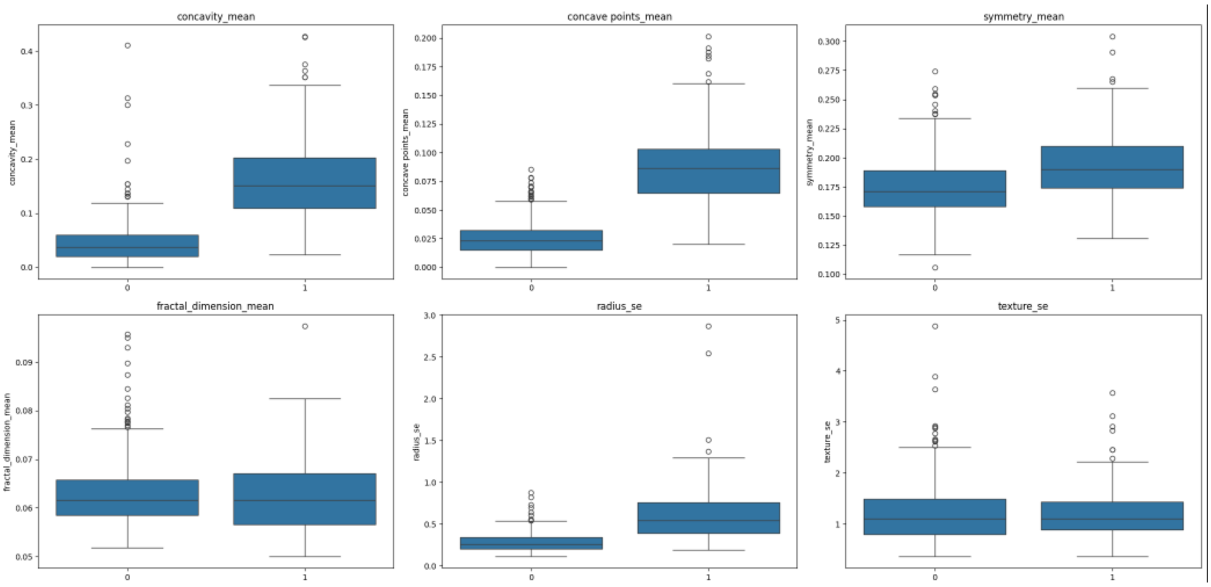


Figure 8: Box plots-Part2.

Concavity_mean and concave points_mean reveal significantly higher medians in malignant tumors, indicating that these features are typically more pronounced in malignant states and might be crucial for diagnosis. Symmetry_mean and fractal_dimension_mean exhibit less differentiation between benign and malignant tumors, but a higher median and several outliers in benign tumors for fractal_dimension_mean suggest that these features could play a role as secondary markers.

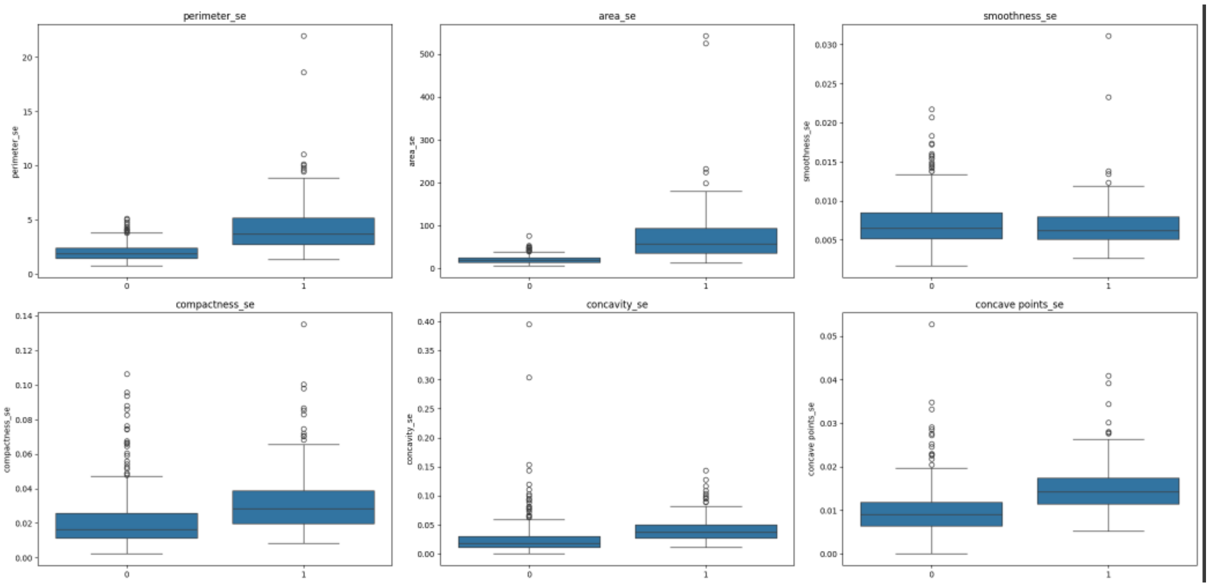


Figure 9: Box plots-Part3.

The box plots for perimeter_se and area_se underscore that malignant tumors tend to exhibit higher measurement variability, aligning with the understanding that malignant tumors often have more complex shapes. Smoothness_se shows a relatively narrow spread of values, suggesting consistency in smoothness variation, while compactness_se indicates a higher median and notable outliers for malignant tumors, reflecting a potential aggressive tumor behavior. Concavity_se and concave points_se emphasize greater irregularity and structural concavities in malignant tumors.

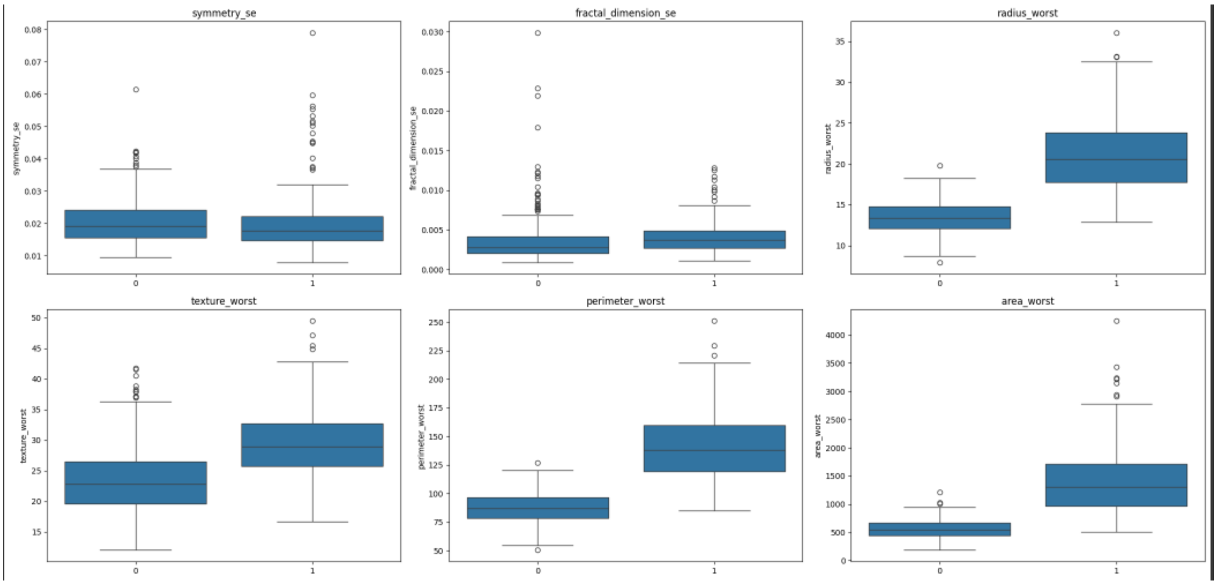


Figure 10: Box plots-Part4.

