جامعة حلوان
HELWAN UNIVERSITY
Faculty of Engineering
Computer Engineering Department

Breast Cancer detection and prediction

Team members:

Omar Khaled Mohamed Hemdan

Hadeer Elsayed Hasanain

Ali Mohamed Ahmed Youssef

Amira Abdelfattah Abdelrazik

Haydy Osama Mahmoud

Under Supervision of:

Professor Dr. Manal Shoman

ACKNOWLEDGMENT

All praises to Allah and His blessing for the completion of this project. We thank God for all the opportunities, trials and strength that have been showered on us to finish this project.

First and foremost, we would like to sincerely thank our supervisor Prof. Dr. Manal Shoman for her guidance, understanding, patience and most importantly, she has provided positive encouragement and a warm spirit to finish this project. It has been a great pleasure and honor to have her as our supervisor.

We are also grateful to all the Teaching staff in the Computer Engineering Department, for their consistent support and assistance. It was great sharing premises with all of you during the last five years.

ABBREVIATIONS' TABLE:

The abbreviation	what is it stands for
BC	Breast Cancer
UI	User Interface
RAM	Random access memory
GPU	Graphics Processing Unit
App	Application
IOS	iPhone Operating System
ID	Identification number
FNA	Fine needle aspiration
SOB	Surgical open biopsy
OS	Operating Systems
SVM	Support vector machine
CNN	Convolution Neural Network
RNN	Recurrent neural network

STUDENT	CONTRIBUTIONS
OMAR KHALED MOHAMED	- Diagnostic test model - Histopathological Image Classification
HADEER ELSAYED HASANAIN	- - Diagnostic test model - Risk assessment model
AMIRA ABDELFATTAH ABDELRAZIK	- Diagnostic test model - Risk assessment model
ALI MOHAMED AHMED YOUSSEF	- Diagnostic test model - Mobile application back-end
HAYDY OSAMA MAHMOUD	- Diagnostic test model - Mobile application front-end

TABLE OF CONTENTS

Acknowledgment 2

Abbreviations' table 3

Table of contents 5

Chapter1: Introduction 6

Abstract 6

Introduction 7

Chapter 2: Related work 9

Chapter 3: Analysis and requirements 15

System requirements 15

Functional requirements specifications 17

Use cases 17

System sequence diagram 20

Class diagram 24

Chapter 4: Implementation 25

Machine learning 25

Deep learning 28

Machine learning and deep learning models implementation 29

Mobile application UI 46

Chapter 5: Testing and validation 59

Requirements Testing 60

Chapter 6: Development and tools technologies 67

Framework 67

Languages 67

Machine learning models deployment 68

Online Tools 69

Libraries 70

Models 72

Chapter 7: References 73

CHAPTER1: INTRODUCTION

ABSTRACT

The number of breast cancer cases is significantly increasing, where breast cancer is the most common in terms of new cases of cancer in 2020 and it is the second most common cancer type for women, also the mortality of BC is very high when compared to other types of cancer due to late and wrong detection. Our project aims to improve the process of breast cancer detection and increase the number of early detected cases of breast cancer. The system should deliver an accurate detection of the type of tumor (benign or malignant).

INTRODUCTION

Breast cancer is the most found disease in women, worldwide, where abnormal growth of a mass of tissue, causing the expansion of malignant cells leads to acute breast cancer.

These malignant cells which are the main reason for breast cancer can be classified into different groups according to their unusual progress and capability affecting other normal cells (whether these malignant cells affect only the local cells or can spread throughout the full body).

It is very important to prevent this spreading effect by a diagnosis of cancer in the early stages using advanced techniques and equipment.

In recent decades, there have been many efforts to employ artificial intelligence and other related methods to assist in the detection of cancer in earlier stages. Early detection of cancer boosts the increase of survival chance to 98%.

This system provides a platform that enables the prediction and empower accurate decision making using the results obtained from a biopsy test.

