Republic of Tunisia
الجمهورية التونسية
Ministry of Higher Education and Scientific Research
وزارة التعليم العالي و البحث العلمي

University of Tunis El Manar
جامعة تونس المنار
National Engineering School of Tunis
المدرسة الوطنية للمهندسين تونس

The End of Second Year Project

Information and Communication Technologies Department

A Machine Learning Based Prototype for Diseases Detection Case of Breast Cancer

Elaborated by :
Amen Allah BERREJEB
Donies HADDAD

Supervised by:
Ms. Wafa MEFTEH

University year: 2022/2023

Abstract

This report focuses on creating a web application to predict breast cancer using machine learning techniques. Patient and tumor data was analyzed and machine learning models were used to predict breast cancer diagnosis. An easy-to-use web application was created to provide a convenient way for users to predict their risk of developing breast cancer.

Keywords: Breast Cancer, Machine Learning, Prediction, Web Application, Streamlit, Data Cleaning, Web Modeling.

Résumé

Ce rapport se concentre sur la création d'une application web pour prédire le cancer du sein en utilisant des techniques d'apprentissage automatique. Les données des patients et des tumeurs ont été analysées et des modèles d'apprentissage automatique ont été utilisés pour prédire le diagnostic du cancer du sein. Une application web facile à utiliser a été créée pour offrir un moyen pratique aux utilisateurs de prédire leur risque de développer un cancer du sein.

Mots Clés : Cancer du sein, Apprentissage Automatique,Streamlit, Prédiction, Application web, Modélisation web.

Acknowledgements

We would like to thank our supervisor, Mrs Wafa MEFTAH, who gave us the opportunity to work on this project. This project allowed us to value the role of engineers in improving health care system. We would also like to thank the administrative staff and teachers of the National School of Engineering in Tunis for providing useful resources and a good education. Finally, we would like to warmly thank the members of the jury who agreed to evaluate this work.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

ML Machine Learning

DCIS Ductal Carcinoma in situ

ICC nvasive Ductal Carcinoma

ILC Invasive Lobular Carcinoma

MRI Magnetic Resonance Imaging

PET Positron Emission Tomography

CT Computed Tomography

KNN k-nearest Neighbor

DT Decision Tree

SVM Support Vector Machines

RF Random Forest

BCWD Breast Cancer Wisconsin Diagnostic

DDSM Digital Database for Screening Mammography

ANN Artificial Neural Network

CNN Convolutional Neural Network

AUC-ROC Area under the Receiver Operating Characteristic Curve

DM Digital Mammography

CAD Computer Aided Detection

DBSCAN Density-Based Spatial Clustering of Applications with Noise

SARSA State-Action-Reward-State-Action

HCA Hierarchical Cluster Analysis

WDBC Wisconsin Diagnostic Breast Cancer

General Introduction

Context

Breast cancer is the leading cause of cancer death among women worldwide. It is the leading cause of cancer death among women in developing countries and the second leading cause among women in developed countries [37, 12, 6]. In developing countries like Tunisia, the incidence of this cancer is still increasing.The International Organization for research on Cancer: World Health Organisation: shows that there are 3,092 new cases of breast cancer in women and 986 deaths from this disease in 2020 [3, 31].
As statistics show, cancer is an aggressive disease with a low median survival rate. Moreover, the treatment process is lengthy and very expensive due to its high recurrence and mortality rates.Hence, early diagnosis of cancer is essential to improve patient survival.

Advances in statistics and computer engineering over the years have encouraged many scientists to apply computer methods such as multivariate statistical analysis to analyze the prognosis of the disease. Furthermore, artificial intelligence, especially machine learning and deep learning, has found popular applications in clinical cancer and health research in recent years. These methods are nowadays playing a major role in improving the accuracy of predictions of cancer susceptibility, recurrence and survival[17].

For our end of second-year project, we chose to create a prototype based on machine learning for disease detection. More specifically, the case of breast cancer.

Problematic

What methods can be used to simplify the context of machine learning and make it more user-friendly for all individuals, while also effectively combining theoretical machine learning with practical applications.

Issues

Several problems are to be highlighted in the context of our work.
Firstly, in the case of breast cancer, the difficulty of diagnosing a cancer specialists are not always able to distinguish a cancer in the results of the analysis, we will try to solve this problem by AI.
Then, we will have to find a ML model that will meet the constraints and objectives that

we have determined. Hence, we must make a meticulous study of the methods to be used in order to obtain a more than reliable solution.
Finally, we will implement the methods and algorithms created to facilitate and restore confidence in women to make more diagnoses at home.

Objectives

Our goal is to develop a cutting-edge web application that can provide a powerful tool for both doctors and patients, allowing for early detection and better treatment options. We believe that by combining the latest in machine learning and data analysis techniques, we can create a solution that will have a meaningful impact on breast cancer outcomes. Our team is dedicated to creating a user-friendly and accessible platform that will help individuals to take control of their health and support medical professionals.

Report Organization

Our report on the development of a web application for predicting breast cancer is structured into five chapters that cover all aspects of the project :

Chapter 1, titled "Preliminary Study," provides a comprehensive overview of breast cancer from a medical standpoint, giving readers an in-depth understanding of the disease and its effects.

Chapter 2, titled "Overview about Related Works and Contribution," explores existing literature and research related to breast cancer prediction and discusses our contributions to this field.

Chapter 3, titled "Modeling and Implementation of the ML Component," explores the world of machine learning and discuss the different models used for breast cancer prediction.

Chapter 4, titled "Modeling and Implementation of the Web Platform," covers the development and implementation of the web application, including Both doctor and patient interfaces.

Finally, Chapter 5, titled "Integration and Recommandation" concludes with an analysis of our results, a discussion of future advancements, and we discuss how we integrated the machine learning model in the web part.

Chapter 1

Preliminary Study

Introduction

Cancer is one of the most dreaded diseases in the history of mankind. It never ceases to test the ability of science to fight it, since it has a unique characteristic: it comes from ourselves. In this chapter, we will define cancer in general and breast cancer in particular under its different types. In addition, we will study the steps that follow from its diagnosis in order to deduce the motivations that determined the introduction of AI in this specific field of medicine. Finally, we will study how to involve machine learning to train algorithms on large datasets and allowing them to make predictions or decisions based on that data.

1.1 Breast Cancer

Breast cancer is a complex and devastating disease that affects millions of people worldwide. In this section, we will explore the medical background of breast cancer and discuss its various types and causes.

1.1.1 The Human Cell

We are made up of about 30 trillion cells, although we started from one cell. This cell replicates itself, and each copy replicates itself, and that happens over and over again. Finally, these cells form a complete human being [13].

Figure 1.1 shows that cells are generally composed of a nucleus that contains DNA and a cytoplasm bounded by a membrane.
A cell becomes cancerous when it changes and multiplies uncontrollably [19].

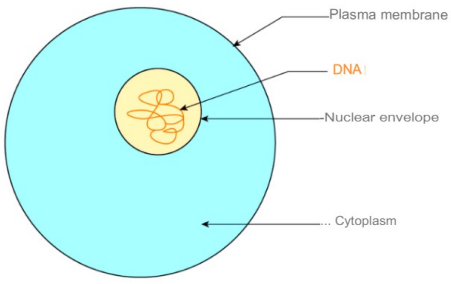

Figure 1.1: The human cell.

Cell Division :

In order to reproduce, a cell subdivides to produce two more daughter cells. This operation is performed to replace 100 billion cells per day in humans [25]. A tumor occurs when there is an error in the copying process, and the replicated cell is defective. The awful mistake is that they continue to replicate without any limits. Moreover, they

do not follow normal life cycles of cells [13].
Figure 1.2 shows the difference between normal and abnormal cells.

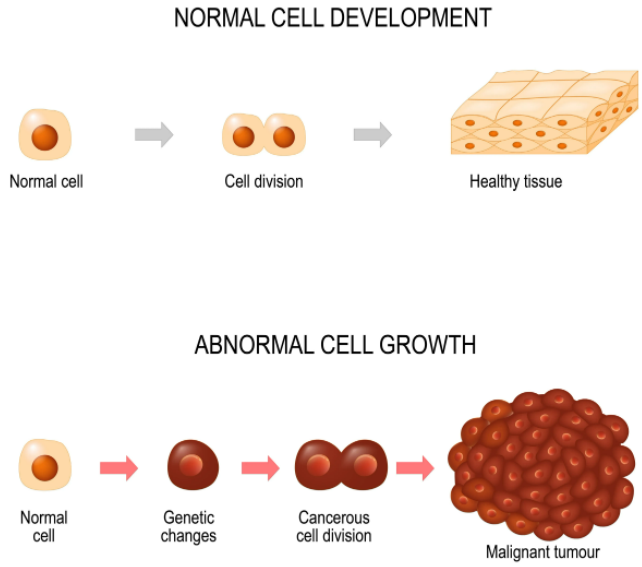

Figure 1.2: Cell development[13].

Tumors develop and behave differently. There are two types of tumors: benign tumours are non-cancerous and do not have harmful health effects, and malignant tumours are cancerous tumours [33].
Figure 1.3 shows the differences between benign and malignant tumors.

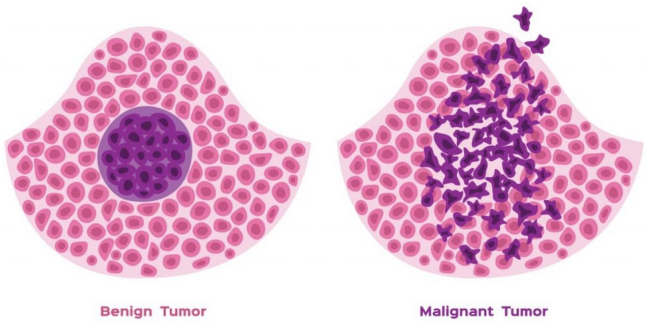

Figure 1.3: Benign and cancerous tumors [9].

1.1.2 Definition

Breast cancer starts in breast cells. The cancerous (malignant) tumor is a group of cancer cells that can invade neighboring tissues and destroy them. It can also spread (metastasize) to other parts of the body [9].

There are several types of breast cancer. There are in breast cancers in order of frequency: infiltrating ductal carcinomas, infiltrating lobulars, papillae and mucinous. This varies according to their histological type, that is to say from which cells the development of breast cancer takes place [27].

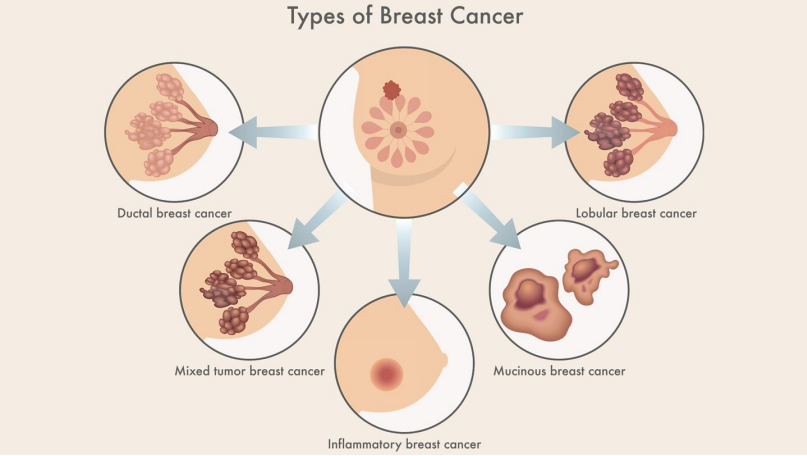

Figure 1.4: Types of breast cancer guide [7].

1.1.3 Anatomy and Physiology of the Breast

The breast is an organ whose structure reflects its particular function: milk production for lactation (breastfeeding) [11].

The structure of the breast is illustrated in the figure 1.5.

- **The lobes** is linked by thin tubes called **ducts**.
- **These ducts** lead to **the nipple** in the center of a dark area of skin called **the areola**.
- **Fat** fills the spaces between **lobules** and **ducts**.
- There are no muscles in the breast, but **muscles** lie under each breast and cover **the ribs**.

1.1.4 Type of Cancer

Non-invasive breast cancer :

Non-invasive cancers are also called in situ cancers, which means that the cancer cells are found only in one place and have not spread to the tissues around the milk ducts or glands (lobules) [26].

- **Ductal carcinoma in situ (DCIS):** refers to the presence of abnormal cells within a milk duct in the breast.These cells do not spread out of the lobules into nearby breast tissue [26].

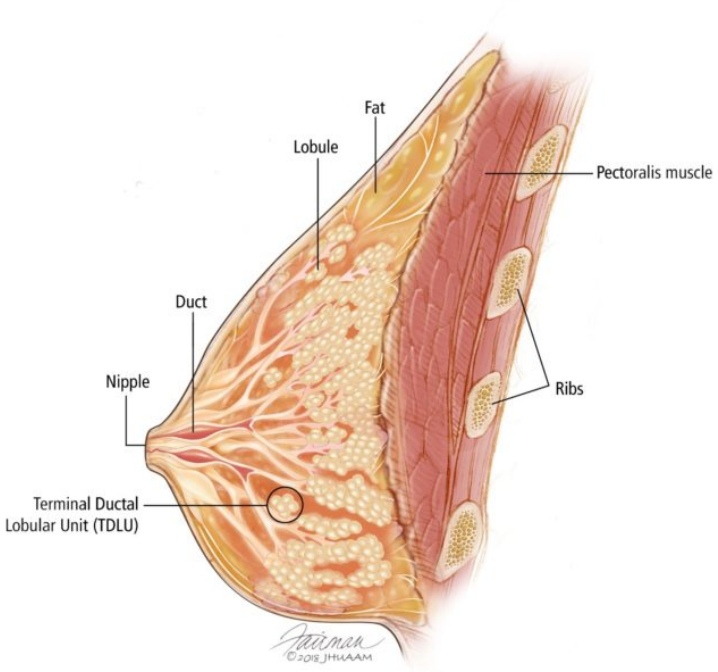

Figure 1.5: Anatomy of the Breast[34].

- **Ductal carcinoma in situ (DCIS):** is a rare condition in which abnormal cells form in the milk glands (lobules) in the breast. DCIS is not a cancer. But the women have an increased risk of developing breast cancer [26].

Invasive Breast Cancer :

Invasive breast cancers, also known as invasive cancers, develop in the normal tissue around a galactophore canal or lobule and can spread to other parts of your body [26].

- **Invasive Ductal Carcinoma (ICC):** forms in a galactophore channel before reaching surrounding tissue [26].

- **Invasive Lobular Carcinoma (CLI):** forms in the lactiferous glands before reaching the surrounding tissues [26].

General Characteristics of Tumors :

There are a multitude of parameters to consider when determining the nature of a histology, the criteria of benignity and malignancy have no exact values [25][22]. In the following we will quote the most common signs.

- **Benign Tumors:**
  - Normal cell nucleus.
  - Slow cell division.

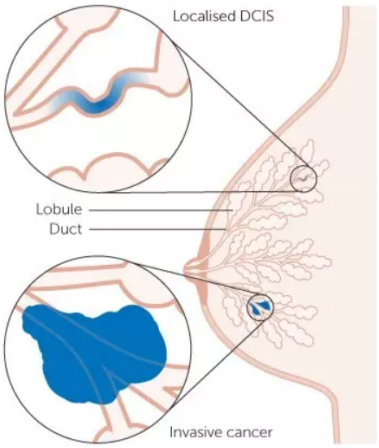

Figure 1.6: Invasive or Non-invasive breast cancer[24].

  - The cancer cell looks like the cells that gave birth to it.
  - No recurrence: surgical removal of a tumor does not lead to a next recurrence.

- **Malignant Tumors :**
  - Abnormally large and dark core.
  - Cells with different sizes and high cell division.
  - Invasion of surrounding tissue (poorly bounded periphery, perforated basement membrane).
  - Irregularity of cell contour. recurrence.
  - Increase of the nucleus/cytoplasm ratio.

1.1.5 The Cancer Detection Process

Breast cancer detection typically involves a combination of screening tests and diagnostic tests.
Screening tests are used to identify potential cases of breast cancer in individuals who do not have any symptoms. The most common screening test for breast cancer is mammography, which uses low-dose X-rays to produce images of the breast tissue. Other screening tests include breast MRI (magnetic resonance imaging) and clinical breast exams.
Diagnostic tests are used to confirm or rule out a diagnosis of breast cancer in individuals who have symptoms or an abnormal screening result. Diagnostic tests may include diagnostic mammography, breast ultrasound, breast biopsy, and other imaging tests such as PET (positron emission tomography) scans or CT (computed tomography) scans.
If breast cancer is detected, additional tests may be ordered to determine the extent of the cancer, including its size, location, and whether it has spread to other parts of the body. This process is called staging and may involve imaging tests such as bone scans or CT scans, as well as blood tests and other diagnostic procedures.
To summarize, the diagnosis of a cancer requires the realization of several examinations biological and imaging examinations. All of these examinations allow us to propose to the

patient the therapeutic strategy most adapted to his case. It is important to note that early detection of breast cancer is key to successful treatment and improved outcomes, so it is recommended that individuals undergo regular screening according to their age and risk factors [15].
Different types of examinations can be performed: clinical, biological or imaging examinations :

Blood Biomarkers :

A blood test is used to evaluate the general health and the functioning of the main organs. The presence of particular blood markers can be looked for. These are usually proteins that are abnormally produced by the cancer cells [29].