The symmetry_se and fractal_dimension_se box plots demonstrate variability within tumor symmetry and fractal dimension, with the presence of outliers in malignant tumors possibly indicative of malignancy. The lower row plots (radius_worst, texture_worst, perimeter_worst, and area_worst) show an increase in median values for malignancy, pointing to larger and more irregular tumor structures, particularly evident in the area_worst plot with its significant outliers.

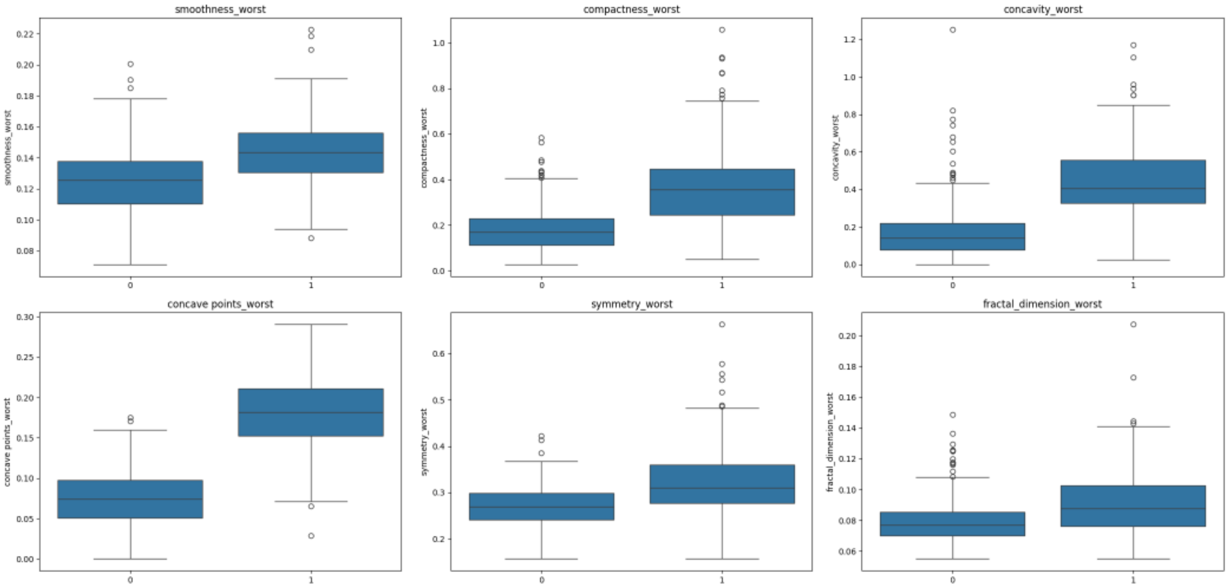


Figure 11: Box plots-Part5.

Smoothness_worst and compactness_worst exhibit higher medians and greater spreads in malignant tumors, consistent with the presence of more aggressive tumor characteristics. Concavity_worst and concave points_worst show a substantial median difference between benign and malignant tumors, reinforcing their importance in indicating malignancy. The symmetry_worst plot displays a slight median increase for malignant tumors, while the fractal_dimension_worst plot's marginal median difference, combined with outliers, might indicate an increased complexity associated with malignant tumor borders.

2.4.3 Correlation Matrix

Correlation matrices reveal the degree to which variables move in relation to one another, which is indispensable for feature selection and understanding potential multicollinearity in the dataset.

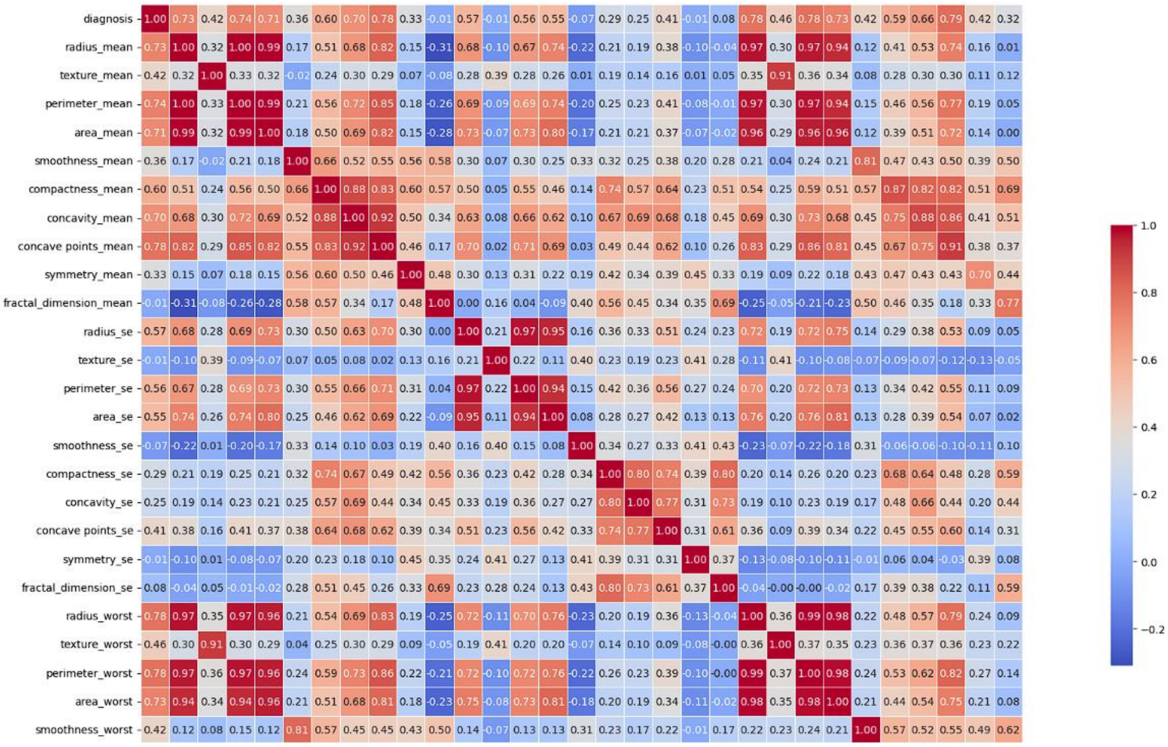
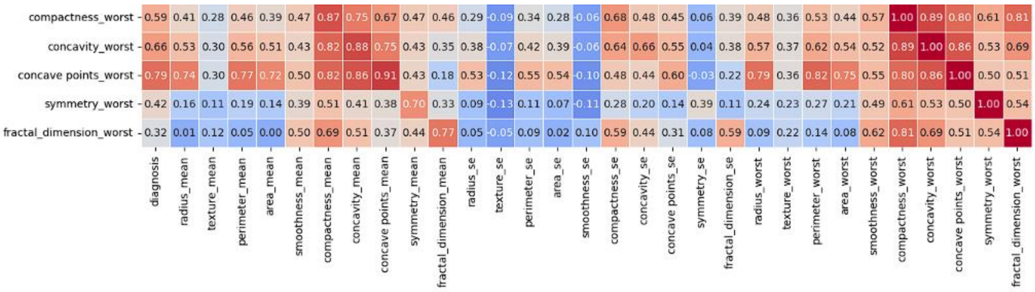