A breast biopsy is a test that removes tissue or sometimes fluid from the suspicious area. The removed cells are examined under a microscope and further tested to check for the presence of breast cancer. A biopsy is the only diagnostic procedure that can definitely determine if the suspicious area is cancerous.

There are different kinds of breast biopsies. Some are done using a hollow needle, and some use an incision (cut in the skin). The type you have depends on several things, like:

- How suspicious the breast change looks or feels

- How big it is
- Where it is in the breast
- If there is more than one suspicious area
- Your overall health
- Your personal preferences

There are three types of biopsies:

- **Core needle biopsy**

  During this procedure, the doctor uses a hollow needle to take out pieces of breast tissue from the area of concern. This can be done with the doctor either feeling the area or while using an imaging test to guide the needle.

- **Fine Needle Aspiration (FNA) of the Breast**

  During a fine needle aspiration (FNA), a small amount of breast tissue or fluid is removed from a suspicious area with a thin, hollow needle and checked for cancer cells.

- **Surgical open biopsy (SOB)**

  In some situations, such as if the results of a needle biopsy aren't clear, you might need a surgical biopsy.

  For this type of biopsy, surgery is used to remove all or part of a suspicious area so it can be checked for cancer cells.

Our project focuses on the latter two (Diagnostics test based on FNA results and Histopathological test based on SOB results), the systems aim to improve the process of breast cancer detection which increases the number of early detected cases of breast cancer by delivering an accurate and simple detection to the type of tumor (benign or malignant) and helping pathologists be more productive.

The system also helps patients by providing videos for self-examination and also to calculate the risk ratio based on a statistical model known as the Gail Model.

The system allows doctors to interact with patients to send the results of the diagnosis directly to patients using National ID and helps patients to contact doctor easily.

CHAPTER 2: RELATED WORK

In the past few years, there's been a constant rise in the medical informatics research/work using machine learning and neural net algorithms. A number of datasets have been used by a number of research scholars in their study and have applied a number of methodologies to get some strong results. Some of them are:

- **Gunjan Rawa,Rakesh Rawal,Hirav Shah,Kamlesh Patel [1],**

  Focused on the implementation of artificial neural network, logistic regression, support vector machines and random forest-based classifiers for the Wisconsin breast cancer diagnostic dataset. From the results obtained, they can say that classifiers based on non-linear algorithms like ANN may have similar results to conventional classifiers like SVM and Random Forest in terms of accuracy, but when performing medical diagnosis sensitivity plays an important role as you have to minimize the misclassification rate of classifying a malignant tumor as benign. In such cases, SVM showed a higher classification rate than ANN, whereas, in terms of specificity i.e. identifying a benign tumor as benign ANN had a higher classification rate than both SVM and Random Forest.

- **Daisuke Kimura [2]**,

  Introduced the application of digital pathological image analysis using machine learning algorithms, addressed some problems specific to such analysis, and proposed possible solutions.

- **Abien Fred M. Agarap [3],**

  Presented a comparison of six machine learning (ML) algorithms: GRU-SVM , Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbor (NN) search, Softmax Regression, and Support Vector Machine (SVM) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset by measuring their classification test accuracy and their sensitivity and specificity values. For the implementation of the ML algorithms, the dataset was partitioned in the following fashion: 70% for the training phase, and 30% for the testing phase. The hyper-parameters used for all the classifiers were manually assigned. Results show that all the presented ML algorithms performed well (all exceeded 90% test accuracy) on the classification task. The MLP algorithm stands out among the implemented algorithms with a test accuracy of ~99.04%.

- **Kader potdar [4],**

  Compared the classification accuracy of 3 Machine Learning algorithms – kNN, ANN and NB on UCI Wisconsin Breast Cancer dataset. The aim of this comparative study was to find the most accurate machine learning tool that can act as a tool for diagnosis of breast cancer. According to the prediction results, ANN has highest accuracy for the given dataset. This shows that ANN can be used for prediction of breast cancer as compared to kNN and NB.