Medical Imagery :

There are currently a number of imaging techniques of the human body commonly used in the medical field. Each of them is sensitive to a particular type of contrast and finds its applications for different organs. Several techniques can also provide complementary information on the same organ. The medical imaging tools used for breast cancer screening and diagnosis are: ultrasound (ultrasound imaging), MRI (Magnetic Resonance Imaging) and mammography (X-ray imaging). In the following, we present the different existing techniques and their characteristics [8].

Radiography :

The X-ray is analyzed with the naked eye by the doctor, hence he cannot identify tumors of very small size. In addition, the presence of gas in the body introduces noise in the resulting image.

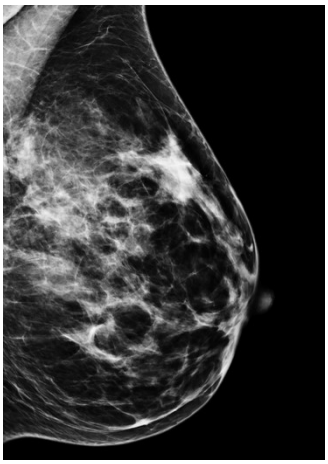

Figure 1.7: A traditional mammogram is taken as an X-ray [8].

Ultrasound :

The principle of ultrasound is to apply a probe against the skin opposite the organ to be explored. This probe emits ultrasounds which pass through the tissues and are then sent back to the are sent back to it in the form of an echo. This signal, once collected, is analyzed by a computer system that transmits a live image on a video screen. Figure 1.7 shows two examples of ultrasound images of the breast with a) the right breast has a malignant lesion and b) left breast of the same patient has benign lesion [8].

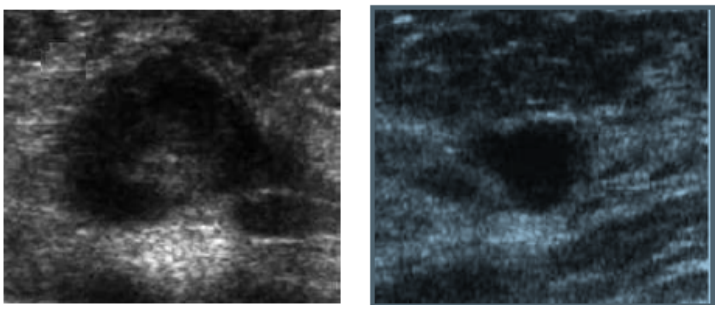

Figure 1.8: Examples of ultrasound images of the breast. a) malignant lesion (right breast) and b) benign lesion (left breast of the same patient) [8].

Magnetic Resonance Imaging IRM:

This technique is based on the use of a magnet and radio frequency waves. Figure 1.9 is showing the MRI image of the left and right breasts of figure 1.8.

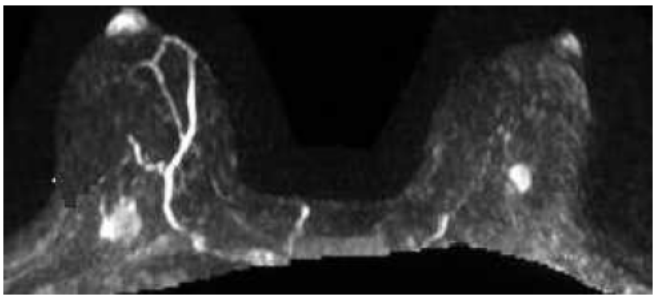

Figure 1.9: Example of breast MRI. a) malignant lesion and b) benign lesion [1].

Breast Biopsy :

It is impossible to know from a screening test whether a lesion is cancerous or not. In order to determine if a lesion is cancerous, the tissue of the lesion must be studied directly under a microscope to see if cancerous cells are present or not. It is therefore necessary to take a tissue sample, that is, a biopsy [30].

After we have learned the process of detection of breast cancer using medical methods, in the third section, we will present the functioning of machine learning algorithms in order to be able to use data to predict this disease.

1.2 Machine Learning

Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science [21].

1.2.1 Definition

Machine Learning is the term most often used to designate artificial intelligence, yet these two notions are not equivalent but intertwined. Machine learning is a field of study in computer science that involves building algorithms and models that enable computers to learn from data without being explicitly programmed. It involves using statistical and mathematical techniques to analyze and identify patterns in data, and then using those patterns to make predictions or take actions on new data [28].

Figure 1.10 illustrates the different subsets that constitute AI and figure 1.11 shows how machine learning can help humans learn.

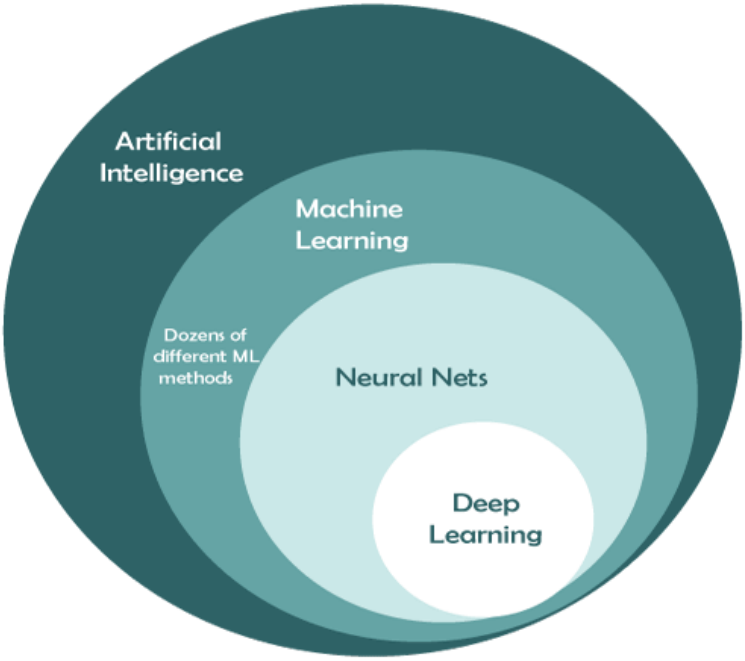

Figure 1.10: Diagram of the components of artificial intelligence.

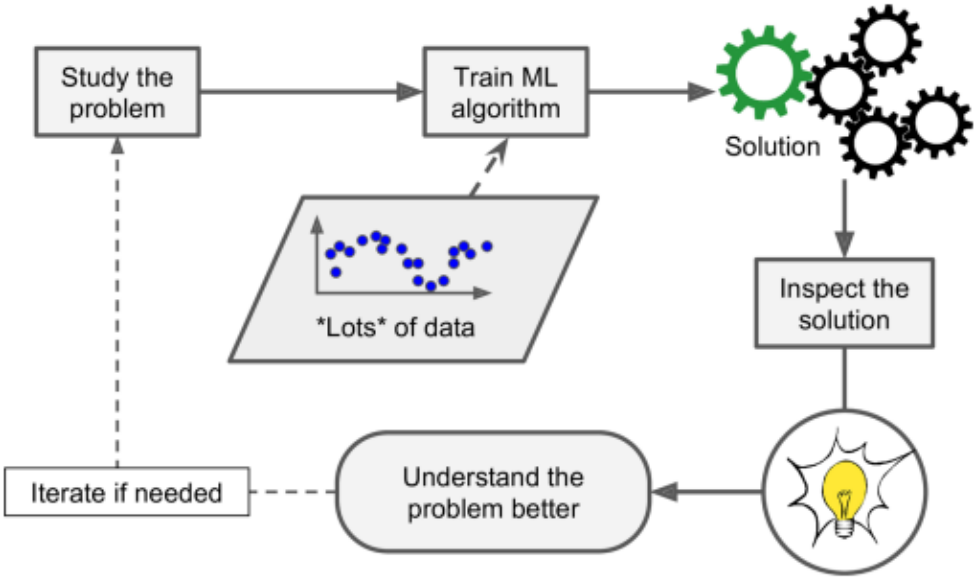

Figure 1.11: Machine Learning can help humans learn [14].

1.2.2 Categories

There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning.

The figure 1.12 shows how machine learning can help humans learn.

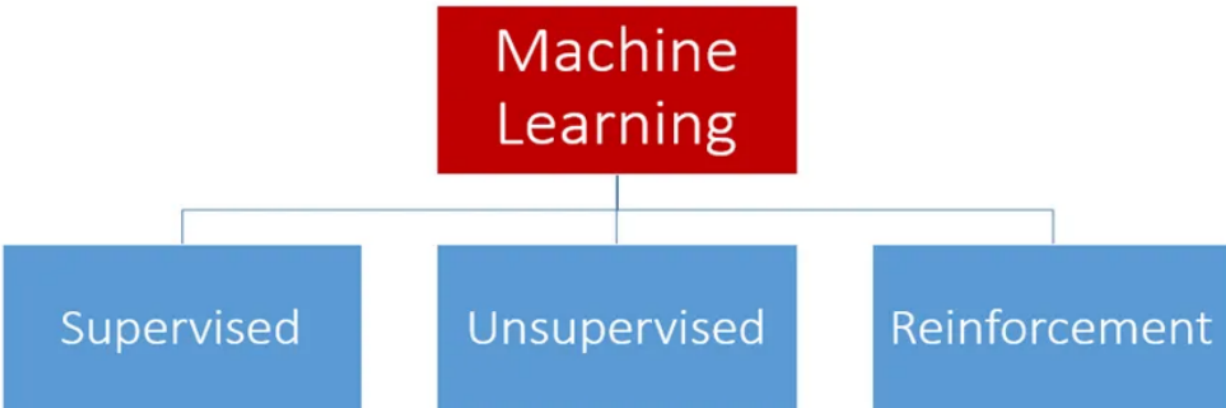

Figure 1.12: Types of Machine Learning.

Supervised Learning :

Supervised learning involves training the algorithm on wanted solutions, called the labels, in order to make an accurate prediction. The main supervised learning tasks are Classification and Regression[14].

Classification Algorithms :

Here are some of the most important supervised learning algorithms (that we might use in our case):

- k-Nearest Neighbors.
- Logistic Regression.
- Support Vector Machines (SVMs).
- Decision Trees and Random Forests.

K nearest Neighbors :

In the k-nearest neighbor or K-nn method, the trained dataset is made up of (input, output) pairs. To rank a new input x, the K-nn accounts for the k nearest neighbor examples of the new entry x and then provides an outcome. The major classification of the k-nearest neighbors is formulated as follows [14]:

$$d(Y, X) = \sqrt{\sum_{i=1}^{n} (x(i) - y(i))^2} \tag{1.1}$$

- X : data vector of the new entry.
- Y : data vector from the training data.

Logistic Regression :

Logistic regression forecasts the outcome of a dependent variable that is categorical. Therefore, the output must be a discrete or categorical value. It can be Yes of No, 0 or 1, True or False, etc. But rather than giving the exact value of 0 and 1, it gives the probabilistic values that fall between 0 and 1 and that are computed by the Sigmoid function as shown below [14]:

$$p(x(i)) = \frac{1}{1 + e^{-x(i)}} \tag{1.2}$$

The figure 1.13 shows that if a probability is over, or under, a certain threshold it then falls into one or the other category.

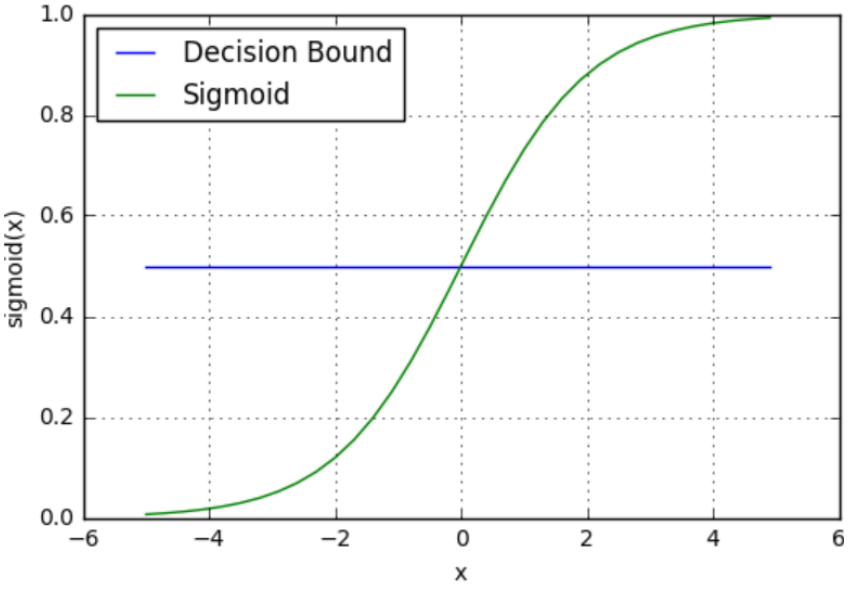

Figure 1.13: The sigmoid function [28].

Support Vector Machines :

The primary goal of the SVM algorithm is to identify a hyperplane that separately divides the data points of the various categories. The hyperplane is localized so that the largest margin is between the classes being considered [14].

As seen in the below figure, the margin refers to the maximum width of the slice that runs parallel to the hyperplane without any internal support vector.

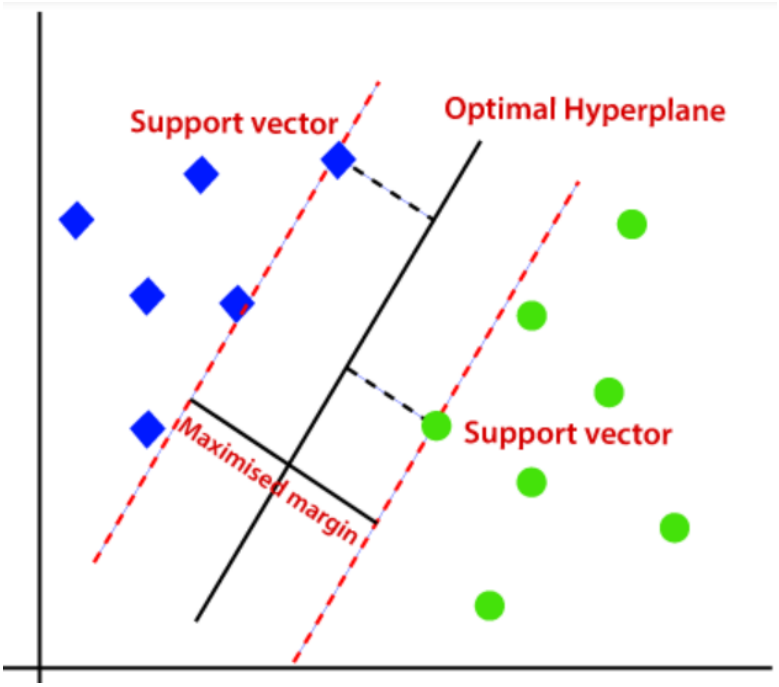

Figure 1.14: SVMs Optimize Margin Between Support Vectors [14].

Decision Trees and Random Forest:

Decision trees are a kind of model that is used for both Classification and Regression. Trees respond to a sequential set of questions that send us on a particular route in the tree based on the response. The model behaves with "if this than that" conditions, which ultimately result in a particular outcome [28].

Random Forest is an ensemble method that uses a combination of multiple decision trees to achieve more accuracy in predictions. In Random Forest, each decision tree is built using a random subset of the features and a random subset of the data for training [28].

The figure 1.15 figure gives an example how Decision Tree algorithm works:

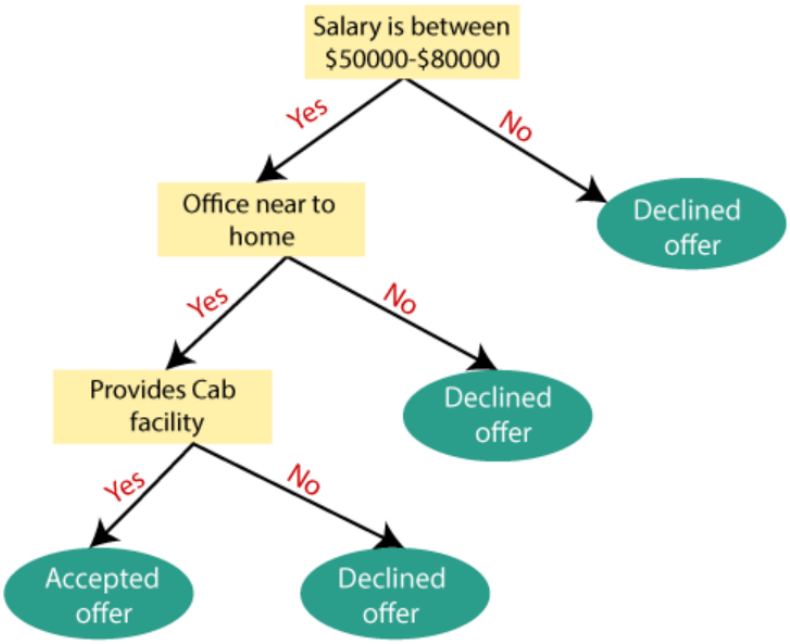

Figure 1.15: Decision Tree real-world example [14].