Figure 12:Correlation matrix.



The correlation analysis reveals significant relationships between individual features and their associations with the diagnosis of breast cancer. Positive correlations suggest that larger values of certain features, such as radius_mean, perimeter_mean, and area_mean, are strongly associated with malignant tumors, while negative correlations indicate lower values may be indicative of benign cases. Additionally, inter-feature correlations highlight clusters of strong relationships among various features. For instance, perimeter_mean exhibits high correlations with radius_mean and area_mean, emphasizing their close associations (Correlation values: 0.998, 0.987 respectively). Similarly, concave points_mean demonstrates strong correlations with radius_mean, perimeter_mean, and area_mean, indicating shared underlying characteristics (Correlation values: 0.823, 0.851, 0.823 respectively).

2.5 Dataset Challenges and Solutions

Challenges

- High Dimensionality: Our dataset’s 30 features present a challenge in machine learning known as the curse of dimensionality. This condition can lead to overfitting, where the model learns the training data too well, including its noise and outliers, reducing its ability to perform well on unseen data. Additionally, with so many features, understanding which ones are most important for the diagnosis becomes complex, and computational resources can be strained.
- Limited Data Size: With 569 instances, the dataset is relatively small compared to the number of features, intensifying the risk of overfitting and making it challenging to train complex models.
- Sensitivity to Outliers: Medical datasets often contain outliers due to individual differences in disease manifestation. These outliers can heavily influence machine learning models, especially in high-dimensional spaces.

Solutions

- Clinical Relevance Over PCA: While PCA is a common technique for reducing dimensions, it was important to us that the features remain interpretable in a clinical setting. Logistic Regression allows us to keep all features and still derive meaningful insights because it handles high-dimensional data efficiently without the need for dimensionality reduction.
- Careful Preprocessing: To address potential overfitting due to a high number of features relative to the sample size, we applied thorough data preprocessing. This included outlier detection and feature scaling, which are essential for logistic regression to perform optimally.

Conclusion

In conclusion, while the high number of features in the Breast Cancer Wisconsin (Diagnostic) dataset posed a challenge, the robustness of logistic regression against overfitting made them a suitable choice for our project. By prioritizing medical relevance and leveraging logistic regression’s strengths, we developed a model that can assist healthcare professionals in diagnosing breast cancer without sacrificing interpretability for accuracy.

3 CHAPTER 3 : MODEL DEVELOPMENT AND EVALUATION

3.1 Model Selection

3.1.1 Logistic Regression

Despite its simplicity, logistic regression can be surprisingly effective, especially when the decision boundary is linear or close to linear. It's interpretable, which could be beneficial for medical applications where understanding the model's decision-making process is essential.

Logistic regression is particularly well-suited for this task due to its simplicity, interpretability, and efficiency. It provides probabilistic predictions, making it easy to interpret the likelihood of a tumor being malignant. Additionally, logistic regression performs well with linearly separable data and is less prone to overfitting when the number of features is relatively small compared to the number of samples.

3.1.2 SVMs

SVMs are known for their effectiveness in handling high-dimensional data and are suitable for binary classification tasks like cancer detection. They can handle both linear and non-linear data and have been successfully applied in medical diagnosis tasks.

SVMs are well-suited for breast cancer detection tasks due to their ability to handle high-dimensional data and nonlinear relationships between features. They are effective in finding complex decision boundaries, which may be necessary for accurately distinguishing between malignant and benign tumors. Additionally, SVMs are robust against overfitting, especially when using a linear kernel with proper regularization.

Both logistic regression and SVMs offer promising solutions for breast cancer detection. While logistic regression is simple and interpretable, SVMs provide flexibility in modeling complex relationships and can handle high-dimensional data efficiently. In the subsequent sections, we will delve into the training, evaluation, and comparison of these models to determine the most suitable approach for our dataset.

3.1.3 Random Forests

Random Forest is a versatile ensemble learning algorithm that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or the mean prediction (regression) of individual trees. It combines the concepts of bagging and feature randomization to build a robust and accurate model.

Random Forests are particularly advantageous for breast cancer detection due to their ability to handle high-dimensional data and nonlinear relationships between features. They excel in capturing complex interactions and dependencies among features, making them effective in distinguishing between malignant and benign tumors. Moreover, Random Forests are less susceptible to overfitting compared to individual decision trees, as they aggregate the predictions of multiple trees, thereby reducing variance and improving generalization.

In the context of breast cancer detection, Random Forests offer several benefits:

Robustness: Random Forests are less sensitive to noise and outliers compared to single decision trees, making them suitable for datasets with irregularities.

Feature Importance: Random Forests provide a measure of feature importance, allowing for the identification of the most relevant features contributing to the classification task.

Nonlinear Relationships: Random Forests can capture nonlinear relationships between features and the target variable, enabling them to model complex patterns in the data effectively.

3.1.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple yet powerful non-parametric algorithm used for both classification and regression tasks. It classifies a data point based on the majority class of its K nearest neighbors in the feature space, where K is a predefined hyperparameter.

KNN is well-suited for breast cancer detection tasks due to its simplicity and intuitive approach. It does not make any assumptions about the underlying data distribution, making it suitable for nonlinear and complex relationships between features. Additionally, KNN can adapt to changes in the dataset structure, making it robust and versatile.

In the context of breast cancer detection, KNN offers several advantages:

Flexibility: KNN can handle both linear and nonlinear decision boundaries, allowing it to capture complex patterns in the data.

Interpretability: KNN's decision-making process is transparent, as it directly reflects the majority class of the nearest neighbors. This interpretability can be valuable in medical applications where understanding the model's rationale is essential.