- **MohammedAmineNajia [5] ,**

  On the Wisconsin Breast Cancer Diagnostic dataset (WBCD) they applied five main algorithms which are: SVM, Random Forests, Logistic Regression, Decision Tree, K-NN, calculate, compare and evaluate different results obtained based on confusion matrix, accuracy, sensitivity, precision, AUC to identify the best machine learning algorithm that are precise, reliable and find the higher accuracy. All algorithms have been programmed in Python using scikit-learn library in Anaconda environment. After an accurate comparison between their models, they found that Support Vector Machine achieved a higher efficiency of 97.2%, Precision of 97.5%, AUC of 96.6% and outperforms all other algorithms. Support Vector Machine has demonstrated its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of accuracy and precision.

- **Fabio Spanhol [6],**

  Presented a set of experiments con-ducted on the BreaKHis dataset using a deep learning approach to avoid hand-crafted features. They have shown that they could use an existing CNN architecture, in that case AlexNet, that has been designed for classifying color images of objects and adapt it to the classification of BC histopathological images. They have also proposed several strategies for training the CNN architecture, based on the extraction of patches obtained randomly or by a sliding window mechanism, that allow it to deal with the high-resolution of these

textured images without changing the CNN architecture designed for low-resolution images.

- **Jitendra Maan, Harsh Maan [7],**

  Studied identification of five diagnostic categories of breast cancer by training a CNN (VGG16, ResNet architecture). They have used BreakHis dataset to train our model. They focused on both detection and classification of cancerous regions in histopathology images.

- **Yassir Benhammoua,b, Boujemˆaa Achchabb, Francisco Herreraa, SihamTabika [8**],

  Defined a taxonomy that categorize this problem into four different reformulations: Magnification-Specific Binary (MSB), Magnification-Independent Binary (MIB),Magnification-Specific Multi-category (MSM) and Magnification-Independent Multi-category (MIM) classifications. They provided a comprehensive survey of all related works. They identified the best reformulation from clinical and practical standpoints. They explored for the first time the MIM approach using deep learning and drew the learnt lessons.

- **Gigi F. Stark [9],**

  Used personal health data and machine learning models to Predict breast cancer risk. Models were trained and evaluated on the PLCO; finally she concluded that logistic regression, linear discriminant analysis, and neural network models with the broader set of inputs predicted five-year breast cancer risk close to the BCRAT did. These results suggest that additional easy-to-obtain inputs can also improve the Gail model's ability to predict breast cancer risk.

- **Geunwon Kim [10] ,**

  Compared Machine learning techniques for personalized breast cancer risk prediction with the BCRAT (Gail model). To provide strong assessment, reliable comparison, and reproducible results, they compared ML-based estimates and estimates from BCRAT model using US population actual observational dataset. They used the same risk factors as BCRAT model, as input for the ML algorithms in the comparison.

Conclusion:

There was an improvement in the accuracy of classification of women with and without breast cancer achieved with ML algorithms compared to the BCRAT (Gail model).

Limitations:

•In the US population-based sample, YBCS had fewer affected relatives than their cancer-free relatives. Thus, number of affected

relatives was detected as an important variable but without external validity in interpretation.

•The inherent complexity of how risk factors interact with each other, their independent effect on the outcome, and how effect sizes are determined within each ML algorithm is not known.