Unsupervised Learning :

Unsupervised learning is a particular type of machine learning whose objective is to find structures or patterns in the dataset without having any prior knowledge of what the outcome must be [14].

Clustering Algorithms:

Here are some of the most important unsupervised learning algorithms used in clustering task:

- k-Means Clustering
- Hierarchical Cluster Analysis (HCA)
- Expectation Maximization

Reinforcement Learning :

Reinforcement learning concentrates on training agents to make good choices in a realistic and dynamic environment. The main objective of the Software Agent is to learn to act in a way that will maximize its expected rewards in the long term [14].

Control Algorithms :

Here are some of the most important reinforcement learning algorithms used in control task:

- Q-Learning
- Actor-Critic Methods
- SARSA(State–Action–Reward–State–Action)

The table below can be used as a useful summary and a reference in choosing the appropriate machine learning technique for a particular task [14].

Summary :

Category	Task	Algorithms
Reinforcement Learning	Control (e.g., robotic arm control)	Q-learning, SARSA, Actor-Critic
Unsupervised Learning	Clustering (e.g., customer segmentation)	K-means, DBSCAN, Hierarchical clustering
Supervised Learning	Classification (e.g., image classification)	Decision trees, SVM, Neural networks

Table 1.1: Task-Based Comparison of Reinforcement, Unsupervised, and Supervised Learning Techniques

1.2.3 Advantages and Inconvenience of each classification algorithm:

The table (1.2) compares different classification algorithms in terms of their advantages and disadvantages. It was helpful for choosing the most suitable algorithm for a given classification problem.

Algorithm	Advantages	Disadvantages
Support Vector Machines	Effective in high dimensional spaces ,can handle non-linear decision boundaries and can be used for both classification and regression	Training can be time-consuming and complex ,can be sensitive to the choice of kernel function and not suitable for large datasets
Logistic Regression	Can handle non-linear decision boundaries with polynomial features and can be used for classification tasks	Assumes a linear relationship between features and target variable , not suitable for many features and Can be affected by outliers
k-Nearest Neighbors	Simple and easy to implement and can handle non-linear decision boundaries	-Can be computationally expensive during testing ,sensitive to the choice of k value and Performance can be affected by irrelevant or noisy features
Decision Trees and Random Forest	Easy to interpret and visualize and can give good results even with very large datasets	-Can overfit on noisy or complex data , performance can be affected by irrelevant features and Random Forest can be computationally expensive during training
K-Means Clustering	Simple and fast clustering algorithm that can handle large datasets	Requires the number of clusters to be specified and can be sensitive to initial centroid placement
Hierarchical Clustering	Does not require the number of clusters to be specified and can create a dendrogram to visualize the clustering structure	Can be computationally expensive and can create long chains of clusters in certain situations
Principal Component Analysis	Can reduce the dimensionality of large datasets while preserving the most important information	May not be suitable for nonlinear relationships between variables and can result in loss of interpretability

Table 1.2: Comparison of different classification algorithms for both supervised and unsupervised learning

1.2.4 Performance Metrics in Machine Learning

Classification models have discrete results, so we need measures that make comparisons between discrete classes in some form or another. The classification metrics assess the behavior of a classification model and don't tell you whether the classification is right or wrong, but each metric evaluates it in a unique way [10].

Here are some fundamental and important classification metrics to consider [10] :

- Confusion Matrix (Fundamental to others).
- Accuracy.
- Precision and Recall.
- F1-score.

Confusion Matrix :

Confusion Matrix is a tabular visualization of the ground-truth labels versus model predictions. Each row of the confusion matrix represents the instances in a predicted class and each column represents the instances in an actual class [10].

The figure 1.16 shows an example of how confusion matrix is build.

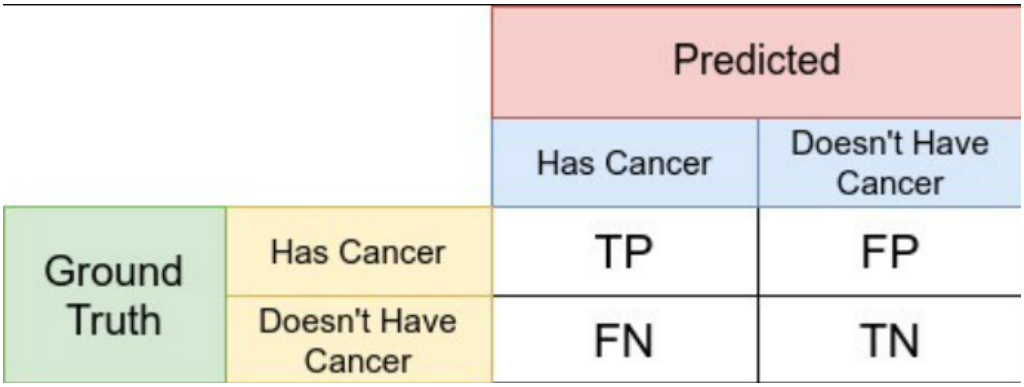

Figure 1.16: Confusion matrix for the Hypothesis:"the individual has cancer".

- TP: are the number of correct positive predictions, the cases where the model predicted a positive outcome and the actual outcome was indeed positive.
- TN: are the number of correct negative predictions, the cases where the model predicted a negative outcome and the actual outcome was indeed negative.
- FP: are the number of incorrect positive predictions, the cases where the model predicted a positive outcome, but the actual outcome was negative.
- FN: are the number of incorrect negative predictions, the cases where the model predicted a negative outcome, but the actual outcome was positive.

Accuracy:

Classification accuracy is the simplest metric to use and implement and is defined as the number of correct predictions divided by the total number of predictions [10] :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.3}$$

Precision and Recall:

Recall is a performance metric that measures the ratio of true positives to the total number of actual positives in the ground truth [10] :

$$Recall = \frac{TP}{TP + FN} \tag{1.4}$$

Precision is the ratio of true positives and total positives predicted [10] :

$$Precision = \frac{TP}{TP + FP} \tag{1.5}$$

F1 score:

The F1-score metric uses a combination of precision and recall[10].

A high F1 score symbolizes a high precision as well as high recall. It presents a good balance between precision and recall and gives good results on classification problems.[10]

$$F1score = \frac{2}{\frac{1}{precision} + \frac{1}{Recall}} \tag{1.6}$$

1.2.5 Overview of Machine Learning in Healthcare :

Machine learning is used in many medical diagnostic and treatment applications. Here are some examples of machine learning applications in this field:

- **Diagnosis:** Machine learning algorithms can be used to analyze clinical data such as symptoms, lab tests, and diagnostic images (X-rays, MRIs, ultrasounds, etc.) to help physicians make more accurate and timely diagnoses. Machine learning models can also be used to predict patients' risk of disease, using risk factors such as age, medical history, lifestyle habits, family history and genetic information.
- **Cancer Detection:** Machine learning algorithms can be used to analyze mammogram, CT scan and other medical images to detect signs of cancer. Machine learning models can also be used to predict the risk of recurrence and to help develop personalized treatment plans.
- **Personalized Treatment:** Machine learning models can be used to develop personalized treatment plans for patients using clinical data, medical history, lifestyle habits and genetic information. Machine learning models can also be used to predict patient response to different treatments and to optimize drug doses.

- **Detection of Rare Diseases:** Machine learning algorithms can be used to help diagnose rare diseases by analyzing patients' clinical and genetic data. Machine learning models can also be used to identify patients who may benefit from experimental treatments.
- **Health Monitoring:** Machine learning algorithms can be used to monitor the health status of patients, using wearable sensors, health tracking devices and other technologies. Machine learning models can also be used to predict potential complications and help healthcare professionals intervene earlier.

In sum, machine learning is a powerful tool for improving medical diagnosis and treatment. Using clinical data, diagnostic images, genetic information, and other data sources, machine learning models can help medical professionals make more accurate and timely diagnoses, develop personalized treatment plans, and monitor patients' health status more effectively.

Conclusion

In conclusion, this preliminary study has provided a foundation for understanding the key components of our breast cancer prediction project. We have discussed the medical background of breast cancer and discuss its various types, the complex process of cancer detection, and the potential for machine learning to enhance diagnostic accuracy and improve patient outcomes. Moving forward, we will provide an overview of related works and our specific contributions to the field.

Chapter 2

Overview about Related Works and Contribution

Introduction

As we saw in the previous chapter, breast cancer is a serious and often deadly disease that affects millions of people worldwide.
Therefore, in our research, we found some articles that showcases the potential of machine learning in improving the accuracy and efficiency of breast cancer detection. For this reason, this chapter discusses some of the existing solutions related to "Machine Learning Based Prototype for Diseases Detection", followed by a synthesis of the current state of the field. Finally, a proposed solution will be presented that builds upon the existing work and addresses the gaps and limitations identified. By examining the existing solutions and proposing a novel approach, this chapter aims to contribute to the advancement of machine learning and provide insights for future research and development.

2.1 Some Existing Solutions

In this section, we will review some of the recent research papers and other real projects that have explored the use of machine learning techniques for breast cancer detection.

2.1.1 Research Papers

The articles I will list discuss different approaches to developing machine learning prototypes, including the use of deep learning techniques, ensemble machine learning, and automated detection using machine learning techniques.The authors used various datasets of mammograms to train and test their models, and achieved high levels of accuracy ranging from 90% to 98.1%.

The figure 2.1 illustrates number of publications on AI in breast cancer diagnosis between 2012–2022.

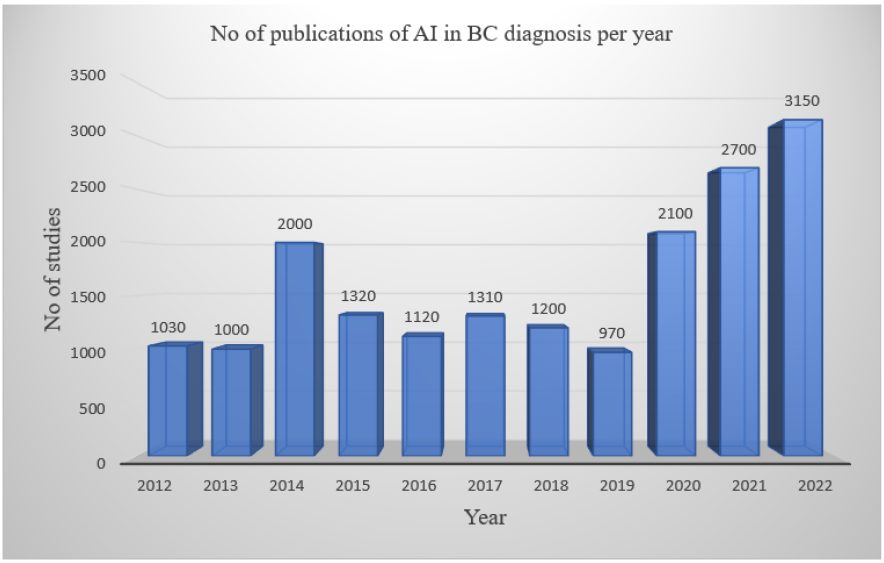

Figure 2.1: Publications on AI in breast cancer diagnosis between 2012–2022 [35].

We will list just a few examples of our research :

- **Breast cancer detection using artificial intelligence techniques: A systematic literature review :** the article is a systematic review of the literature on the use of artificial intelligence techniques for breast cancer detection.

  The authors analysed scientific literature on the use of different artificial intelligence techniques, such as deep learning, neural networks, decision trees and classification methods, for the detection of breast cancer.

  The results of the study showed that the use of artificial intelligence to detect breast cancer is a growing area of research. The authors identified several data sets used for model training, such as the Breast Cancer Wisconsin Diagnostic Data Set (BCWD) and the Digital Database for Screening Mammography (DDSM).

  The results also showed that the performance of artificial intelligence methods varied considerably depending on different factors, such as data quality, data set size, the choice of characteristics and the choice of hyperparameters.

- **An automated computational biomechanics workflow for improving breast cancer diagnosis and treatment :** this article presents an automated workflow for analyzing breast cancer using computational biomechanics.

  The proposed workflow consists of three main stages: image segmentation, finite element modeling, and mechanical analysis. In the first stage, the breast MRI image is segmented to obtain the geometry and material properties of the breast tissue. In the second stage, the finite element model is created to simulate the mechanical behavior of the breast tissue. In the third stage, the mechanical analysis is performed to analyze the deformation and stress distribution of the breast tissue under different loading conditions.

  The authors tested their workflow on a dataset of 20 breast MRI images and compared the results with manual segmentation and analysis. The results showed that the automated workflow was accurate and efficient in analyzing the breast tissue mechanics.

  The authors also demonstrated the potential applications of their workflow in breast cancer diagnosis and treatment. They suggested that the workflow could be used to identify early-stage breast cancer, assess tumor response to treatment, and optimize treatment planning.

  To conclude, the article highlights the potential of using computational biomechanics to improve breast cancer diagnosis and treatment. The proposed workflow may help in the development of personalized treatment plans and improve patient outcomes.

  The figure 2.2 shows the architectural of a prototype of their automated workflow which was implemented at Auckland City Hospital in New Zealand.

- **Breast cancer data classification using ensemble machine learning :** presents a study on the use of ensemble machine learning techniques for breast cancer classification.

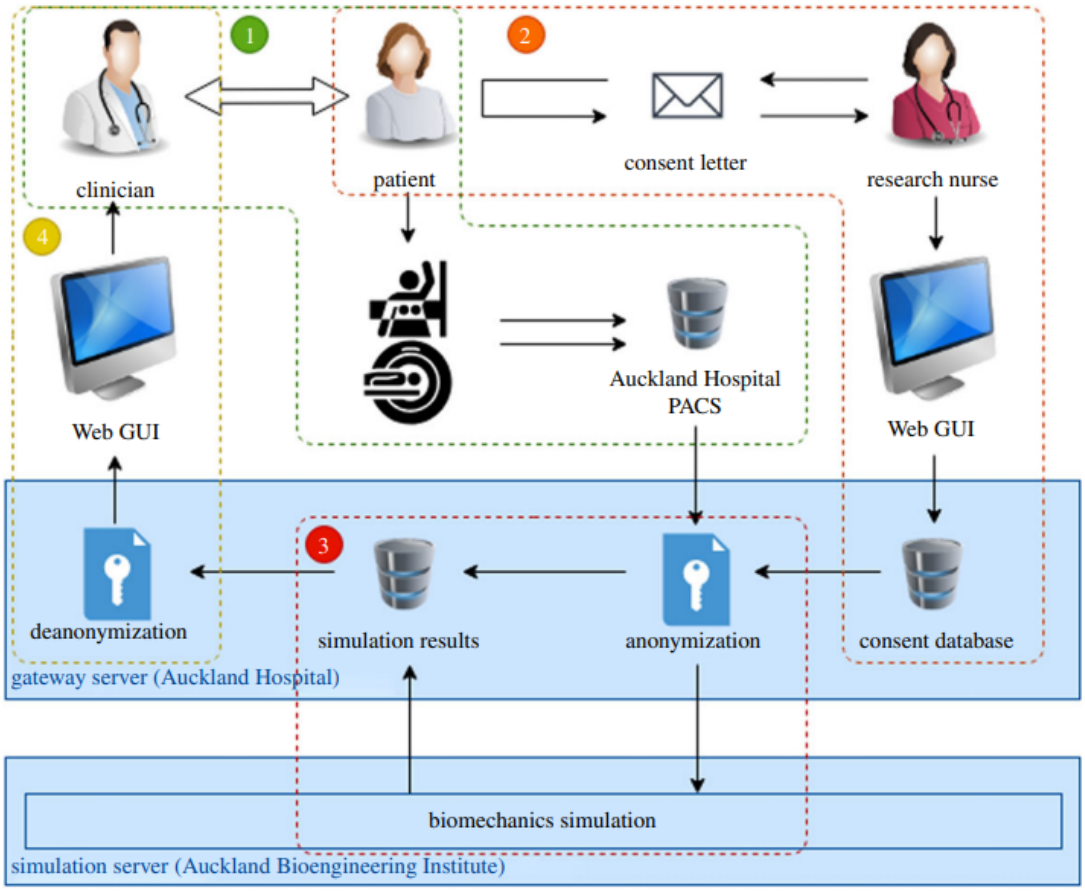

Figure 2.2: Implementation of an automated image analysis workflow at the Auckland City Hospital, illustrating the steps involving: (1) consultation; (2) consent; (3) simulation; (4) presentation of results [4].

The authors used four different machine learning algorithms: k-nearest neighbor (KNN), support vector machine (SVM), decision tree (DT), and random forest (RF) to classify breast cancer data. They also applied three ensemble learning techniques: bagging, boosting, and stacking to improve the accuracy of classification.

The authors used the Wisconsin Breast Cancer Dataset (WBCD), which contains 699 instances with 11 features, to train and test their models. They evaluated the performance of each model using various metrics such as accuracy, precision, recall, and F1 score.