Robustness: KNN is robust to outliers and noise in the data, as it relies on local information from neighboring data points rather than global assumptions about the data distribution.

Overall, Random Forests and KNN offer complementary approaches to breast cancer detection, each with its own strengths and limitations. In the subsequent sections, we will explore the training, evaluation, and comparison of these models alongside logistic regression and SVMs to determine the most suitable approach for our dataset.

3.2 Model Training

Logistic Regression

Training Process:

For logistic regression, the training process involves fitting the model to the training data using an optimization algorithm such as gradient descent. The model learns the coefficients for each feature, which define the decision boundary separating the two classes. During training, we iterate over the training data multiple times, adjusting the coefficients to minimize the logistic loss function.

Implementation:

We initialize a logistic regression model and train it using the scaled training data. The model learns the optimal coefficients that best fit the training data and minimize the error between predicted and actual outcomes. The trained logistic regression model can then be used to make predictions on new data.

Support Vector Machines (SVMs)

Training Process:

Training an SVM involves finding the optimal hyperplane that maximizes the margin between different classes in the feature space. This process is achieved by solving a convex optimization problem, where the goal is to minimize the classification error while maximizing the margin. SVMs can handle both linearly separable and non-linearly separable data by using different kernel functions.

Implementation:

We initialize an SVM classifier and train it using the scaled training data. The SVM algorithm learns the optimal decision boundary that separates malignant and benign tumors in the feature space. By adjusting the hyperparameters such as the choice of kernel and regularization parameter, we can optimize the performance of the SVM model.

Random Forests

Training Process:

Random Forests train multiple decision trees on random subsets of the training data and features. Each tree is trained independently, and predictions are made based on the majority vote of all trees (classification) or the average prediction (regression). This ensemble approach helps reduce overfitting and improve generalization.

Implementation:

To implement Random Forests, we initialize an ensemble of decision trees and train them using the training data. Each tree learns from a random subset of the data and features. Predictions are then aggregated to make the final prediction.

K-Nearest Neighbors (KNN)

Training Process:

KNN classifies new data points based on the majority class of their k nearest neighbors in feature space. During training, KNN stores the entire training dataset and uses it for classification. The choice of k affects the model's bias-variance tradeoff, with smaller k values leading to more flexible decision boundaries.

Implementation:

To implement KNN, we initialize the model and store the training data. When making predictions, KNN identifies the k nearest neighbors of each data point and assigns the majority class label among them. The optimal value of k is typically determined through cross-validation.

3.3 Model Evaluation

In the model evaluation section, we assessed the performance of four classification models: logistic regression, support vector machines (SVMs), K-Nearest Neighbors (KNN), and Random Forest, for the task of breast cancer detection. Each model was evaluated based on metrics such as accuracy, precision, recall, and F1-score.

The logistic regression model achieved an accuracy of approximately 0.9766 on the testing data, with precision, recall, and F1-score all around 0.9683. Similarly, the SVM model demonstrated high performance, with an accuracy of approximately 0.9825, and precision, recall, and F1-score all above 0.95. The KNN model reported an accuracy of 0.9591, precision of 0.9516, recall of 0.9365, and an F1-score of 0.944. The Random Forest model showed superior metrics, with an accuracy of 0.9708, precision of 0.9833, recall of 0.9365, and an F1-score of 0.9593.

Furthermore, we conducted an analysis to check for overfitting. The logistic regression model showed a slight indication of potential overfitting, with a training accuracy of 0.9868 compared to the testing accuracy of 0.9766. In contrast, the SVM model exhibited similar training and testing accuracies, indicating a lower risk of overfitting. Both KNN and Random Forest models demonstrated robust performance without significant signs of overfitting.

Considering the overall performance and potential for overfitting, all models appear promising for breast cancer detection. However, logistic regression may be preferred due to its simplicity and interpretability, while SVMs could offer better generalization performance for complex datasets. KNN provides a good balance between simplicity and performance, especially when the data exhibits non-linear relationships. Random Forest stands out for its robustness and ability to handle large feature spaces effectively.

3.4 Final Model Choice

After careful consideration of the performance, interpretability, and potential for overfitting of the logistic regression, support vector machines (SVMs), K-Nearest Neighbors (KNN), and Random Forest models for breast cancer detection, we have chosen logistic regression as the preferred model.

Logistic regression, SVMs, KNN, and Random Forest all demonstrated high accuracy, precision, recall, and F1-score on the testing data, indicating their effectiveness in classifying benign and malignant tumors. However, upon analyzing for overfitting, the SVM model showed a slight indication of potential overfitting with higher training accuracy compared to the testing accuracy. In contrast, the logistic regression model exhibited similar accuracies on both training and testing data, suggesting better generalization performance and a lower risk of overfitting. Both KNN and Random Forest models performed robustly but did not show significant advantages over logistic regression in terms of simplicity and interpretability.

While SVMs, KNN, and Random Forest may provide better generalization performance for complex datasets, logistic regression offers simplicity and interpretability, making it easier to understand the underlying relationships between features and the target variable. Given the slight indication of potential overfitting in the logistic regression model and the strong performances of other models, the decision to prioritize interpretability and transparency in the final model choice was clear.

Therefore, logistic regression is selected as the final model for breast cancer detection. Its transparent nature allows for easier understanding of the underlying factors contributing to tumor classification decisions, facilitating insights into the diagnostic process. Additionally, logistic regression's simplicity makes it well-suited for deployment in clinical settings, where interpretability and ease of implementation are paramount.

Conlusion:

Equipped with a model that carefully balances the intricacies of machine learning with the nuances of medical data, we proceeded to the next phase: the development of a user-friendly application. This application, created with Streamlit and deployed in the cloud, brings our logistic regression model directly into the hands of clinicians and patients.

4 CHAPTER 4 : APPLICATION DEVELOPMENT AND DEPLOYMENT

4.1 Choice of Technology

4.1.1 Introduction to Streamlit

Streamlit is a Python library that allows for rapid prototyping of data-driven web applications. It simplifies the process of creating interactive web apps by enabling developers to focus on writing Python code, it also provides user-friendly applications.

Figure 13:Streamlit logo.

4.1.2 Advantages for prototyping

-Streamlit provides a simple and intuitive way to create web applications using Python. Developers can focus on writing code without needing expertise in web development technologies like HTML, CSS, or JavaScript.

-With Streamlit, developers can quickly prototype ideas and concepts, allowing for faster iteration and experimentation. Changes made to the code are immediately reflected in the browser, streamlining the development process.