CHAPTER 3: ANALYSIS AND REQUIREMENTS

SYSTEM REQUIREMENTS

1.1. Functional requirements

Requirements	Description	Actor
1. Sign up	User should create an account once, to be able to use the application, so he should input the required credentials to become a user.	Doctor & normal user
2. Log in	After registration, user will be able to login whenever he needs, by entering the email and password he registered with.	Doctor & normal user
3. Check breast cancer risk	This function enables user to answer some questions and get the risk ratio of having breast cancer in the future.	Normal user
4. Tutorial for self-examination	This function enables user to examine himself by watching some videos	Normal user
5. contact a doctor	this function enables the user to contact a registered doctor on the app	normal user
6. Breast cancer detection using fine needle aspirate test	This function enables doctors to detect the type of tumor by deducing features computed from a digitized image of a fine needle aspirate of breast mass and entering these values to the application to detect type of tumor	Doctor
7. Breast cancer detection using histopathological images	This function enables doctors to know the type of cancer by entering the histopathological images from surgical open biopsy test to the application	Doctor

1.2. Non-functional requirements

Requirements	Description
1- Availability	Our app is available every time and wherever there is an internet connection, and it is a great privilege as some apps would have downtime for updates, backups and recovery
2- Usability	Our app is easy to use, without too many instructions or obstacles
3- Performance	Our app has short response time to respond to user's request and low utilization of phone resources.
4- Reliability	Our app keeps updating its performance and accuracy of its results and responses by the time. That avoids failure and downtimes
5- Efficiency	Our app uses low resources of storage space, memory and bandwidth
6- Security	Our app provides user verification by email, and allows easy password changing in case of user forgets his password or try to protect himself from hacking

FUNCTIONAL REQUIREMENTS SPECIFICATIONS

1.3.Stakeholders

- Doctor
- Patient

1.4.Actors and goals

- Doctor

- Enter the numeric features from the biopsy test to get the diagnosis and then save the result using the patient id.
- Upload the histopathological image to get the diagnosis and then save the results using the patient id.

- Patient
  - Answer the questions to get the risk ratio of having breast cancer in the future.
  - Get the instructions for self-examination.

USE CASES

- Use cases description
  - **Register:** Users should create an account once, to be able to use the application.

    Requirements: insert the required credentials.

- **Login:** Each user can Login to the system by entering the email and password which make the system Provides Authentication and Authorization.

  Requirements: login

- **Get the risk ratio**: the patient gets the risk ratio of having breast cancer in the future**.**

  Requirements: login, answer 9 questions about her medical history.

- **Self-examination instruction:** the patient could see a video or read written instructions about how to examine herself.

  Requirements: login

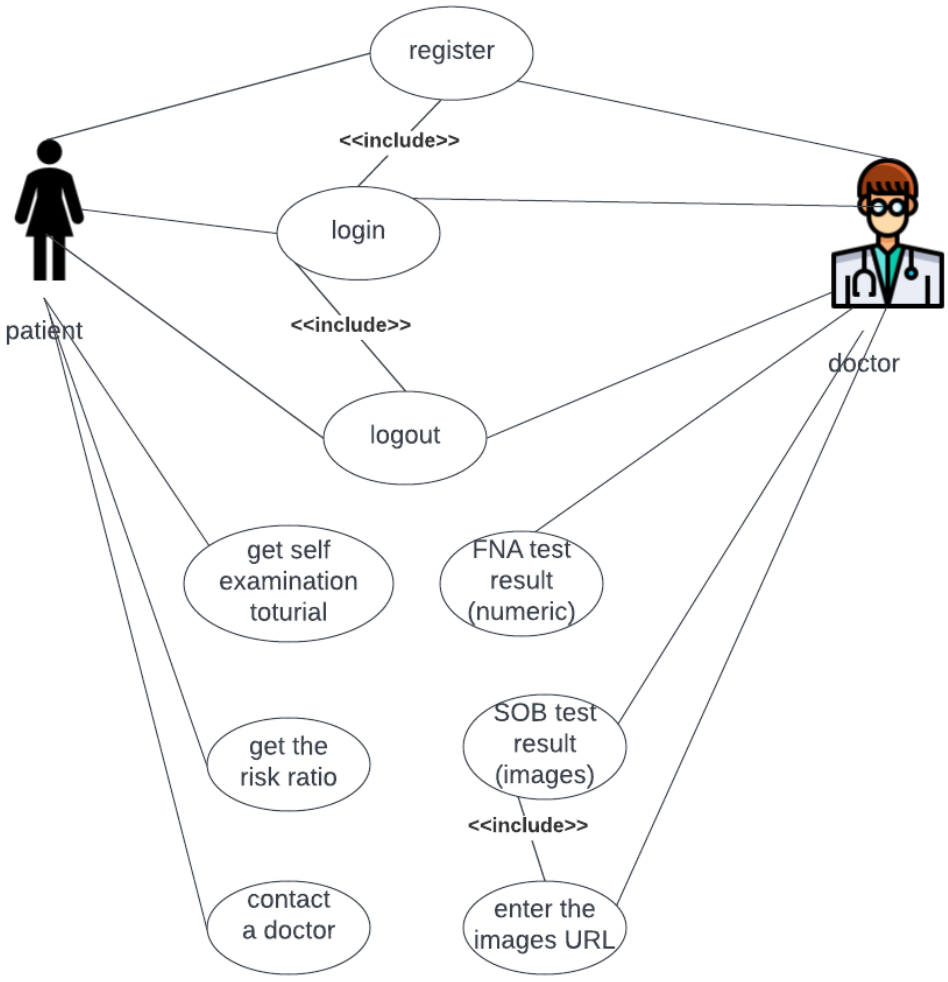
- **Get FNA test result (numeric):** the doctor gets the diagnosis of FNA results and then he can save it using the patient id.