The results showed that the ensemble learning techniques significantly improved the accuracy of the classification models. The best performance was achieved using the stacked ensemble model, which had an accuracy of 97.57%, precision of 97.51%, recall of 97.49%, and F1 score of 97.5%.

Overall, the article demonstrates the effectiveness of using ensemble machine learning techniques for breast cancer classification. The authors suggest that the proposed methodology could be useful in assisting medical professionals in the early detection and diagnosis of breast cancer.

The process map for the approach suggested in this article is illustrated in Figure2.3.

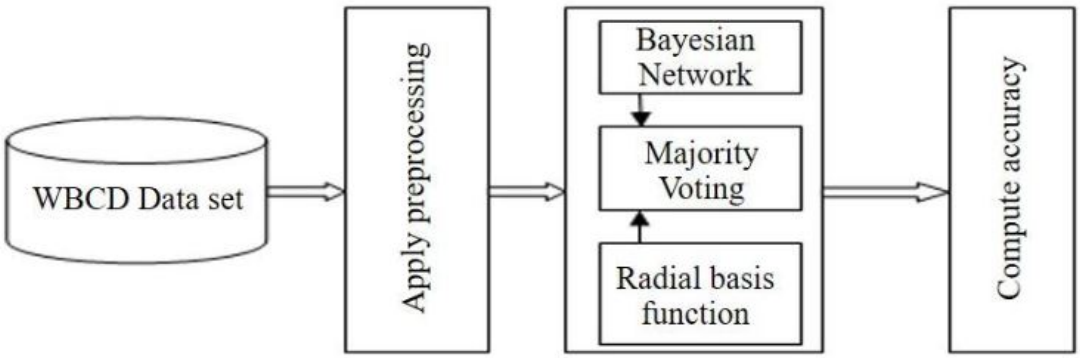

Figure 2.3: Process Flow Diagram of the Proposed Approach [20]

- **A Theoretical Methodology and Prototype Implementation for Detection Segmentation Classification of Digital Mammogram Tumor by Machine Learning and Problem Solving Approach :** proposes a methodology for detecting, segmenting, and classifying tumors in digital mammograms using machine learning and problem-solving approaches.

  The proposed methodology consists of three stages: tumor detection, segmentation, and classification. In the first stage, an adaptive thresholding algorithm is used to identify the possible tumor regions in the mammogram. In the second stage, an edge detection algorithm is applied to segment the tumor region from the background. In the third stage, a feature extraction algorithm is used to extract features from the segmented tumor region, which are then used to classify the tumor as benign or malignant using a machine learning algorithm.

  The authors implemented a prototype of their methodology and tested it on a dataset of 322 mammograms. The results showed that the proposed methodology achieved an accuracy of 93.62% in tumor detection, 91.86% in tumor segmentation, and 93.47% in tumor classification.

  To conclude, the article demonstrates the potential of using machine learning and problem-solving approaches to improve the accuracy of tumor detection, segmentation, and classification in digital mammograms. The proposed methodology and prototype implementation may be useful in assisting radiologists in the early detection and diagnosis of breast cancer.

  The figure 2.4 illustrates the overall block diagram of the method for classifying weights and the figure 2.5 illustrates the Case Base Reasoning Approach.

- **Breast Cancer Detection in Saudi Arabian Women Using Hybrid Machine Learning on Mammographic Images :** presents a study on the application of hybrid machine learning techniques for the early detection of breast cancer in Saudi Arabian women using mammographic images.

  The authors used a dataset of 310 mammograms from Saudi Arabian women to train and test their machine learning model. They used two feature extraction methods: texture features and shape features, to extract features from the mammograms. They then used a hybrid machine learning algorithm, which combines k-nearest

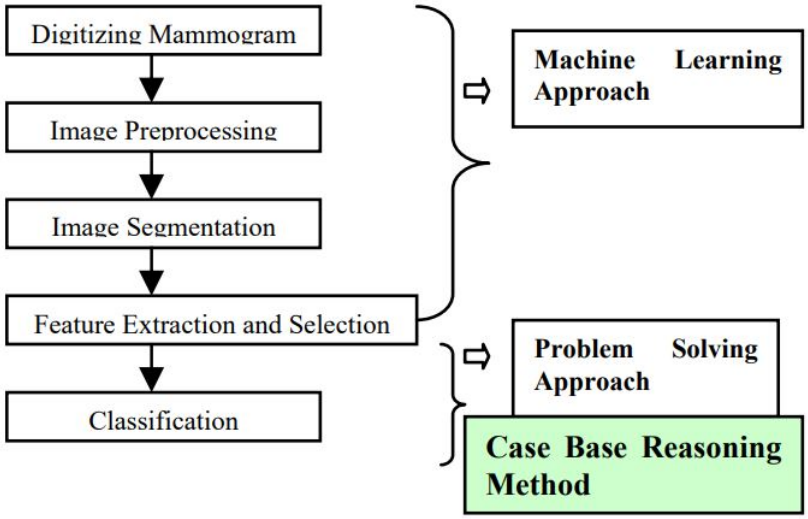

Figure 2.4: The overall block diagram of the method for classifying weights [36]

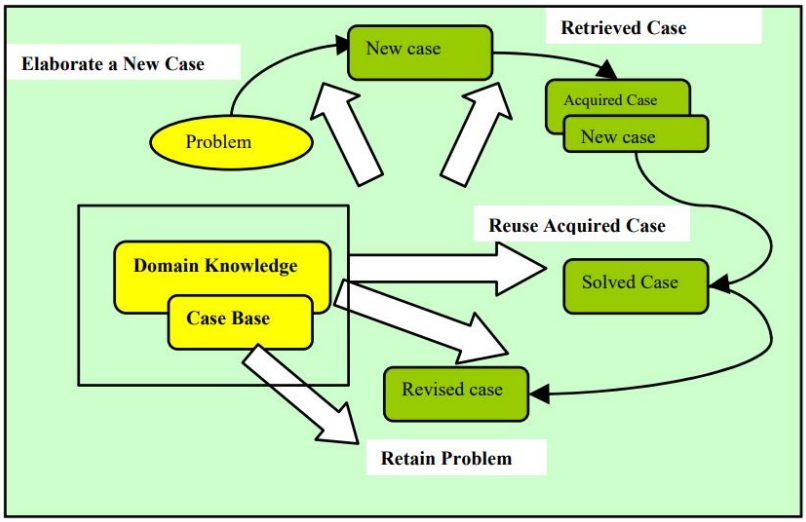

Figure 2.5: The Case Base Reasoning Approach [36]

neighbor (KNN) and artificial neural network (ANN) algorithms, to classify the mammograms as either benign or malignant.

The authors evaluated the performance of their model using various metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC).

The results showed that the hybrid machine learning algorithm outperformed the individual KNN and ANN algorithms in terms of accuracy, sensitivity, specificity, and AUC-ROC. The best performance was achieved using the hybrid algorithm with texture and shape features, which had an accuracy of 98.1%, sensitivity of 97.1%, specificity of 99.0%, and AUC-ROC of 0.998.

Overall, the study demonstrates the potential of using hybrid machine learning techniques and mammographic images for the early detection of breast cancer in Saudi Arabian women. The proposed approach is based on three models (models 1, 2 and 3) to get the results of the whole model. These models have different

structures with different layers of parameters, functionality, learning rates, steps and filter sizes as described in Figure 2.6. This last may help in improving the accuracy of breast cancer diagnosis and assist radiologists in making better clinical decisions.

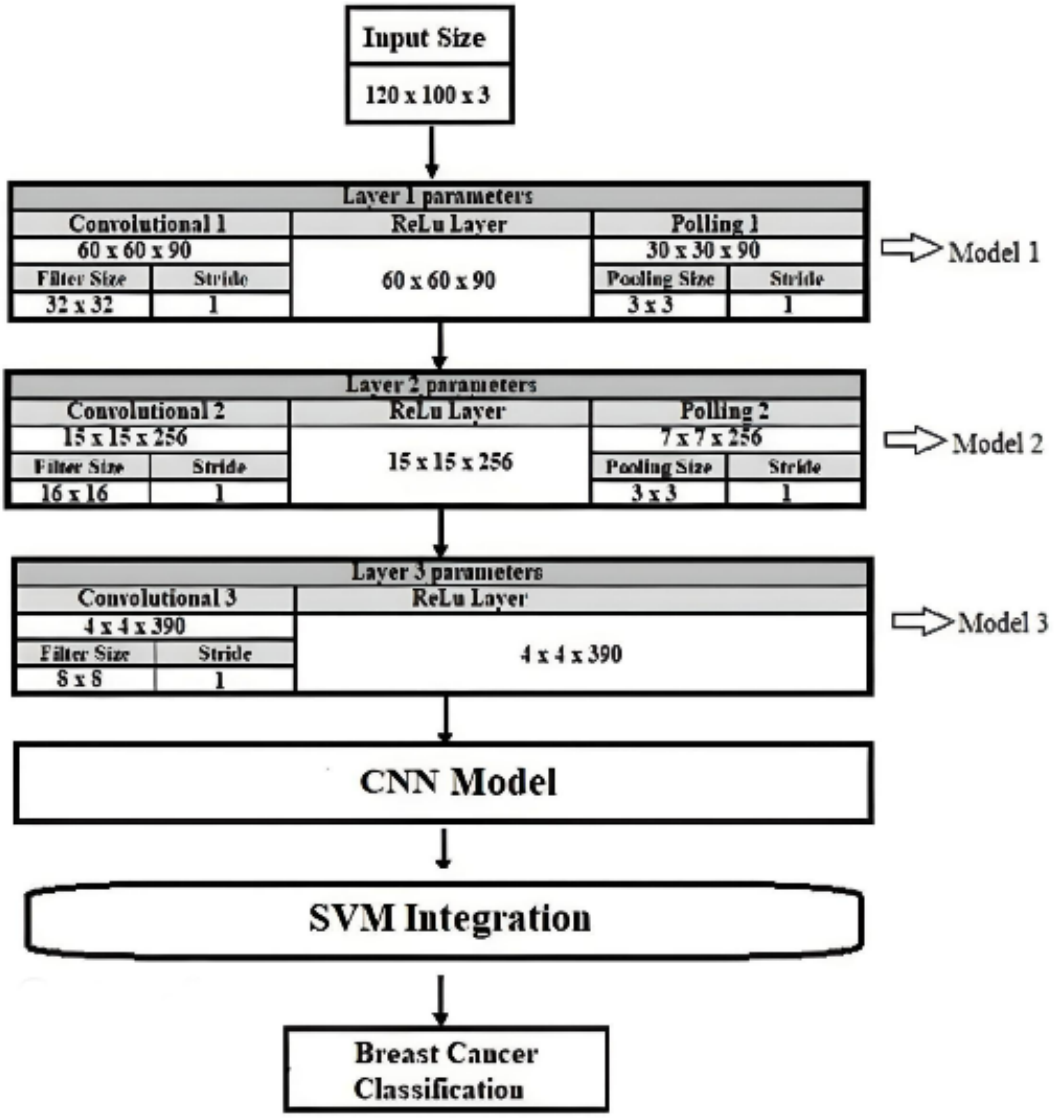

Figure 2.6: Proposed model architecture [2].

2.1.2 Projects

- **Digital Mammography DREAM Challenge:** This Challenge was a collaborative effort among several institutions and research organisations to develop machine learning algorithms for breast cancer detection. The challenge used a large dataset of digital mammograms to train and test the algorithms, and the winning solutions were able to achieve high levels of accuracy in detecting breast cancer.
  The goal of the Digital Mammography (DM) DREAM Challenge was to apply an open science, crowd-sourced approach to develop and assess algorithms for risk stratification of screening mammograms that can be used to improve breast can-

cer detection. These algorithms have the potential to improve the interpretation of other tumour images, affecting a wide range of cancerous patients. The DM Challenge encouraged teams to apply deep learning approaches to a large set of mammographic images of over 640,000 images of 80,000 women. Dozens of teams participated in the Challenge, leading to the development of many innovative approaches to cancer detection and the setting of public standards and benchmarks [5].

- **iCAD SecondLook Digital:** iCAD SecondLook Digital is a computer-aided detection (CAD) system that uses machine learning algorithms to analyze digital mammograms for signs of breast cancer. The system has been approved by the FDA and is being used in clinical practice to help radiologists detect breast cancer at an early stage.
  SecondLook is based on sophisticated patented algorithms that analyze the data, automatically identifying and marking suspicious regions in 2D mammography images. The solution provides the radiologist with a "second look" which helps the radiologist detect actionable missed cancers earlier than screening mammography alone. SecondLook detects and identifies suspicious masses and micro-calcifications utilizing image processing, pattern recognition and artificial intelligence techniques. Information from thousands of mammography images are incorporated into these algorithms enabling the product to distinguish between characteristics of cancerous and normal tissue [18].

- **Breast Cancer Risk Assessment Tool:** The Breast Cancer Risk Assessment Tool, also known as the Gail Model, is a machine learning-based tool used to estimate a woman's risk of developing breast cancer. The tool uses various risk factors such as age, family history, and personal health history to predict a woman's likelihood of developing breast cancer over a certain period of time.
  The breast cancer risk assessment tool allows health professionals to estimate the risk of invasive breast cancer in a woman over the next five years and up to age 90 (lifetime risk).
  The tool uses a woman's personal medical and reproductive history and the history of breast cancer in her first-degree parents (mother, sisters, daughters) estimate the absolute risk of breast cancer — its likelihood of developing invasive breast cancer within a defined age range [32].

- **CancerSEEK:** is a blood test that uses machine learning algorithms to detect multiple types of cancer, including breast cancer. The test analyzes various biomarkers in the blood to identify cancer at an early stage when it is more treatable.
  CancerSEEK involves "the use of a single, non-invasive, multi-analyte test that simultaneously assesses the presence of mutations and eight biomarkers of cancer-associated proteins in the blood," says co-developer Anne Marie Lennon. PCR multiplex analysis of cell-free circulating tumour DNA (TDNA) detects mutations at 2,001 genomic positions in 16 genes, while protein biomarker levels are evaluated using immunoassays [23].

- **Hologic Genius 3D Mammography:** Hologic Genius 3D Mammography is a digital mammography system that uses machine learning algorithms to detect breast

cancer. The system creates 3D images of the breast tissue, which can help improve the accuracy of breast cancer detection compared to traditional 2D mammography. The new 3Dimensions system is designed to provide better 3D images to radiologists, a more comfortable mammography experience for patients and improved workflow for technologists [16].

2.2 Synthesis

The table 2.1 shows The table provides information on the use of machine learning models for tumor classification. The data includes the reference, publication date, classification task, tumor type, number of study images, type of model used, the size of the training set, as well as sensitivity, specificity, and accuracy of the model.The results indicate that multiple models were used with varying training set sizes and tumor types. The models achieved different levels of sensitivity, specificity, and accuracy, but overall, the results suggest that the models were able to perform accurate tumor classifications with levels of accuracy up to 98.2%.Additionally, the results show that the size of the training set appears to play an important role in the performance of machine learning models. In summary, the information provided by the table highlights the potential benefits of using machine learning models for tumor classification, but also underscores the importance of selecting the appropriate training set size to achieve the best results.

The Table 2.2 contains a few questions related to the detection of breast cancer using machine learning techniques. The following table 2.3 provides the reference that answered these questions, along with the corresponding question numbers from the previous table. Overall, the information provided by the two tables highlights the increasing importance of using machine learning techniques for breast cancer detection and classification, while also underscoring the persistent challenges associated with using these techniques in medical practice.

The Table 2.3 presents several machine learning projects along with details on their topic, type of data used, and feedback received. The projects include the Digital Mammography DREAM Challenge, iCAD SecondLook Digital, Breast Cancer Risk Assessment Tool, CancerSEEK, and Hologic Genius 3D Mammography. The types of data used vary across the projects and include mammographic images, breast cancer risk data, blood analysis, and financial market data. The feedback received includes evaluations of model accuracy, recommendations to improve the models, and comments on data quality. The projects cover a wide range of topics, from breast cancer risk analysis to early cancer detection. Overall, the information provided by the table highlights the diversity of machine learning projects applied to cancer detection and prevention, while underscoring the importance of continual improvement of models and data quality.

Reference	Date	Task	Tumor Type	No Images Studies	Type of Model	Size of Input Training	Sensitivity (%)	Specificity (%)	Accuracy (%)
Tomoyuki Fujioka, Kazunori Kubota et al	2020	Differentiation	Breast cancer	576.6	CNN	48 benign, 72 malignant	95.8, 58.3, 91.7	92.5, 60.4, 77.1	92.5, 65.8, 79.2
Wei M	2020	Classification	Breast tumor	1061	SVM classifier	589 malignant, 472 benign	87.0	87.6	87.3
Tomoyuki Fujioka, Kazunori Kubota et al.	2019	Classification	Breast tumor	1536	CNN + Res152	897 malig, 639 benign	90.9	87.0	NA
Kayode et al.	2019	SVM classifier	Breast	322	Texture Feature + SVM	126 Normal, 60 benign, 48 malignant	94.5	91.3	NA
E.P.V Le	2019	classification	Breast imaging	640,000	CNN + DREAM Challenge	318,000	84.0	91	NA
Fujioka et al.	2019	Breast mass segmentation	Breast US	947	DCNN	480 benign, 467 cancerous	NA	NA	NA
Chougrad et al.	2018	Comparative image classification	Mammography + X-ray	6116	DDSM, INbreast, BCDR	641.688DDSM, 300.300INbreast, 344.300BCDR	NA	NA	98.2
Mohammed et al.	2018	classification	Ultrasound Breast Images	1393	ANN classifier	900.300	79.4	84.76	82.0

Table 2.1: Various algorithms of AI applications and their performances in breast cancer diagnosis [35].