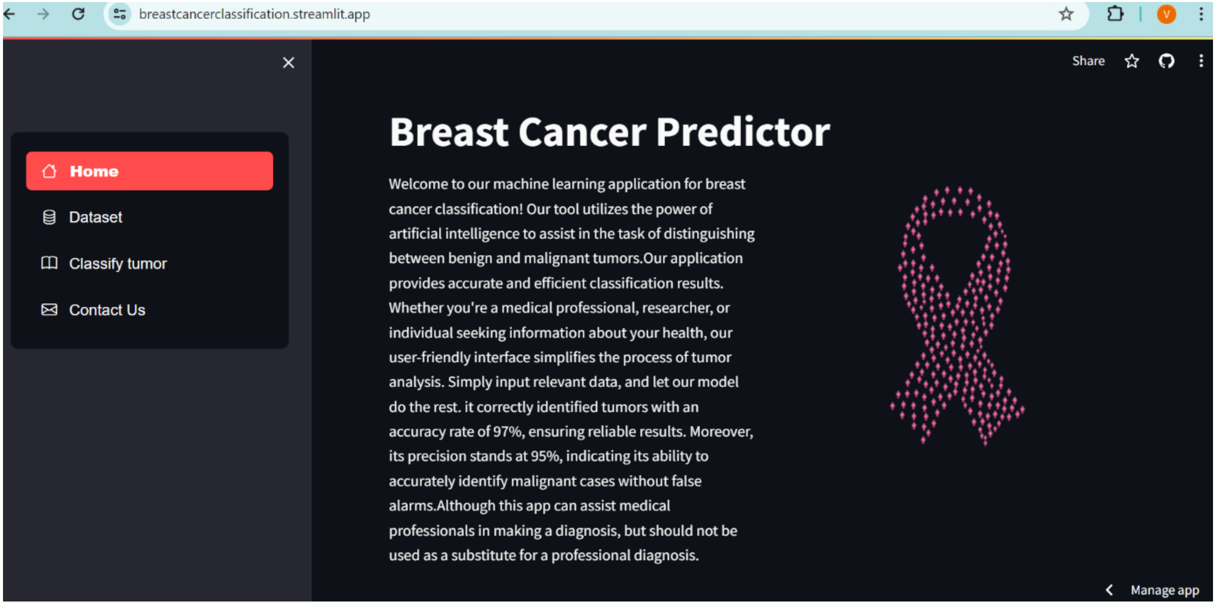
-Streamlit applications can be easily deployed on various platforms, including cloud services like Heroku, AWS, and Google Cloud Platform, as well as streamlit community cloud where you can deploy your application for free in just a few minutes. This flexibility makes it easy to share applications with others or deploy them for production use.

4.2 Web Application Functionality

4.2.1 User interface

- **Home :**

The "Home" menu option serves as the landing page of our web application. It provides an introduction to the application, its purpose, and relevant informations for users. This section includes a brief overview of the breast cancer classification task.



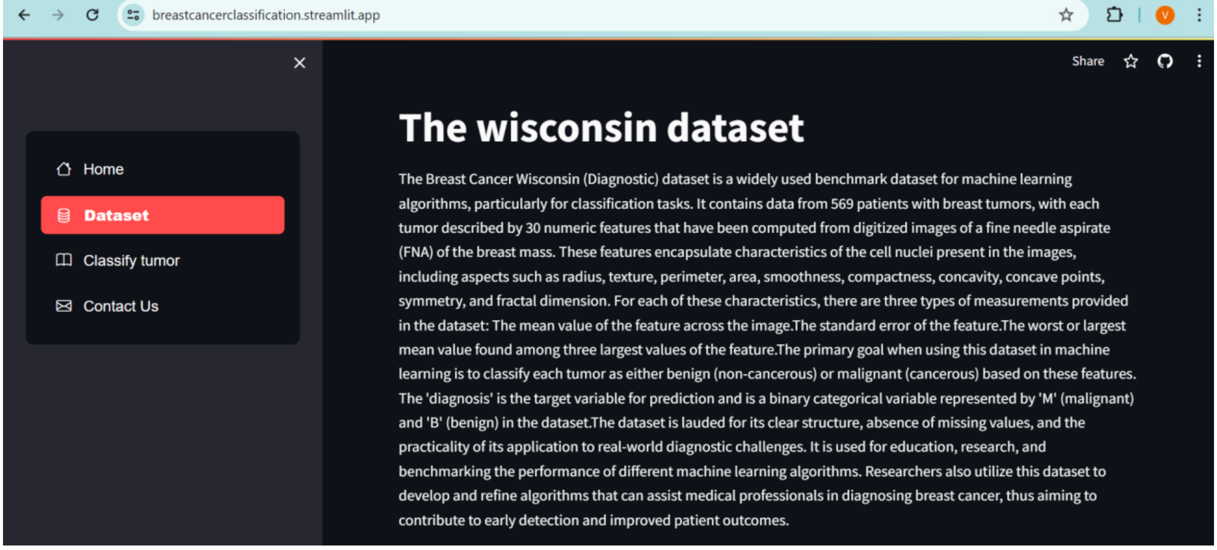
Breast Cancer Predictor

Welcome to our machine learning application for breast cancer classification! Our tool utilizes the power of artificial intelligence to assist in the task of distinguishing between benign and malignant tumors.Our application provides accurate and efficient classification results. Whether you're a medical professional, researcher, or individual seeking information about your health, our user-friendly interface simplifies the process of tumor analysis. Simply input relevant data, and let our model do the rest. it correctly identified tumors with an accuracy rate of 97%, ensuring reliable results. Moreover, its precision stands at 95%, indicating its ability to accurately identify malignant cases without false alarms.Although this app can assist medical professionals in making a diagnosis, but should not be used as a substitute for a professional diagnosis.

Figure 14:Home page.

- **Dataset**

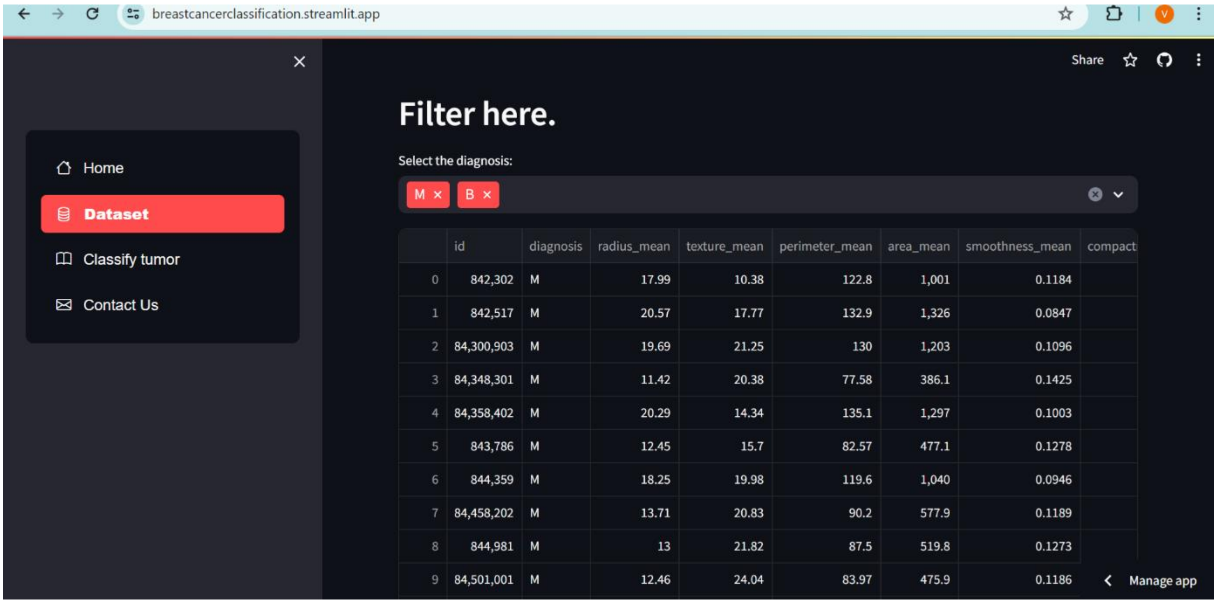
The "Dataset" menu option allows users to explore the dataset used for training and testing the breast cancer classification model. This section provides transparency into the data used to build the model, including details such as the number of samples, features, and classes. Users can download the dataset for further analysis or reference. Providing access to the dataset enhances the transparency and reproducibility of our model's results.