  Requirements: login, enter the numeric value of each feature required.

- **Get Surgical open biopsy test result (image):** the doctor gets the diagnosis of FNA image result and then he can save it using the patient id.

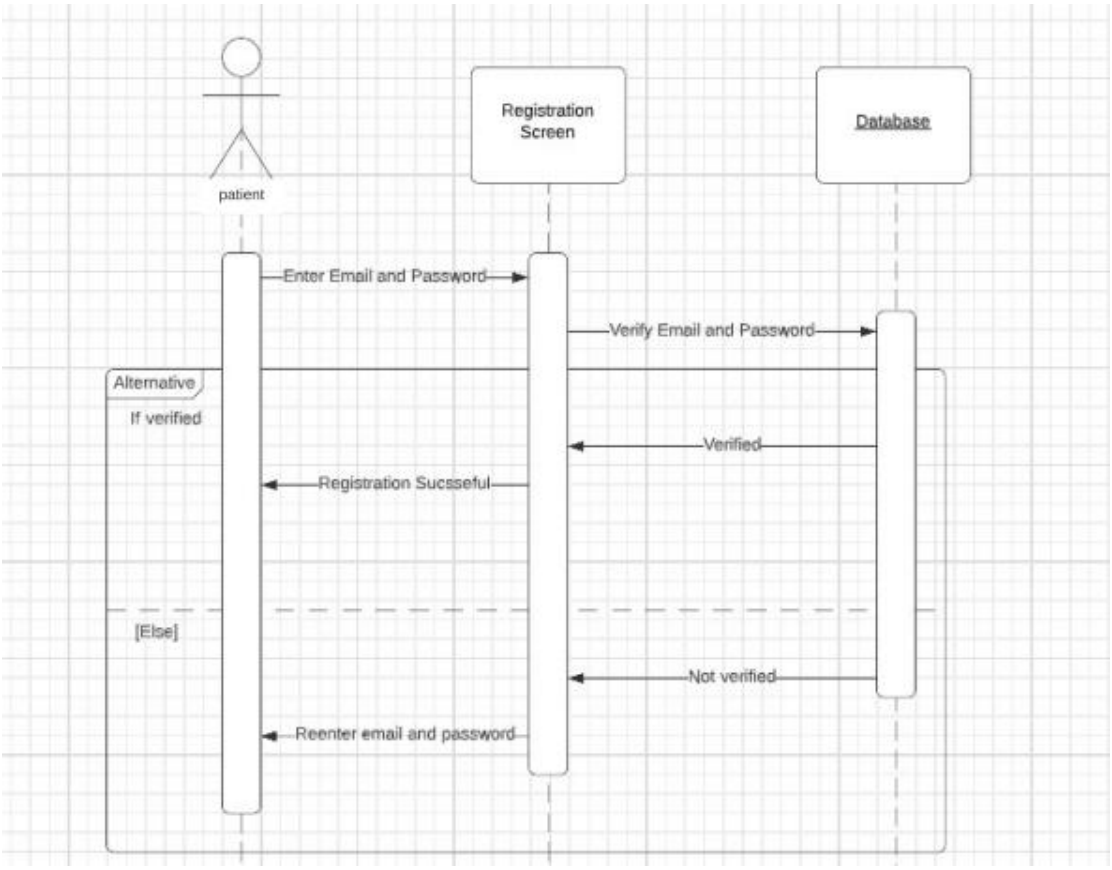
  Requirements: login, upload the image URL.