Questions	Research questions
Q1	How a machine learning approach can be used to improve the accuracy of breast cancer detection on mammographic images ?
Q2	How does the machine learning approach compare to traditional diagnostic methods for detecting breast cancer ?
Q3	What factors contribute to the success of the machine learning approach for breast cancer detection ?
Q4	Can the proposed machine learning approach be extended to other medical image analysis tasks for different types of cancer detection?
Q5	How does the proposed machine learning approach compare to other machine learning approaches in terms of accuracy, sensitivity and specificity breast cancer detection?
Q6	What are the challenges of breast cancer detection using machine learning ?

Table 2.2: Research questions related to breast cancer detection.

Reference	Q1	Q2	Q3	Q4	Q5	Q6
Breast cancer detection using artificial intelligence techniques: A systematic literature review	✓	✓	✓	✓	✓	✓
An automated computational biomechanics workflow for improving breast cancer diagnosis and treatment	×	×	✓	✓	✓	×
Breast cancer data classification using ensemble machine learning	✓	✓	×	×	✓	✓
A Theoretical Methodology and Prototype Implementation for Detection Segmentation Classification of Digital Mammogram Tumor by Machine Learning and Problem Solving Approach	✓	✓	×	✓	✓	×
Breast Cancer Detection in Saudi Arabian Women Using Hybrid Machine Learning on Mammographic Images	×	✓	×	×	✓	✓

Table 2.3: Comparative analysis of survey papers on breast cancer detection in terms of research questions

Projects	Topic	Type of Data
Digital Mammography DREAM Challenge	Analysis of digital mammograms, identifying new approaches for detecting early-stage breast cancer and reducing the number of false positives.	Large,high-quality datasets of digital mammograms,relevant clinical and demographic information.
iCAD SecondLook Digital	Developing and implementing advanced image analysis algorithms, identify potential areas of concern in digital mammograms, reduce false negatives and improve diagnostic accuracy.	Large, high-quality datasets of digital mammograms,relevant clinical and demographic information,including previous mammograms and biopsies.
Breast Cancer Risk Assessment Tool	Analyzing a range of risk factors, including age, family history, reproductive history, lifestyle factors, identify patterns and relationships between these factors and the likelihood of developing breast cancer.	large datasets , including medical histories, family histories, lifestyle factors, genetic data and other biomarkers.
CancerSEEK	Develop a blood test that can detect multiple types of cancer at an early stage.	Large datasets of patient blood samples,genetic data and other biomarkers.
Hologic Genius 3D Mammography	Analyze 3D mammography images, identify potential tumors and abnormalities that may not be visible on traditional 2D mammography images	Large datasets of 3D mammography images

Table 2.4: Comparative analysis of projects

2.3 BreastCancerCheck - ML based Web Platform for Cancer Detection

We suggest a Streamlit web application that helps doctors and make it simple for patients to predict the likelihood of breast cancer based on a number of risk factors.Four pages are defined in our app:

- **The "About Us" page:** Provides information about the developers of the app, our passion for improving healthcare using machine learning and latest technology, and the purpose of the project.
- **The "Breast Cancer Sensitization" page:** Includes a video that covers the

basics of breast cancer, including the signs and symptoms to look out for, how to perform a self-examination, and the importance of regular mammograms

- **The "Predict for Doctors" page:** Allows doctors to input tumor information for various features and get a prediction of whether the tumor is malignant or benign.The doctor can input the tumor information for the specified features.
- **The "Predict for Patients" page:**The user is prompted to input personal information and blood test results, which are used to make a prediction about the likelihood of breast cancer.If the prediction is positive, the user is advised to contact a doctor for further evaluation and guidance.

Conclusion

In this chapter, we explored some existing solutions for breast cancer detection and discussed their strengths and limitations. We also presented our contribution to this field, the BreastCancerCheck platform, which is a machine learning-based web application that aims to assist in early breast cancer detection.

After reviewing various solutions, we created a summary table that compares and contrasts the features of the existing solutions. We believe that our proposed platform, BreastCancerCheck, stands out from the others in terms of its simplicity and ease of use. Our platform provides a user-friendly interface that allows patients to input their personal and blood test information and receive an accurate prediction of their breast cancer risk.

In summary, the existing solutions for breast cancer detection vary in terms of their approach and features. Our BreastCancerCheck platform aims to fill the gap in this field by providing a easy-to-use solution for early breast cancer detection.

Chapter 3

Modeling and Implementation of the ML Component

Introduction

The ability to accurately detect and diagnose diseases is of critical importance in healthcare. Machine learning has emerged as a powerful tool for disease detection, and has the potential to significantly improve diagnostic accuracy and efficiency. In this chapter, we focus on the modeling and implementation of the ML component in our prototype for breast cancer detection. The chapter is divided into three main sections: Choices, Modeling, Implementation and Evaluation. In the Choices section, we discuss the key decisions that were made during the design phase, including the selection of appropriate algorithms . In the Modeling section, we describe the process of how treated data from data analysis to the model training and finally model deployment . Finally, in the Implementation and Evaluation section, we explain how the code for machine learning works in the overall system architecture, also we assess the performance of the machine learning model , both in terms of its accuracy and its computational efficiency.

3.1 Choices

To implement the ML component , we used two datasets: one containing personal data like Age and several blood test results about individuals, including whether they are breast cancer patients or not, and the other containing extracted from medical MRI images about made for doctors, indicating whether a tumor is malignant or benign.

Based on the labeled data (0 for malignant and 1 for benign in the doctor dataset, and 1 for healthy and 2 for patients in the patient dataset),we need supervised learning algorithms so we trained four algorithms from the desired category on our data: kNN, SVM, Logistic Regression, and Random Forest (or Gradient Boosting)

3.2 Modeling

3.2.1 General Design

For each dataset, we began the machine learning process by importing the dataset into our coding environment. After this, we performed exploratory data analysis to understand the relationships between the different columns in the 2 datasets. Then, we trained four supervised learning models on the data to identify which model performs best at classifying breast tumors.

To choose the best model we calculated metrics including accuracy which allowed us to evaluate the models' performance. Based on these metrics, we selected the model that provided the most accurate predictions for the 2 datasets and we deployed the best model.

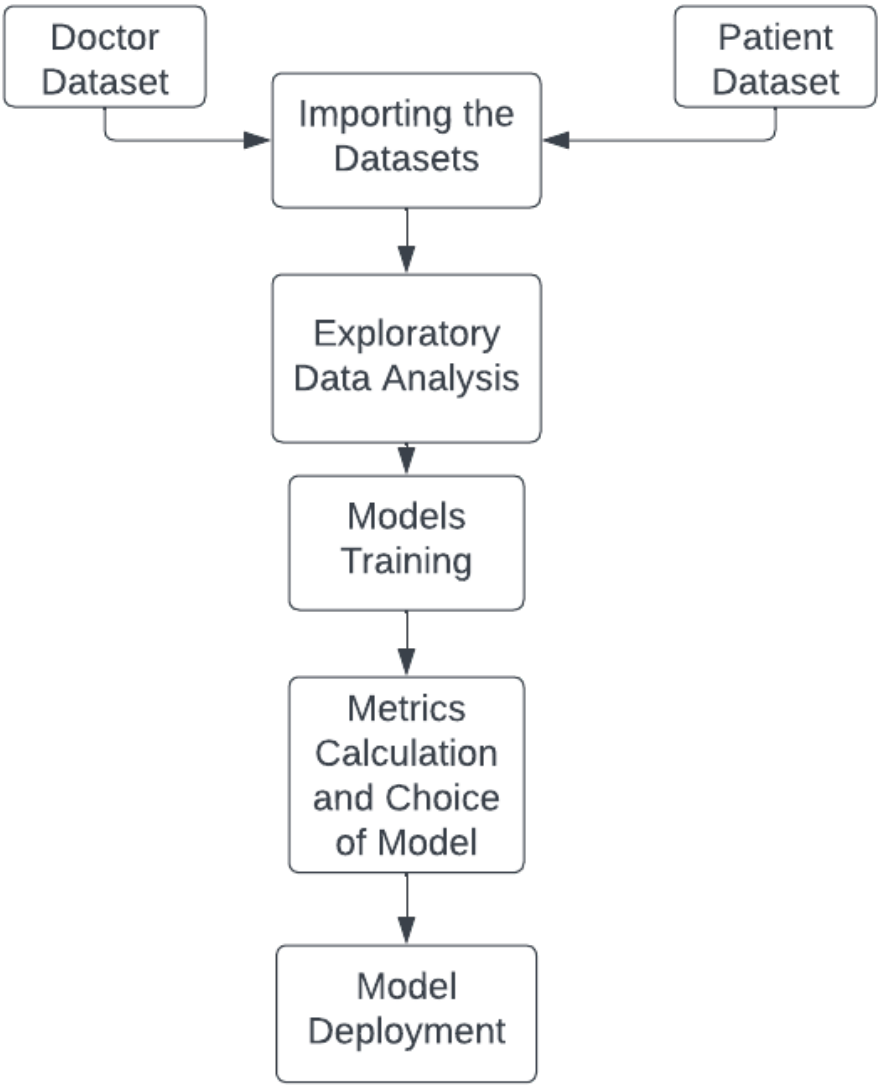

Figure 3.1: Phases of our ML Component

3.2.2 Specific Details about Model Selection

To select the best model, we evaluated each algorithm's accuracy and computational time. After evaluating the metrics for each model, we chose the one that achieved the highest accuracy while still being computationally efficient. Our model selection process ensures that we have a highly accurate and efficient model that can effectively classify breast cancer cases (Figure 3.2)

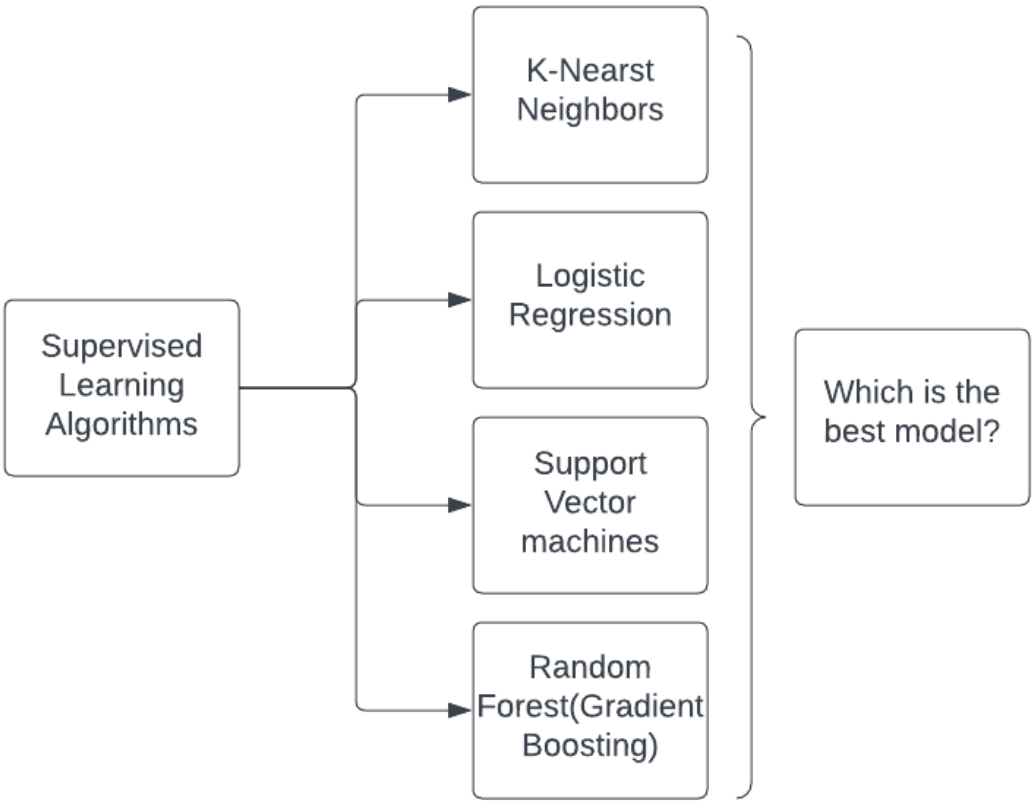

Figure 3.2: Model selection

3.3 Implementation

3.3.1 Hardware and Software Environments

Jupyter Notebook :

For the creation and manipulation of ML component, we chose Jupyter Notebook, which is an interactive development environment for the creation of documents containing code, visualizations and textual explanations.

Jupyter Notebook is particularly useful for data analysis, data manipulation and visualization creation.

Hardware Environment :

In order to carry out the work during the project, we used:

- Processor :Intel i5

- Memory: 16 GB
- Hard drive: 512 SSD
- Operating System: Windows 11

3.3.2 The Datasets

The first dataset used in this project is the Wisconsin Breast Cancer Dataset (WDBC)

The dataset contains :

- 569 samples of breast mass lesions from patients in Wisconsin
- 33 columns (diagnosis , radius mean ,etc...)

The second dataset contains informations about patient and their blood test results to detect breast cancer

3.3.3 Exploratory Data Analysis and Data Pre-processing

Both dataset were processed in a different .ipynb (jupyter file)file The first step to the model realization is to import the dataset. Therefore, we used the python library pandas as shown in figure 3.3 and 3.4

As shown in figure 3.4 we changed one column name so it will be compatible in the web part.

```
df=pd.read_csv("pfa.csv")#header=0 , which means the first row of the CSV
```

Figure 3.3: Importing the Wisconsin dataset

```
df = pd.read_csv("patient.csv",header = 0)#header=0 , which means the
column_names = list(df.columns)
df = df.rename(columns={'MCP.1': 'MCP_1'})
```

Figure 3.4: Importing the patient dataset

To have a clear idea about the data at hand we proceeded at showing the first (and last) five rows of the dataset as shown in figure 3.5 and figure 3.6.

```
df['diagnosis'] = df['diagnosis'].map({'M':1,'B':0})
df.tail()
```

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_me
564	1	21.56	22.39	142.00	1479.0	0.11100	0.115
565	1	20.13	28.25	131.20	1261.0	0.09780	0.103
566	1	16.60	28.08	108.30	858.1	0.08455	0.102
567	1	20.60	29.33	140.10	1265.0	0.11780	0.277
568	0	7.76	24.54	47.92	181.0	0.05263	0.043

5 rows × 31 columns

Figure 3.5: Exploring the last five rows of the Wisconsin dataset

```
df.head()
```

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP_1	Classification
0	48	23.500000	70	2.707	0.467409	8.8071	9.702400	7.99585	417.114	1
1	83	20.690495	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	1
2	82	23.124670	91	4.498	1.009651	17.9393	22.432040	9.27715	554.697	1
3	68	21.367521	77	3.226	0.612725	9.8827	7.169560	12.76600	928.220	1
4	86	21.111111	92	3.549	0.805386	6.6994	4.819240	10.57635	773.920	1

Figure 3.6: Exploring the first five rows of the patient dataset

Next we generated a summary statistics of both datasets, including count, mean, standard deviation, minimum and maximum values(figure 3.7 and figure 3.8).

```
df.describe()
```

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.
mean	0.372583	14.127292	19.289649	91.969033	654.889104	0.096360	0.
std	0.483918	3.524049	4.301036	24.298981	351.914129	0.014064	0.
min	0.000000	6.981000	9.710000	43.790000	143.500000	0.052630	0.
25%	0.000000	11.700000	16.170000	75.170000	420.300000	0.086370	0.
50%	0.000000	13.370000	18.840000	86.240000	551.100000	0.095870	0.
75%	1.000000	15.780000	21.800000	104.100000	782.700000	0.105300	0.
max	1.000000	28.110000	39.280000	188.500000	2501.000000	0.163400	0.