The wisconsin dataset

The Breast Cancer Wisconsin (Diagnostic) dataset is a widely used benchmark dataset for machine learning algorithms, particularly for classification tasks. It contains data from 569 patients with breast tumors, with each tumor described by 30 numeric features that have been computed from digitized images of a fine needle aspirate (FNA) of the breast mass. These features encapsulate characteristics of the cell nuclei present in the images, including aspects such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. For each of these characteristics, there are three types of measurements provided in the dataset: The mean value of the feature across the image.The standard error of the feature.The worst or largest mean value found among three largest values of the feature.The primary goal when using this dataset in machine learning is to classify each tumor as either benign (non-cancerous) or malignant (cancerous) based on these features. The 'diagnosis' is the target variable for prediction and is a binary categorical variable represented by 'M' (malignant) and 'B' (benign) in the dataset.The dataset is lauded for its clear structure, absence of missing values, and the practicality of its application to real-world diagnostic challenges. It is used for education, research, and benchmarking the performance of different machine learning algorithms. Researchers also utilize this dataset to develop and refine algorithms that can assist medical professionals in diagnosing breast cancer, thus aiming to contribute to early detection and improved patient outcomes.

Figure 15: The Wisconsin dataset introduction page.



Filter here.

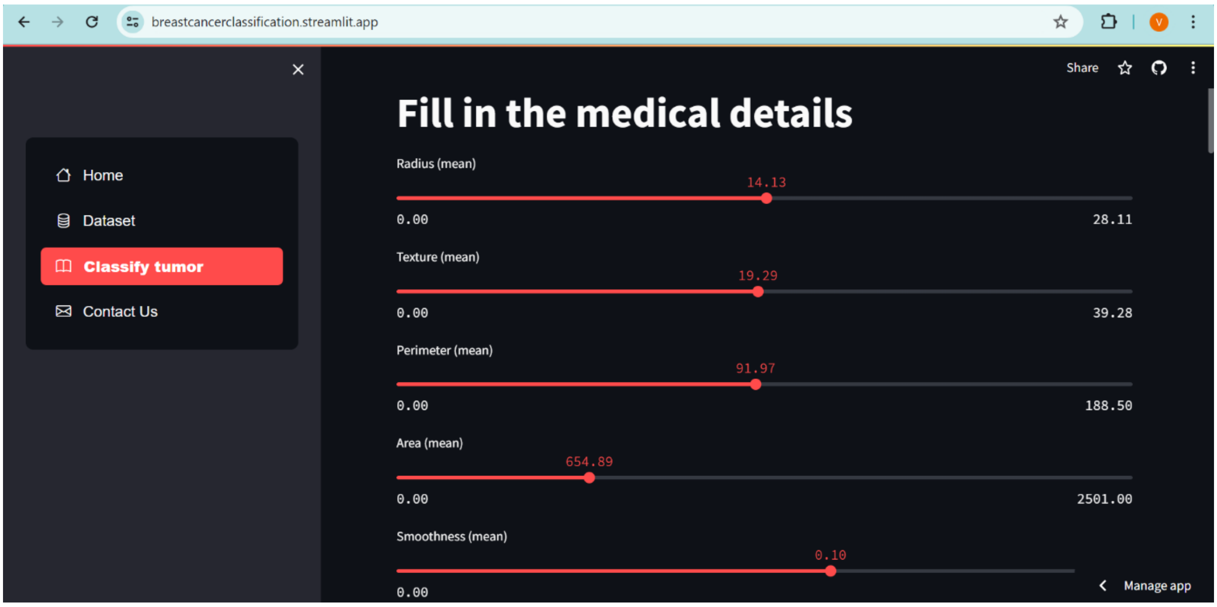
Select the diagnosis: M × B ×

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compact
0	842,302	M	17.99	10.38	122.8	1,001	0.1184	
1	842,517	M	20.57	17.77	132.9	1,326	0.0847	
2	84,300,903	M	19.69	21.25	130	1,203	0.1096	
3	84,348,301	M	11.42	20.38	77.58	386.1	0.1425	
4	84,358,402	M	20.29	14.34	135.1	1,297	0.1003	
5	843,786	M	12.45	15.7	82.57	477.1	0.1278	
6	844,359	M	18.25	19.98	119.6	1,040	0.0946	
7	84,458,202	M	13.71	20.83	90.2	577.9	0.1189	
8	844,981	M	13	21.82	87.5	519.8	0.1273	
9	84,501,001	M	12.46	24.04	83.97	475.9	0.1186	

Figure 16: Dataset visualization page.

- ### Classify tumor

The "Classify Tumor" menu option is the core functionality of our web application. It enables users to interactively classify breast cancer tumors using the deployed machine learning model. Users may input relevant features of a tumor, such as its radius, perimeter, texture, and symmetry, through intuitive form fields. Upon submission, the model predicts the likelihood of the tumor being malignant or benign and presents the results to the user. Clear and informative visualizations, such as prediction probabilities can aid users in understanding the model's predictions.