Figure 3.7: Exploring the first dataset statistics

```
df.describe()
```

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP_1	Class
count	116.000000	116.000000	116.000000	116.000000	116.000000	116.000000	116.000000	116.000000	116.000000	116
mean	57.301724	27.582111	97.793103	10.012086	2.694988	26.615080	10.180874	14.725966	534.647000	1
std	16.112766	5.020136	22.525162	10.067768	3.642043	19.183294	6.843341	12.390646	345.912663	0
min	24.000000	18.370000	60.000000	2.432000	0.467409	4.311000	1.656020	3.210000	45.843000	1
25%	45.000000	22.973205	85.750000	4.359250	0.917966	12.313675	5.474283	6.881763	269.978250	1
50%	56.000000	27.662416	92.000000	5.924500	1.380939	20.271000	8.352692	10.827740	471.322500	2
75%	71.000000	31.241442	102.000000	11.189250	2.857787	37.378300	11.815970	17.755207	700.085000	2
max	89.000000	38.578759	201.000000	58.460000	25.050342	90.280000	38.040000	82.100000	1698.440000	2

Figure 3.8: Exploring the second dataset statistics

Next, we created a histogram of the 'diagnosis' and the 'classification' columns in the datasets to calculate the frequency of malignant and benign tumors , healthy and patient individuals and compare them as shown in figure 3.9 and 3.10.

```
df.describe()
plt.hist(df['diagnosis'])
plt.title('Diagnosis (M=1 , B=0)')
plt.show()
```

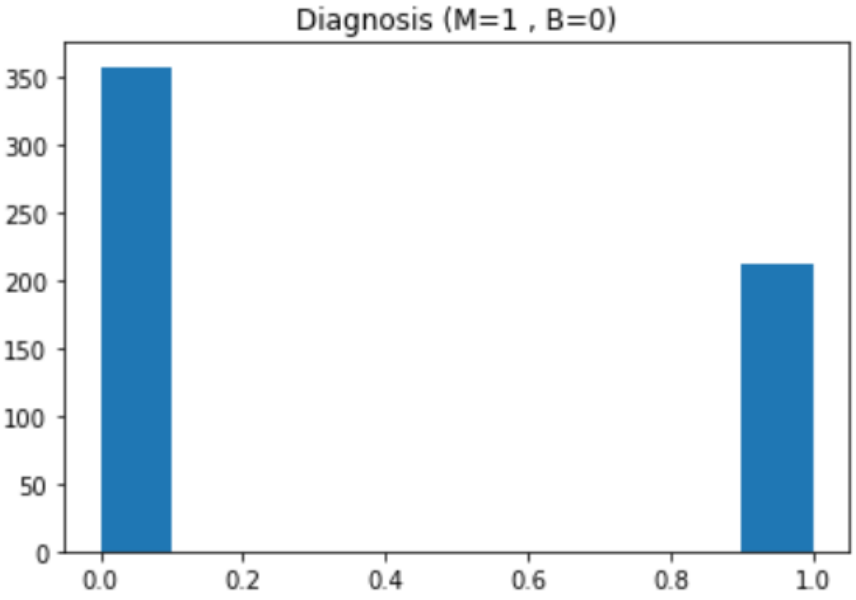

Figure 3.9: The frequency of both tumors types.

```
df.describe()
plt.hist(df['Classification'])
plt.title('Classification (Patients=2 ,Healthy=1)')
plt.show()
```

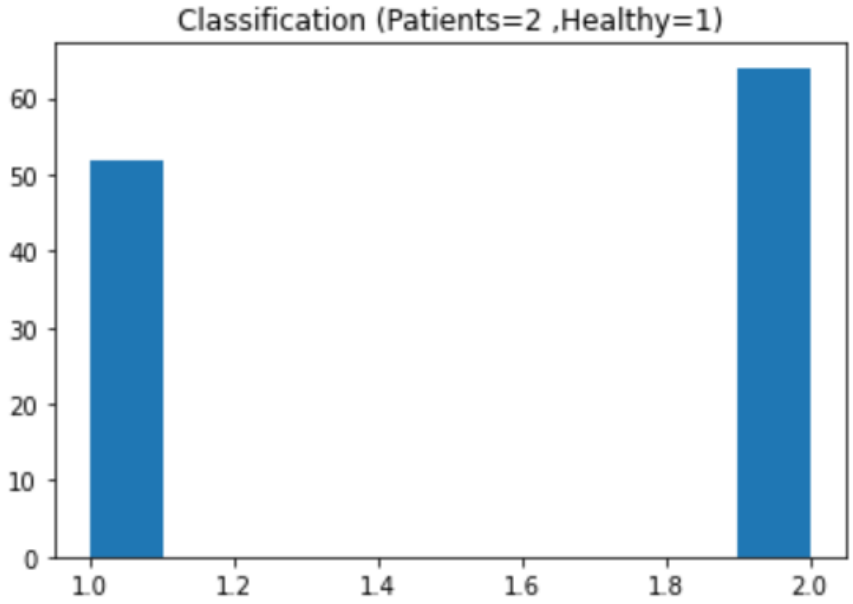

Figure 3.10: The frequency of healthy and patient individuals.

Our next step is to create mini histograms for each feature in the feature mean list and shows the distribution of the values for the two classes of the datasets with plot a stacked histogram of the mean features in the datasets (figure 3.11 and 3.12).

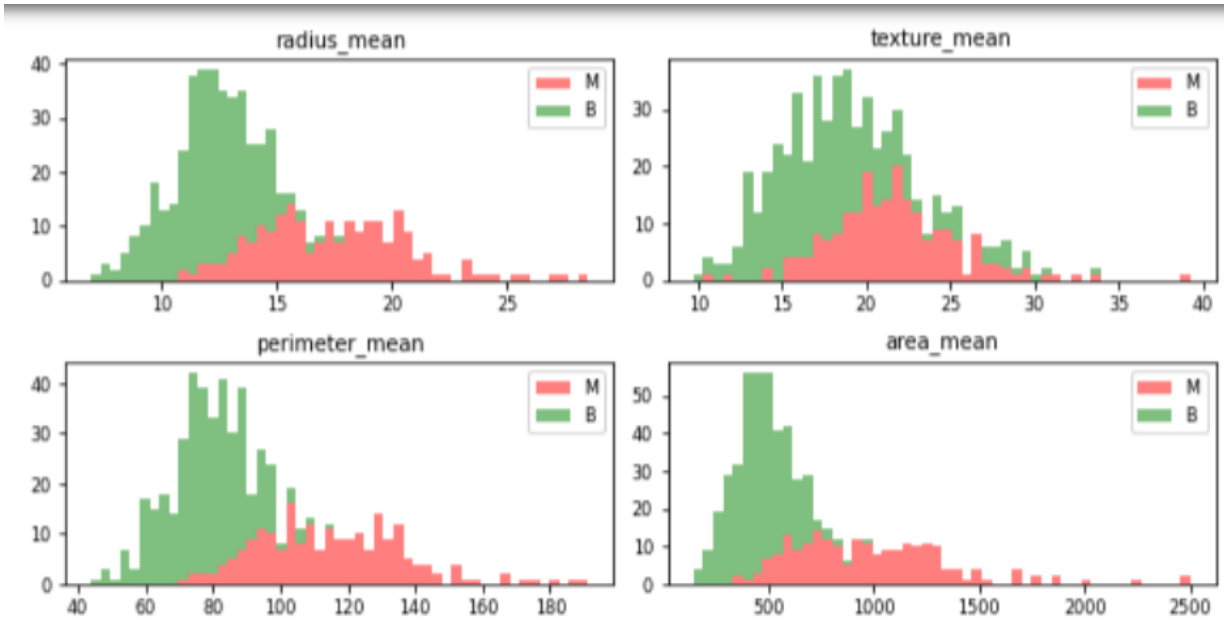

Figure 3.11: The first dataset histograms

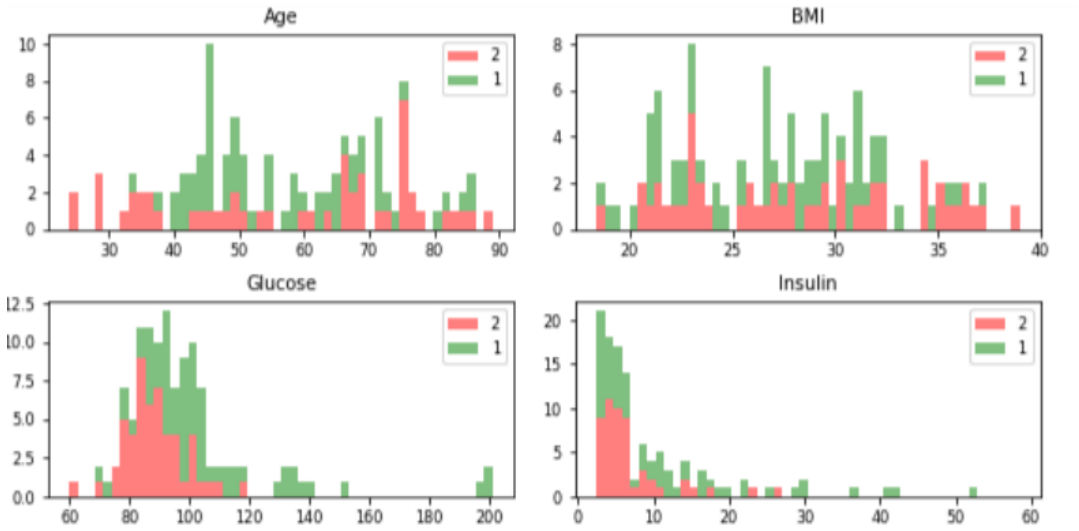

Figure 3.12: The second dataset histograms.

Our next step is to visualise the missing values using missingno python library (each white hyphen is a missing value) as shown in figures 3.13 and 3.14.

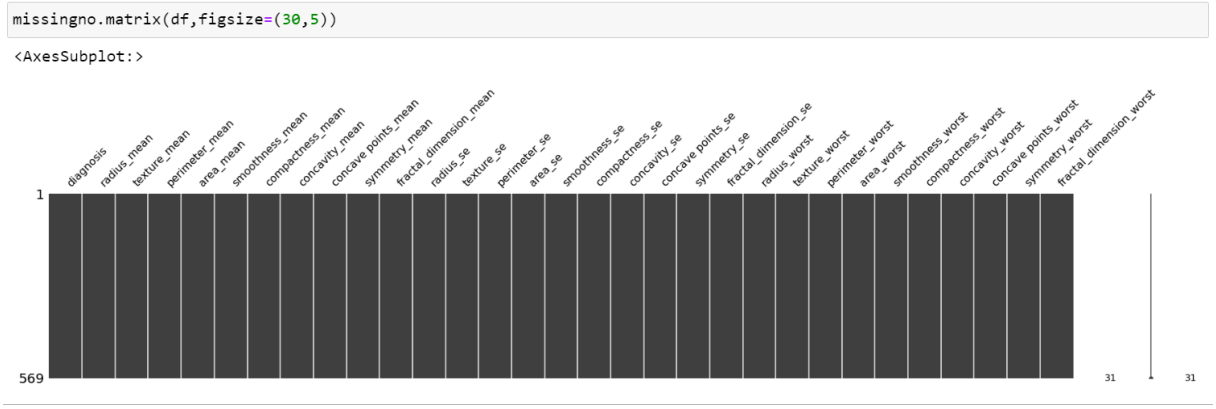

Figure 3.13: The Wisconsin dataset missing values (no white hyphen = no missing values).

Figure 3.14: The second dataset missing values (no white hyphen = no missing values)

Our last step is to create a training data for the model by separating the datasets into inputs and outputs as shown in figure 3.15 and 3.16

```
x_train=df.drop('Classification',axis=1)
y_train=df.Classification
```

Figure 3.15: creating a training set for the second dataset

```
x_train=df.drop('diagnosis',axis=1)
y_train=df.diagnosis
```

Figure 3.16: creating a training set for the Wisconsin dataset

3.3.4 Models Training and Selection

The first step in the model training code is to define a function that fits a machine learning algorithm to the training datasets , performs a cross-validation on it, and returns the training predictions, accuracy on the training set, and accuracy on the cross-validation set (figure 3.17)

```
def fit_ml_algo(algo,x_train,y_train,cv):
  #One Pass
  model=algo.fit(x_train,y_train)
  acc=round(model.score(x_train,y_train)*100,2)
  #Cross Validation
  train_pred=sklearn.model_selection.cross_val_predict(algo,x_train,y_train,cv=cv )
  #Cross Validation accuarcy metric
  acc_cv=round(metrics.accuracy_score(y_train,train_pred)*100,2)
  return train_pred ,acc, acc_cv
```

Figure 3.17: The model training function

Next step is to fit the models mentioned in the modeling part : Logistic Regression ,SVM, KNN , Gradient Boosting (or Descision Tree) into the training data and to return the predicted values, accuracy score for training set and accuracy score for cross-validation set and we visualise the results as shown in the next figures in the Logistic Regression example (figure 3.18 , 3.19 and 3.20).

```
#Logistic Regression
start_time=time.time()
train_pred_log,acc_log,acc_cv_log=fit_ml_algo(LogisticRegression(),x_train,y_train,10)
log_time=time.time()-start_time
print('Accuracy (percent)',acc_log)
print('Accuracy for cross validation(percent)',acc_cv_log)
print('Runing Time',datetime.timedelta(seconds=log_time))
```

Figure 3.18: Logistic Regression training

```
Accuracy (percent) 78.45
Accuracy for cross validation(percent) 64.66
Runing Time 0:00:00.525859
```

Figure 3.19: Results for the patient dataset

```
Accuracy (percent) 78.45
Accuracy for cross validation(percent) 64.66
Runing Time 0:00:00.525859
```

Figure 3.20: Results for the Wisconsin dataset

The same method is used for the other ML models . The figure 3.21 shows the Random Forest algorithm training on the Wisconsin dataset and its results

```
start_time=time.time()
train_pred_dt,acc_dt,acc_cv_dt=fit_ml_algo(RandomForestClassifier(),x_train,y_train,10)
log_time=time.time()-start_time
print('Accuracy (percent)',acc_dt)
print('Accuracy for cross validation(percent)',acc_cv_dt)
print('Runing Time',datetime.timedelta(seconds=log_time))
```

```
Accuracy (percent) 100.0
Accuracy for cross validation(percent) 71.55
Runing Time 0:00:03.175891
```

Figure 3.21: Random Forest training and the results

Our next objective is to create a DataFrame with the scores of different machine learning models in accuracy, and sort it in descending order based on the scores.

```
models=pd.DataFrame({
    'Model':['KNN','Logistic Regression ', 'Linear svc ','Decision Tree '],
    'Score':[acc_knn,acc_log, acc_svc,acc_dt]
})
models.sort_values(by='Score',ascending=False)
```

Figure 3.22: creating a DataFrame for the models accuracy

We did the same method mentioned above with the cross-validation accuracy scores.

	Model	Score
3	Decision Tree	100.00
0	KNN	94.73
1	Logistic Regression	94.73
2	Linear svc	91.04

Figure 3.23: Accuracy scores for the Wisconsin dataset

	Model	Score
3	Random forest	100.00
1	Logistic Regression	78.45
0	KNN	70.69
2	Linear svc	61.21

Figure 3.24: Accuracy scores for the patient dataset

```
models=pd.DataFrame({
    'Model':['KNN','Logistic Regression ', 'Linear svc ','Decis
    'Score':[acc_cv_knn,acc_cv_log, acc_cv_svc,acc_cv_dt]
})
models.sort_values(by='Score',ascending=False)
```

Figure 3.25: Creating a DataFrame for the models cross-validation accuracy

	Model	Score
3	Random forest	71.55
1	Logistic Regression	64.66
2	Linear svc	56.90
0	KNN	47.41

Figure 3.26: Cross validation accuracy scores for the first dataset

	Model	Score
1	Logistic Regression	94.20
0	KNN	92.97
3	Decision Tree	91.56
2	Linear svc	89.81

Figure 3.27: Cross Validation accuracy scores for the second dataset

Based on the results of both cross-validation accuracy and accuracy scores , we noticed that the Random Forest model is overfitting the training data in both datasets (100 % accuracy) so w chose the second in the ranking which is Logistic Regression model and we saved the model in two separate .pkl file using dump function from the pickle library as shown in figure 3.28 and 3.29.

This choice was based on a combination of high accuracy scores and the fact that the model is not overfitting the data.

```
selector=LogisticRegression()
selector.fit(x_train,y_train)
with open('model1.pkl', 'wb') as f:
    pickle.dump(selector, f)
```

Figure 3.28: The chosen model for the Wisconsin dataset

```
selector=LogisticRegression()
selector.fit(x_train,y_train)
with open('model6.pkl', 'wb') as f:
    pickle.dump(selector, f)
```

Figure 3.29: The chosen model for the patient dataset

Conclusion

In conclusion, the modeling and implementation of the ML component in our breast cancer detection prototype has been a rigorous and multi-faceted process. In the Choices section, we discussed the importance of selecting appropriate algorithms and other key design decisions. The Modeling section detailed the various steps involved in the data analysis and model training processes.The implementation part shows how the jupyter notebook code works.

Chapter 4

Modeling and Implementation of the Web Platform

Introduction

During our deep research in this theme. It was noted that the solutions remain theoretical and are not really exploited by patients and doctors. For this, we tried to bring together the latest research and technology for ordinary people in a simpler way.
This chapter provides a comprehensive overview of our web platform modelling and implementation process, detailing the choices, methods and techniques we have used to create a tailored solution that meets our customer's needs.

4.1 Choices

For the creation of our web application, we chose to use Visual Studio Code as an integrated development environment (IDE). VS Code is a smart choice for Python projects because of its excellent Python support, advanced debugging features and integration with tools like Git and GitHub.

When it comes to programming language, Python was chosen for its simplicity, versatility and vast library of modules. Python is also known for its readability and code clarity, which facilitates maintenance and understanding of the code.