Fill in the medical details

Radius (mean) 14.13 (0.00 – 28.11)

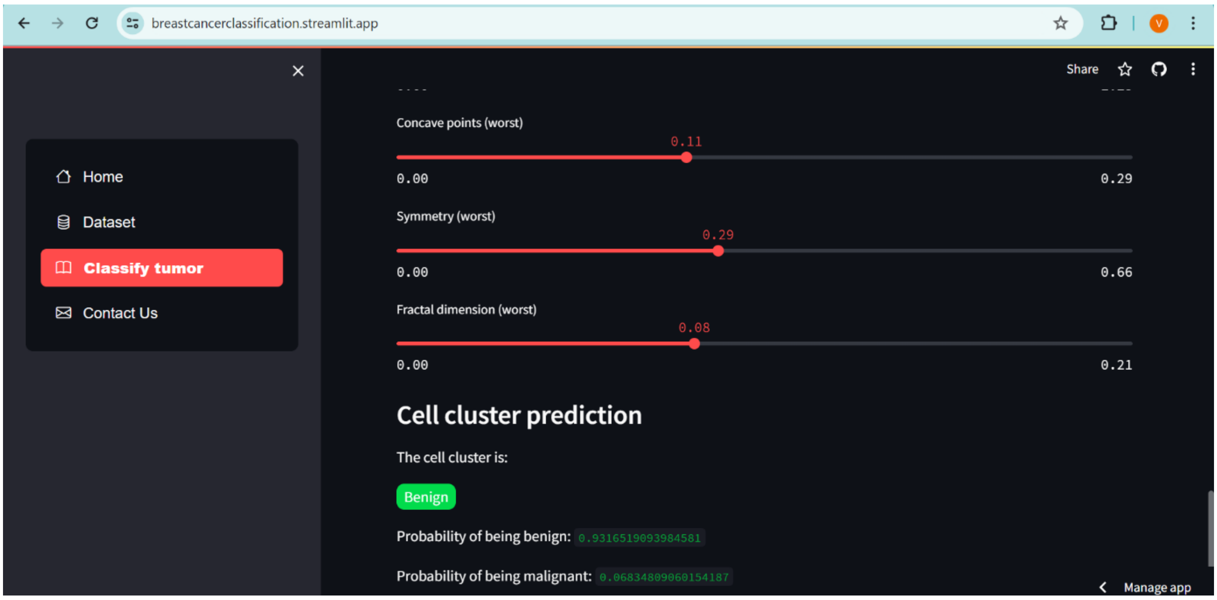
Texture (mean) 19.29 (0.00 – 39.28)

Perimeter (mean) 91.97 (0.00 – 188.50)

Area (mean) 654.89 (0.00 – 2501.00)

Smoothness (mean) 0.10 (0.00 –

Figure 17: Classify tumors page part1.



Concave points (worst) 0.11 (0.00 – 0.29)

Symmetry (worst) 0.29 (0.00 – 0.66)

Fractal dimension (worst) 0.08 (0.00 – 0.21)

Cell cluster prediction

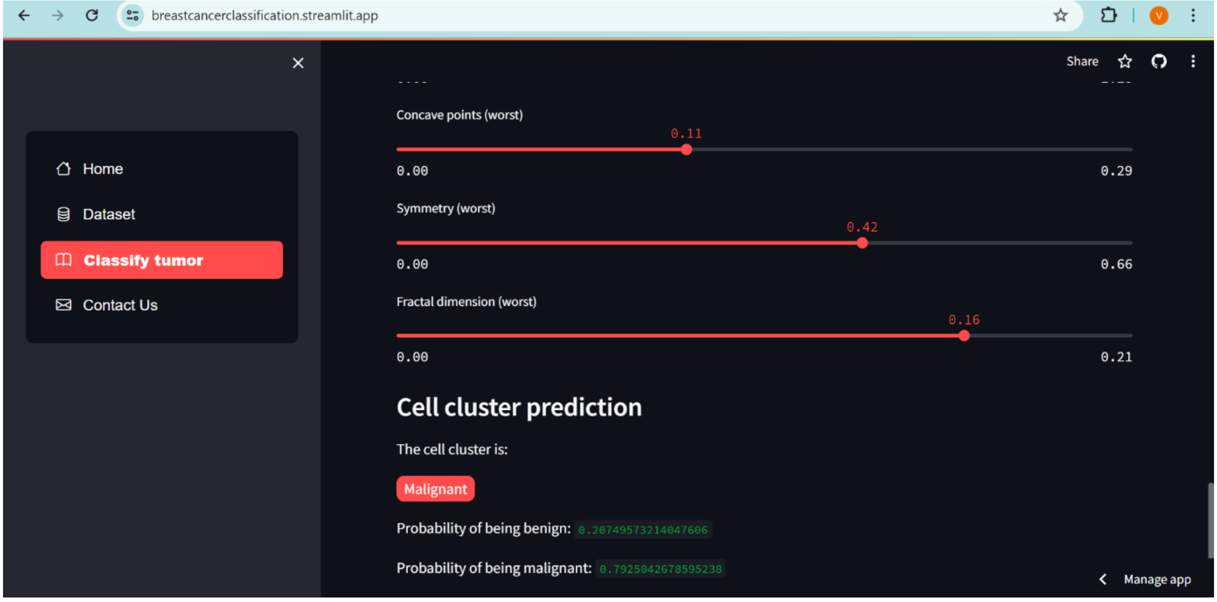
The cell cluster is:

Benign

Probability of being benign: `0.9316519093984581`

Probability of being malignant: `0.06834809060154187`

Figure 18: Classify tumors page part2.



Concave points (worst) 0.11 (0.00 – 0.29)

Symmetry (worst) 0.42 (0.00 – 0.66)

Fractal dimension (worst) 0.16 (0.00 – 0.21)

Cell cluster prediction

The cell cluster is:

Malignant

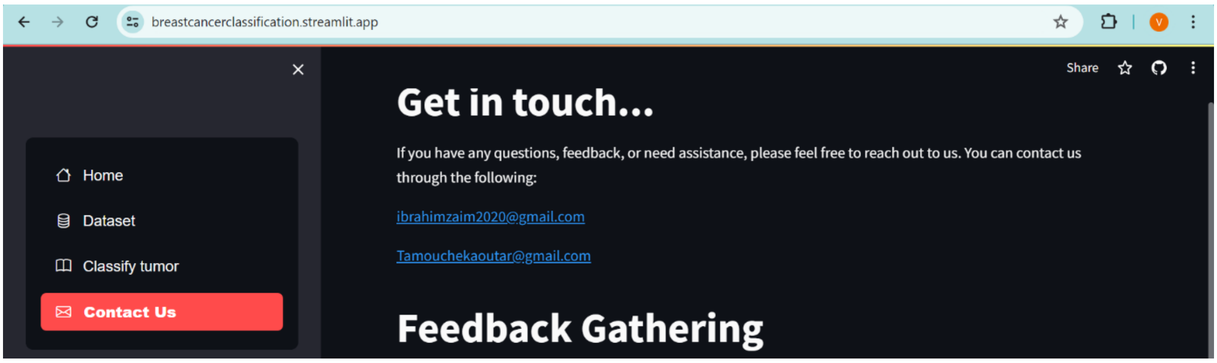
Probability of being benign: `0.20749573214047606`

Probability of being malignant: `0.7925042678595238`

Figure 19: Classify tumors page part3.

- Contact us

In the "Contact Us" section of our breast cancer classification application, we offer users the opportunity to rate our application, provide feedback, ask questions, or reach out for further assistance. This section serves as a channel for direct communication between users and our team.



Get in touch...

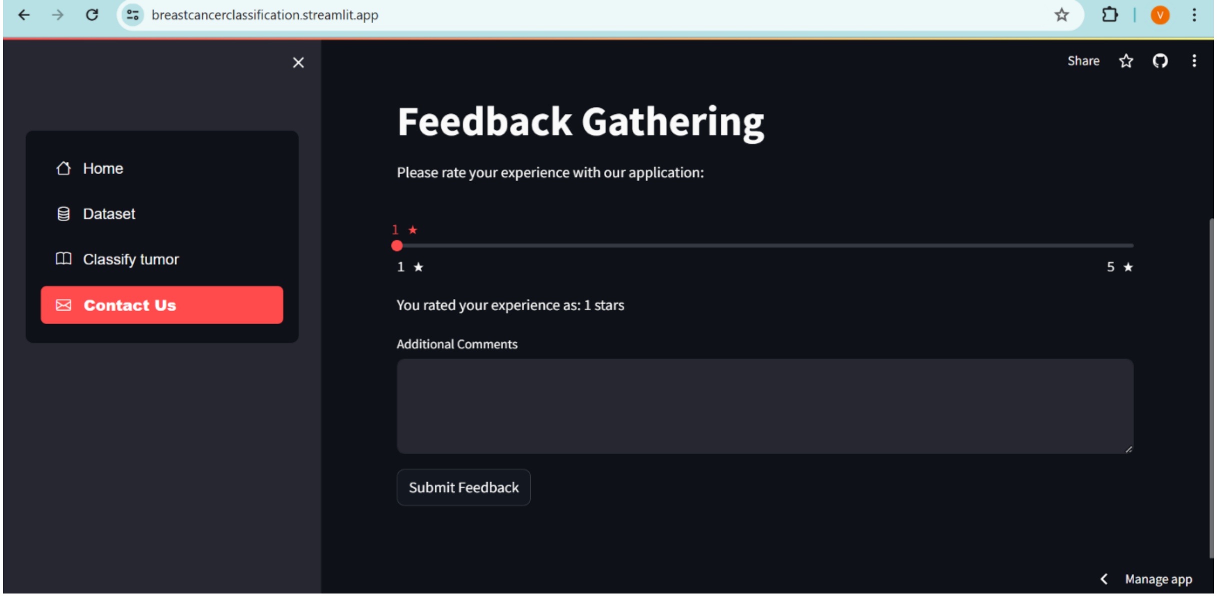
If you have any questions, feedback, or need assistance, please feel free to reach out to us. You can contact us through the following:

ibrahimzaim2020@gmail.com

Tamouchekaoutar@gmail.com

Feedback Gathering

Figure 20: Contact page « Get in touch ».



Feedback Gathering

Please rate your experience with our application:

You rated your experience as: 1 stars

Additional Comments

Figure 21:Contact us page "Give feedback".

4.2.2 Feature interactivity

The features interactivity in our breast cancer classification application refers to the interactive elements and functionalities that enable users to actively engage with the application and manipulate data in real-time. Here are some examples of the features interactivity in our application:

- Data Exploration: Users can explore the dataset used for training the classification model through interactive data exploration features to gain insights into tumor characteristics and classification outcomes.

- Input Form: Users can interact with the application by inputting relevant data related to breast tumor characteristics, such as size, shape, texture, and margin. This input form allows users to customize their queries and receive personalized predictions based on their specific data.

- Prediction Results: After submitting their input data, users receive real-time predictions regarding the likelihood of the tumor being malignant or benign. The application displays prediction results dynamically, allowing users to see how changes in input parameters affect the classification outcome.

-Feedback Mechanism: Establishing a feedback loop with users rating our application besides adding additional comments which allows for continuous improvement of the application. By actively soliciting and incorporating user feedback, developers can identify areas for enhancement, prioritize feature requests, and address usability issues to enhance the overall user experience. The feedbacks are stored in a json-server.

4.3 Deployment Strategy

4.3.1 Deployment process

In order to deploy this application in the streamlit Community Cloud, we have connected our streamlit Community Cloud account directly to our GitHub repository. After making updates to our application code locally and pushing them to GitHub, the Streamlit Community Cloud automatically triggers the deployment process, fetching the latest code, building the Streamlit application, and deploying it to the cloud environment. Users can then access the deployed application via a unique URL, where they can interact with it in real-time and provide feedback. Additionally, scalability options ensure that our application can handle increased demand as it grows in popularity.

4.3.2 Deployment challenges

To deploy this application, we have encountered several challenges:

-Limited Customization: Streamlit is designed for rapid prototyping and ease of use, which may result in limited customization options compared to more traditional web development frameworks.

- Dependency Management: Managing dependencies and ensuring compatibility with Streamlit can be challenging, especially when working with many libraries or complex project requirements. We have encountered conflicts between Streamlit and other Python libraries.

4.4 Future Enhancements

User Authentication and Authorization: Implementing robust user authentication and authorization mechanisms to ensure secure access to sensitive data and features within the application. This includes user registration, login/logout functionality, and role-based access control to protect patient information and maintain compliance with data privacy regulations.

Real-time data: data entered by the user is seamlessly added to our dataset by continuously updating our dataset with new user-contributed data in real-time, we can enhance the accuracy of our classification model. The addition of fresh data allows for ongoing model retraining, ensuring that the model remains up-to-date and capable of capturing the latest trends and patterns in breast tumor characteristics. The ability for users to see their contributions reflected in real-time within the application creates a sense of ownership and engagement.

Users become more invested in the application's success and are motivated to provide accurate and relevant data.

Conclusion:

In conclusion, our journey through the development of our breast cancer classification application involved precise data handling, model selection, evaluation, and deployment.

Firstly, we carefully selected and preprocessed the data to ensure its quality and relevance to our task. This step involved cleaning the dataset, handling missing values, and normalizing features to create a solid input for our models.

Next, we explored various machine learning algorithms, including logistic regression, SVMs, KNN, and random forests. After thorough experimentation and evaluation using metrics such as accuracy, precision, F1-score, and recall. As well as taking in other factors for instance the overfitting and the complexity of each model, we found logistic regression to be the most suitable model for our application.

Furthermore, we deployed our application on Streamlit Community Cloud, ensuring accessibility and usability for healthcare professionals and patients. This deployment process involved packaging our model into a user-friendly interface, allowing users to interact with the application seamlessly.

In summary, our breast cancer app represents a big step in our machine learning journey. It's our first project out in the real world, and it's taught us a lot about how AI can help in healthcare. By making it easier to spot breast cancer early. This project has been a great learning experience, showing us how to handle data, train models, build websites, and get them out there for people to use. Looking ahead, we are excited to keep learning and using AI to make a positive impact in the world.