For the development of the user interface of your web application, we used the Streamlit library, which is an open source Python library for the creation of web data applications. Streamlit allows you to create interactive web applications with very little code and is ideal for data science and machine learning projects.

In short, the choice of these software, libraries and languages was dictated by their ease of use, their flexibility, their user-friendliness for developers, their adaptability to data science projects and their compatibility with the creation of interactive web applications.

4.2 Modeling

In this section we have defined two UML diagrams:Use case diagram and the Sequence diagram. These UML models are widely used in IT projects to improve understanding of requirements, document system architecture, facilitate communication, validate specifications and facilitate system maintenance.

4.2.1 Use case diagram:

this diagram describes the different actions that users can perform on the web application. It can help understand the features and features of the application.
For this, Figure 4.1 shows the actions of our application.

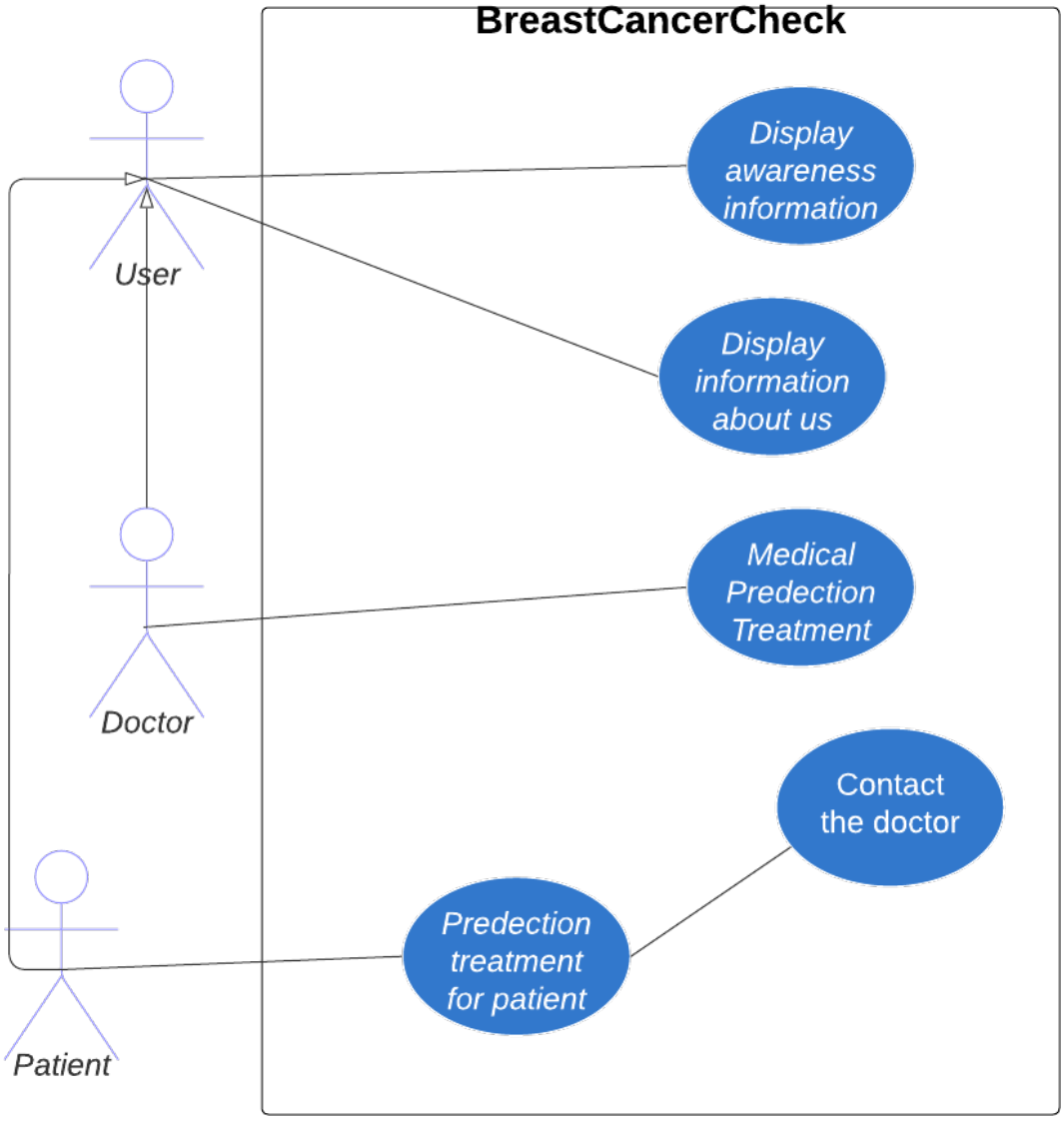

Figure 4.1: Use case diagram.

4.2.2 Sequence diagram:

This diagram can be used to describe the interactions between the different components of the web application. It can show how the application responds to user requests, how it processes data and how it returns results.
We started with the actions of a normal user and Figure 4 describes these details(Figure 4.2) . After the Doctor's actions (Figure 4.3). Finally, we ended with those of patient(Figure 4.4).

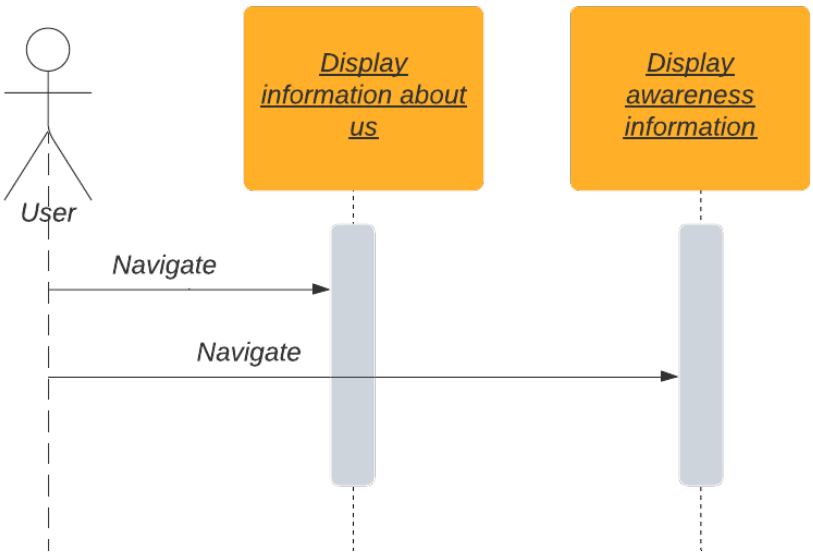

Figure 4.2: User Sequence diagram.

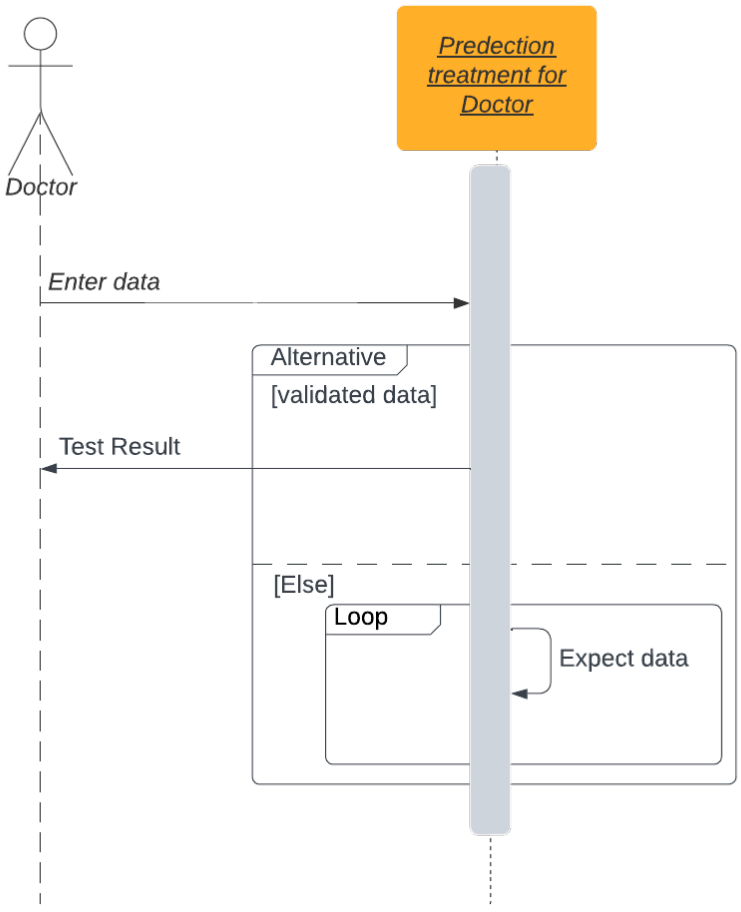

Figure 4.3: Doctor Sequence diagram.

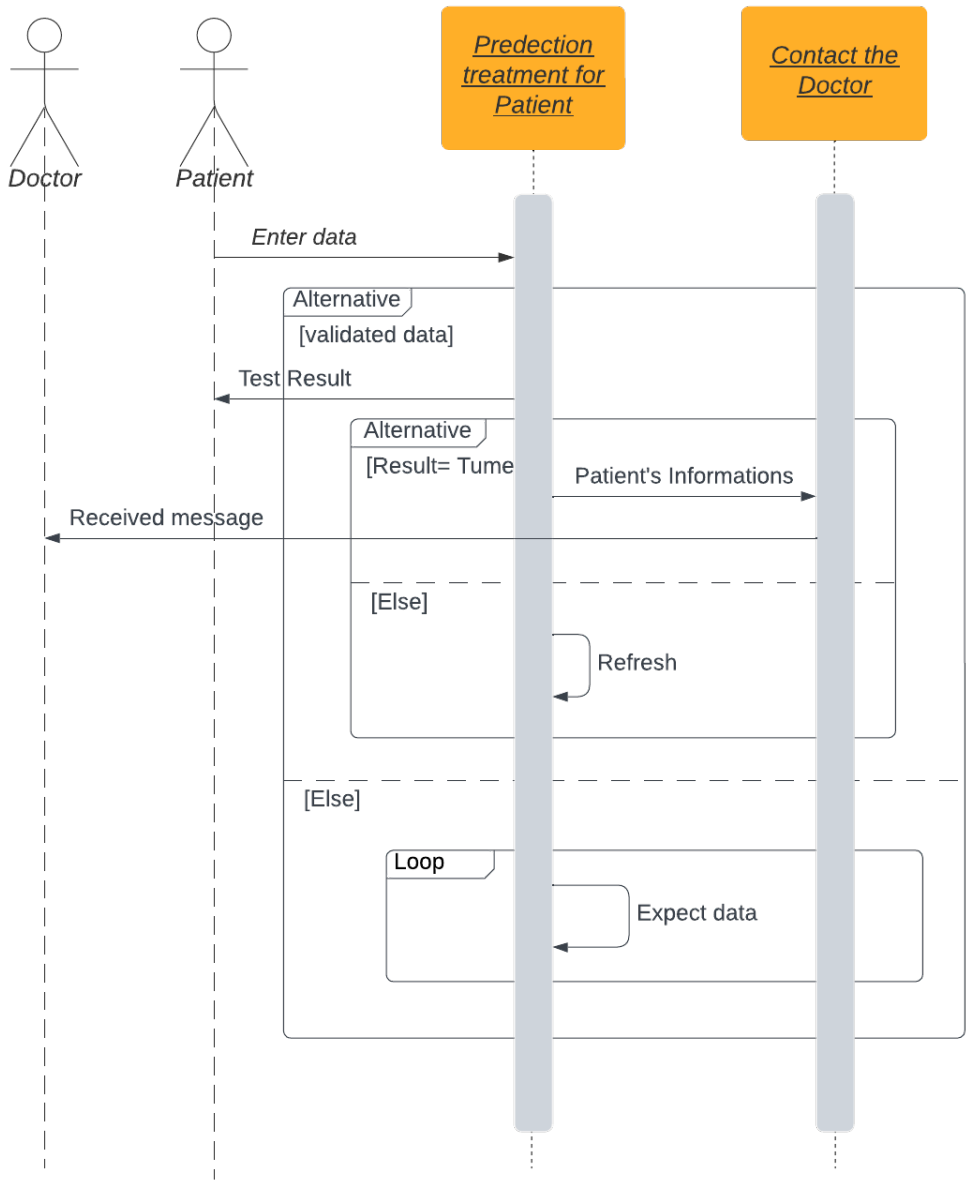

Figure 4.4: Patient Sequence diagram.

4.3 Implementation

Our breast cancer detection web application was developed using Visual code and Streamlit library in python. And as mentioned, we chose Streamlit because it is a library that allows you to create user-friendly web applications in Python with great ease.
This application uses machine learning algorithms to predict the presence or absence of breast cancer in patients. These algorithms were trained using data from previous patients. The prediction results are then displayed on the screen.

4.3.1 Application's Pages

We create a menu in which we can choose one of these :

- About Us
- Breast Cancer Sensitization
- Predict for Doctors
- Predict for Patients

In addition, the app's homepage "About Us" page contains information about the application developers and their motivation for creating this app (Figure 4.5).

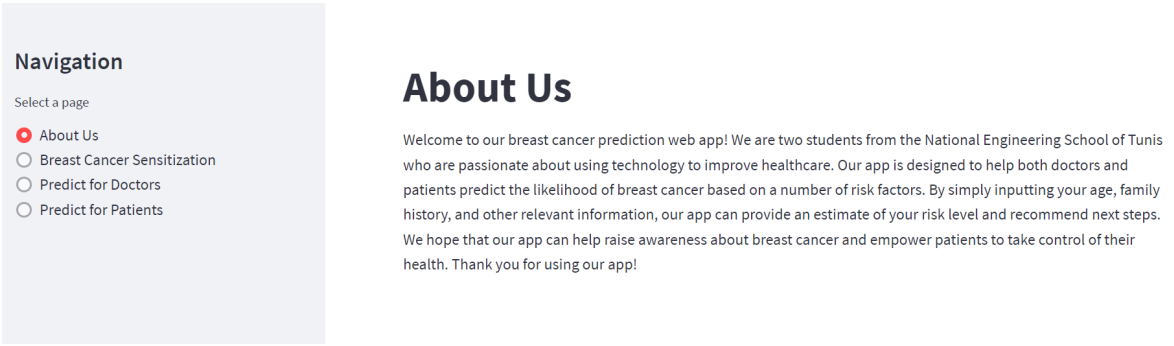

Figure 4.5: Homepage capture

Moreover, you find the page of "Breast Cancer Sensitization" is designed to educate the general public about breast cancers. This page includes information about symptoms, risk factors, and treatment options (Figure 4.6).

Right now, for the predictions, we have a dedicated page for doctors and one for patients.
The Doctor Space page allows doctor to enter data extracted from patients' medical images. This data is then used to predict the presence or absence of breast cancer in patients (Figure 4.7).
The patient data entry page allows patients to enter data from blood tests. This data is also used to predict the presence or absence of breast cancer in patients (Figure 4.8). If the result is breast cancer (Figure 4.9), the patient has the option to contact a doctor by sending personal information and a message via the app. In addition, he or she can then make an appointment with the doctor to follow the medical treatment process (Figure 4.10).

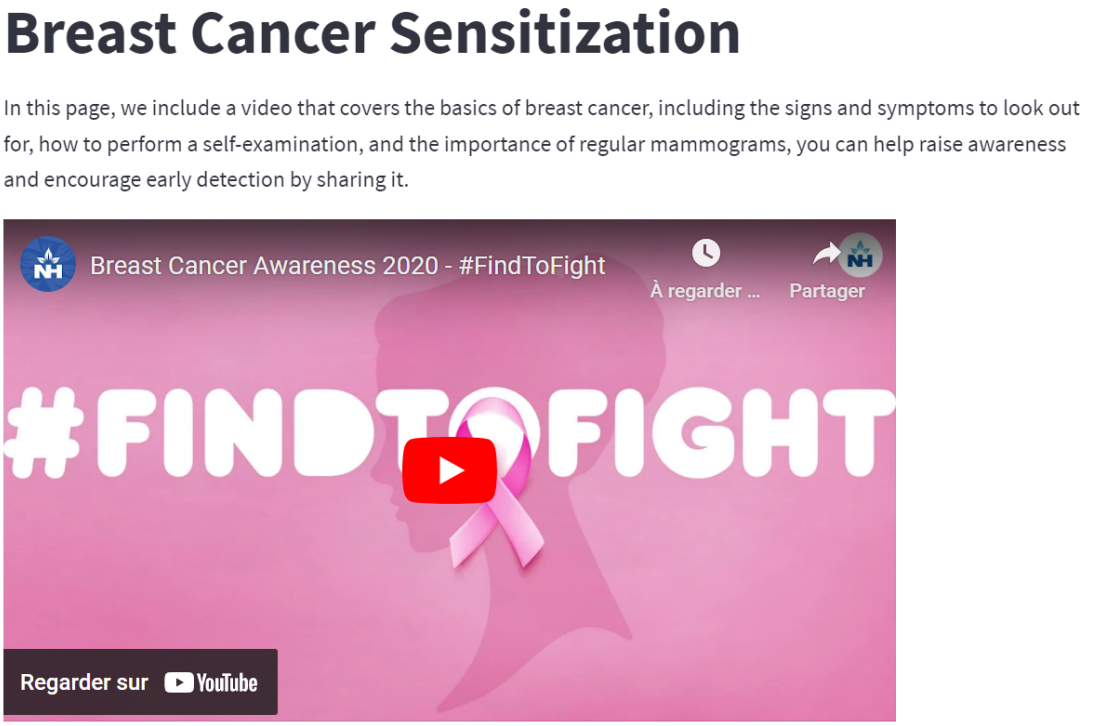

Figure 4.6: Page Cancer Sensitization.

Figure 4.7: Doctor Space.

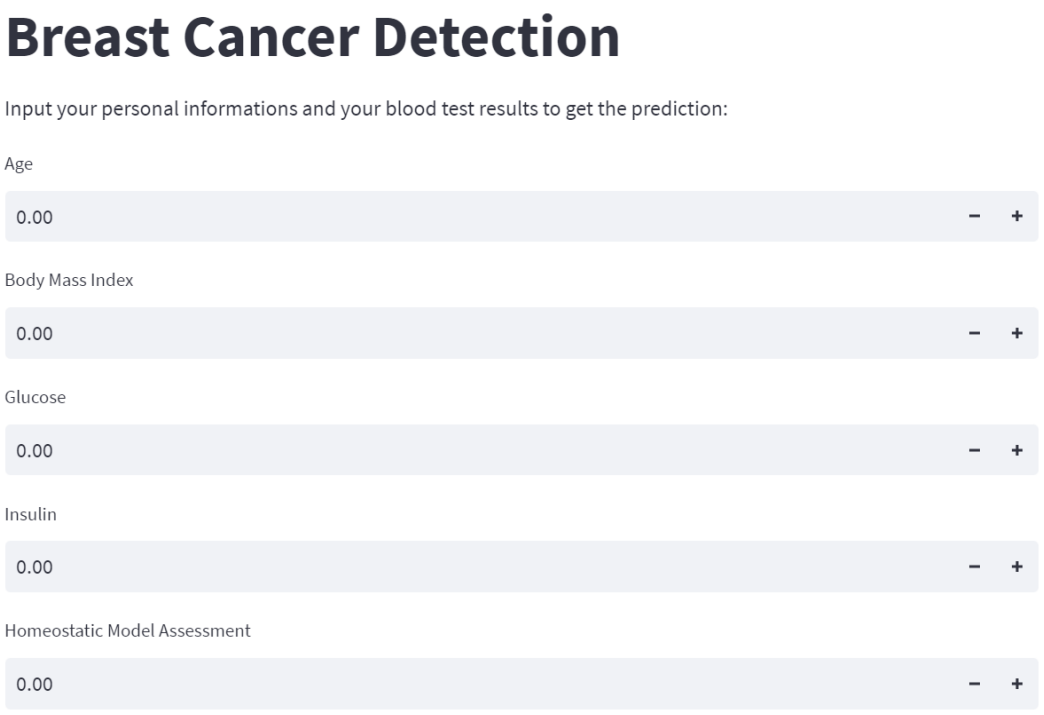

Figure 4.8: Patient Space.

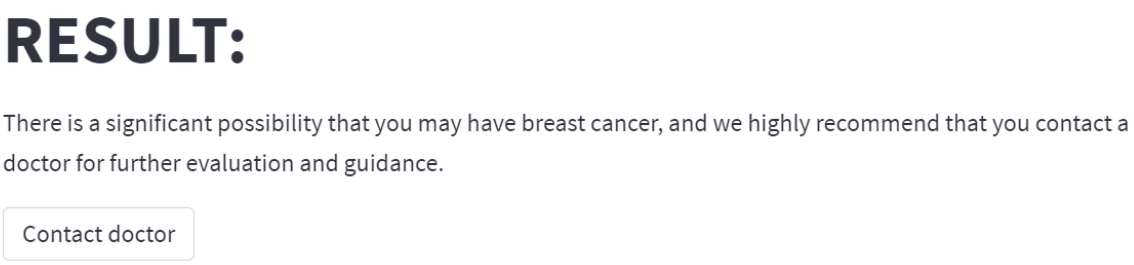

Figure 4.9: Example Test Result

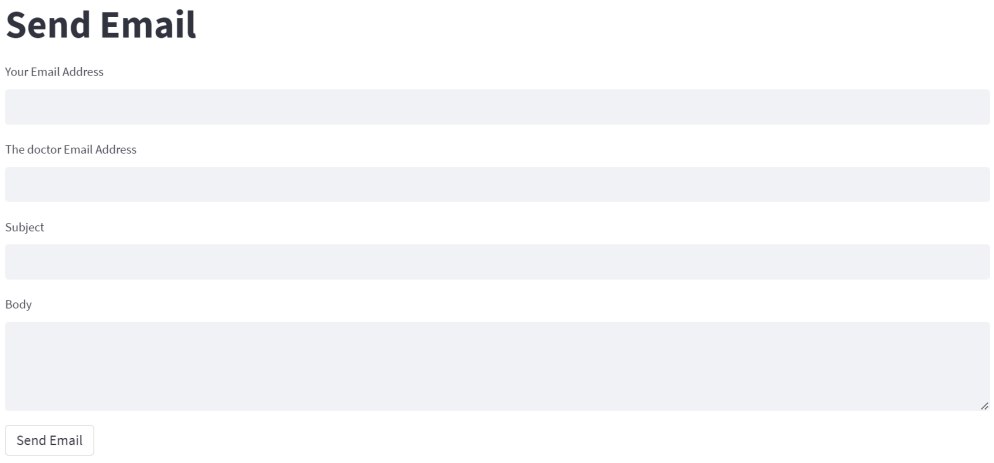

Figure 4.10: Send Email Page

4.3.2 Code generated

As far as our code is concerned, we will describe it in a general way.
The code was written using Python and the Streamlit library to build the user interface. We started by importing several libraries and modules that were used to create the web application. For example, streamlit for creating the web application, pandas for data manipulation, pickle for serializing and deserializing objects. Then we define many functions. Each function represents a whole page.

Conclusion

In conclusion, the modeling and implementation of our web platform has been a comprehensive process that involved careful consideration of various choices, methods, and techniques. By utilizing use case diagrams and sequence diagrams, we were able to accurately model the functionality of our platform and ensure it met the needs of our users.
The implementation phase involved the creation of application pages and code generation, which was crucial in bringing our platform to life.
Throughout the process, we kept the end-user in mind, ensuring that the platform was easy to use and accessible for patients and doctors alike.

Chapter 5

Integration and Recommendation

Introduction

To ensure that the application works seamlessly, we have integrated different components of the application, including the patient and doctor interfaces and machine learning model.
The integration part involved the explanation of how the machine learning component was integrated in the web part.
In the recommandation section, we provide insights into the application's performance and how we can improve it.

5.1 Integration

In the integration part , we used the pickle library which is often used in combination with other libraries and frameworks to save and load machine learning models, store user preferences, and implement other data persistence tasks.
Our main objective of the integration part is to load the pre-trained machine learning models (mentioned in the chapter 3) from a two separate files named model6.pkl for the patient space in our application and model1.pkl for the doctor space in our application the and assigning it to a variable named model.
This loaded model can then be used to make predictions on new data by the user of the app.

```
# Load the trained model
    model = pickle.load(open('model1.pkl', 'rb'))
```

Figure 5.1: Model Integration in the Doctor space

```
# Load the trained model
model = pickle.load(open('model6.pkl', 'rb'))
```

Figure 5.2: Model Integration in the Patient space

5.2 Recommendation

Building a web application is often an iterative process that requires continuous development to add functionality and improve the user experience. However, due to time or resource constraints, there may be missing features that could improve the application. In our breast cancer detection project, there are several missing features that could be added to improve the user experience and quality of service.

First of all, the login functionality is missing in our application. This can be solved by adding a login page for each type of user (normal user, physician, and patient) to secure personal information and ensure proper access to each functionality. Second, although our application uses pre-existing medical imaging data, we could add functionality to extract medical imaging data directly from radiology equipment. This would improve the efficiency of our application by reducing the time required to extract the data and allowing for real-time analysis. In addition, we could add calendar functionality to organize appointments between patients and physicians. This would allow for better appointment management and time organization for both parties. Finally, a chatbot or smart assistant could be added to the app to immediately answer users' questions and guide them through the various features of the app. This would enhance the user experience and provide additional support for users who need help. Although these features were not included in our current app, we believe they could be important improvements for the future version of our breast cancer detection app.

To continue the development of our web application, we plan to add the missing features in the next steps. We will start with the implementation of the login functionality, which will allow to separate the dedicated space for each user according to his functionalities. Then, we will work on the extraction of medical imaging data using an image segmentation method that will be integrated into our application. We also plan to set up a calendar and a dashboard to organize users' spaces, which will allow a better management of appointments between patients and doctors. For this, we will use open source scheduling tools like FullCalendar and Metabase. Finally, we will implement an intelligent assistant (ChatBot) to answer users' questions instantly.We will use natural language processing (NLP) tools like Rasa to create an assistant that will be able to answer frequent questions from users. We plan to complete these missing features in the next 6 months to improve the user experience and increase the accuracy of our application.

In sum, adding these missing features will allow our breast cancer detection app to be more comprehensive and user-friendly. Implementing these features will take time and resources, but it will be worth it to provide a superior service to patients and physicians. We are confident in our ability to successfully implement these features using web development best practices and modern tools to ensure the quality of the application. We hope that our application will be useful in the fight against breast cancer and that it will help save lives.
Ultimately, we remain committed to the further development of our application to continuously improve the user experience and provide an effective solution for early breast cancer detection.

Conclusion

In the chapter 5 , we discussed the integration part and we wrote some ideas how we can improve the web app and make it more efficient and beneficial for users.

General Conclusion

In conclusion, this report presented the development of a machine learning based prototype for breast cancer detection. This work was motivated by the need to develop a system that is accessible and usable for non-technical users.

We can say that the project was divided into five main steps: a literature search, data collection and preparation, development of the breast cancer detection model, development of the web application and integration of machine learning into the application.
First, a literature search was conducted to identify existing solutions in breast cancer detection, medical detection methods and machine learning. This research helped to understand the advantages and limitations of the different existing solutions and to establish a comparison between them. This comparison revealed that the existing solutions were mainly theoretical and not accessible to non-technical users and not exploited on the web. This step identified the need for an accessible breast cancer detection system that can be used by all users. Then, data were collected and prepared to form the breast cancer detection model. In the third step, a breast cancer detection model was developed using machine learning algorithms. The choice of the algorithm was based on a comparison of the performance of different algorithms. In the fourth step, a web application was developed to allow users to test the breast cancer detection model. The user interface was designed to be simple and easy to use for non-technical users. The functionality of the application allowed doctors to enter information extracted from a mammographic image, patients to enter information extracted from a blood test after obtaining a prediction of the breast cancer detection model. Finally, in the fifth step, the breast cancer detection model was integrated into the web application. This integration was done using libraries such as pickle. The application was tested to verify that the model worked correctly.

However, this work also has limitations and room for improvement. Future studies could explore the possibility of using other machine learning techniques to improve the performance of the breast cancer detection model. Hence, this work paves the way for future studies on the use of machine learning to improve breast cancer detection solutions and to help patients fight this disease.

References

[1] Mentum healthcare urology marketing agency. *What is a Breast Xray? | Two Views.* (accessed: 13.02.2023).

[2] Yassir Edrees Almalki et al. "Breast Cancer Detection in Saudi Arabian Women Using Hybrid Machine Learning on Mammographic Images." In: *Computers, Materials & Continua* (2022).

[3] Houda Ben Ayed et al. "National Burden of Cancers in Tunisia: A Disability Adjusted Life–year Approach." In: *Dr. Sulaiman Al Habib Medical Journal* 2 (3 2020), pp. 115–122. ISSN: 2590-3349. DOI: 10.2991/dsahmj.k.200901.001.

[4] Thiranja Prasad Babarenda Gamage et al. "An automated computational biomechanics workflow for improving breast cancer diagnosis and treatment." In: *Interface Focus* 9.4 (2019), p. 20190034.

[5] Sage Bionetworks. *IBM and Sage Bionetworks announce winners of first phase of DREAM Digital Mammography Challenge.* (accessed: 28.03.2023).

[6] Freddie Bray et al. "Global estimates of cancer prevalence for 27 sites in the adult population in 2008." In: *International journal of cancer* 132.5 (2013), pp. 1133–1145.

[7] Cathy Cassata. *Types of Breast Cancer.* (accessed: 11.02.2023).

[8] Imen Cheikhrouhou. "Description et classification des masses mammaires pour le diagnostic du cancer du sein." PhD thesis. Université d'Evry-Val d'Essonne, 2012.

[9] doctorfly. *What are the differences between benign and cancerous tumors.* (accessed: 11.02.2023).

[10] Neslihan Dogan and Zuhal Tanrikulu. "A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness." In: *Information Technology and Management* 14 (2013), pp. 105–124.

[11] Harold Ellis and Vishy Mahadevan. "Anatomy and physiology of the breast." In: *Surgery (Oxford)* 31.1 (2013), pp. 11–14.

[12] Jacques Ferlay et al. "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012." In: *International journal of cancer* 136.5 (2015), E359–E386.

[13] National Cancer Foundation. *What is cancer?* (accessed: 11.02.2023).

[14] Aurélien Géron. *Hands-on machine learning with scikit-learn and tensorflow: Concepts.* O'Reilly Media, 2017.

[15] Olivier Hermine. “Cancer : le diagnostic.” In: *dictionnaire médicale flammarion 8 ème édition* 50 (2020), p. 14.

[16] Hologic. *3Dimensions® Mammography System | Hologic*. (accessed: 28.03.2023).

[17] Shigao Huang et al. “Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges.” In: *Cancer letters* 471 (2020), pp. 61–71.

[18] Inc. iCAD. *SecondLook for 2D Mammography*. (accessed: 28.03.2023).

[19] The French National Cancer Institute. *cellule*. (accessed: 11.02.2023).

[20] Meerja Akhil Jabbar. “Breast Cancer Data Classification Using Ensemble Machine Learning.” In: *Engineering and Applied Science Research* 48.1 (Jan. 27, 2021), pp. 65–72.

[21] Michael I Jordan and Tom M Mitchell. “Machine learning: Trends, perspectives, and prospects.” In: *Science* 349.6245 (2015), pp. 255–260.

[22] BENTALHA KHADIDJA and ABDELBAKI ACHWAK RANIA. “L’incidence De Cancer Du Poumon Et Les Facteurs De Risque Dans La Wilaya De Tlemcen.” PhD thesis.

[23] David Killock. “CancerSEEK and destroy—A blood test for early cancer detection.” In: *Nature Reviews Clinical Oncology* 15.3 (2018), pp. 133–133.

[24] Ronald FS Lee et al. “The differential distribution of RAPTA-T in non-invasive and invasive breast cancer cells correlates with its anti-invasive and anti-metastatic effects.” In: *International journal of molecular sciences* 18.9 (2017), p. 1869.

[25] Kamel Mesbahi et al. “Un modèle d’apprentissage automatique pour la prédiction des maladies, cas d’étude: cancer du sein.” PhD thesis. Université Mouloud Mammeri, 2020.

[26] MonTraitement-CancerduSein.fr. *Types et stades du cancer du sein*. (accessed: 11.02.2023).

[28] Chigozie Nwankpa et al. “Activation functions: Comparison of trends in practice and research for deep learning.” In: *arXiv preprint arXiv:1811.03378* (2018).

[29] Fondation pour la Recherche Médicale. *Cancers : du dépistage au diagnostic, au plus tôt !* (accessed: 12.02.2023).

[30] Roche. *Diagnostic cancer du sein - la biopsie mammaire*. (accessed: 13.02.2023).

[31] Rajesh Sharma et al. “Mapping cancer in Africa: a comprehensive and comparable characterization of 34 cancer types using estimates from GLOBOCAN 2020.” In: *Frontiers in Public Health* 10 (2022).

[32] Breast Cancer Risk Assessment Tool. *Breast Cancer Risk Assessment Tool*. (accessed: 28.03.2023).

[33] Cancer Research UK. *How cancers grow*. (accessed: 11.02.2023).

[34] Johns Hopkins University. *Overview of the Breast - Breast Pathology | Johns Hopkins Pathology*. (accessed: 11.02.2023).

[35] Dilber Uzun Ozsahin et al. “The Systematic Review of Artificial Intelligence Applications in Breast Cancer Diagnosis.” In: *Diagnostics* 13.1 (2022), p. 45.

[36] Raman Valliappan, Sumari Putra, and Rajeswari Mandava. “A theoretical methodology and prototype implementation for detection segmentation classification of digital mammogram tumor by machine learning and problem solving approach.” In: *International Journal of Computer Science Issues (IJCSI)* 7.5 (2010), p. 38.

[37] Imen Zemni et al. “Breast cancer incidence and predictions (Monastir, Tunisia: 2002–2030): A registry-based study.” In: *Plos one* 17.5 (2022), e0268035.