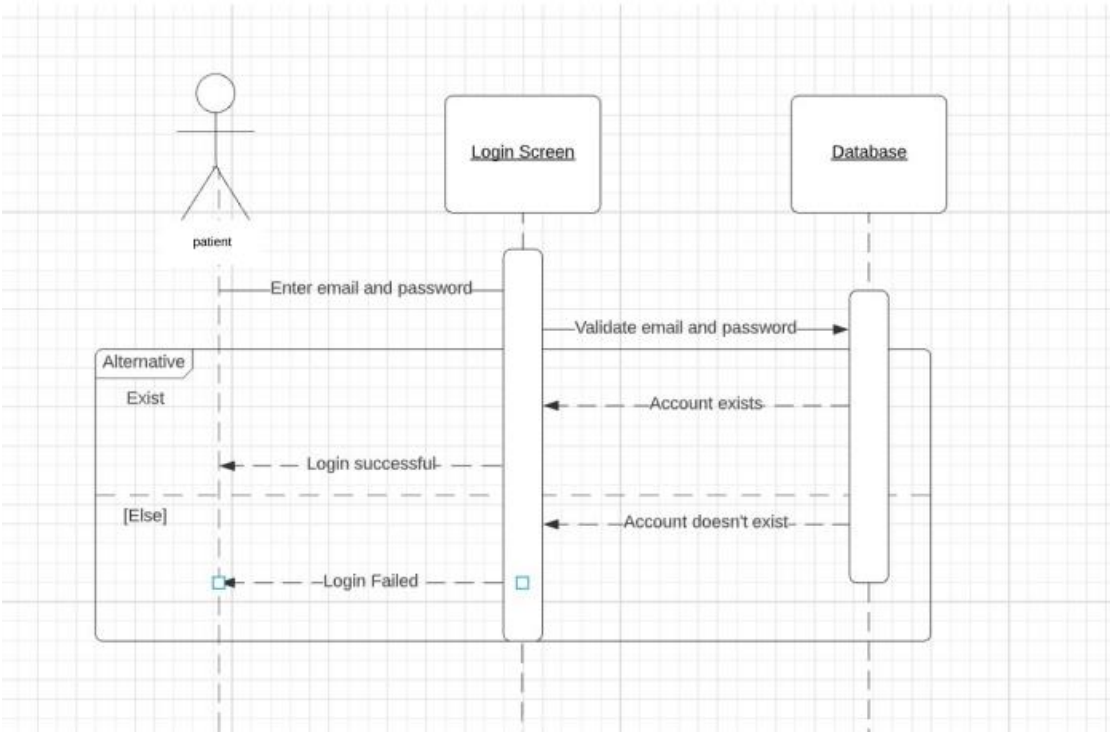
- **Use case diagram**



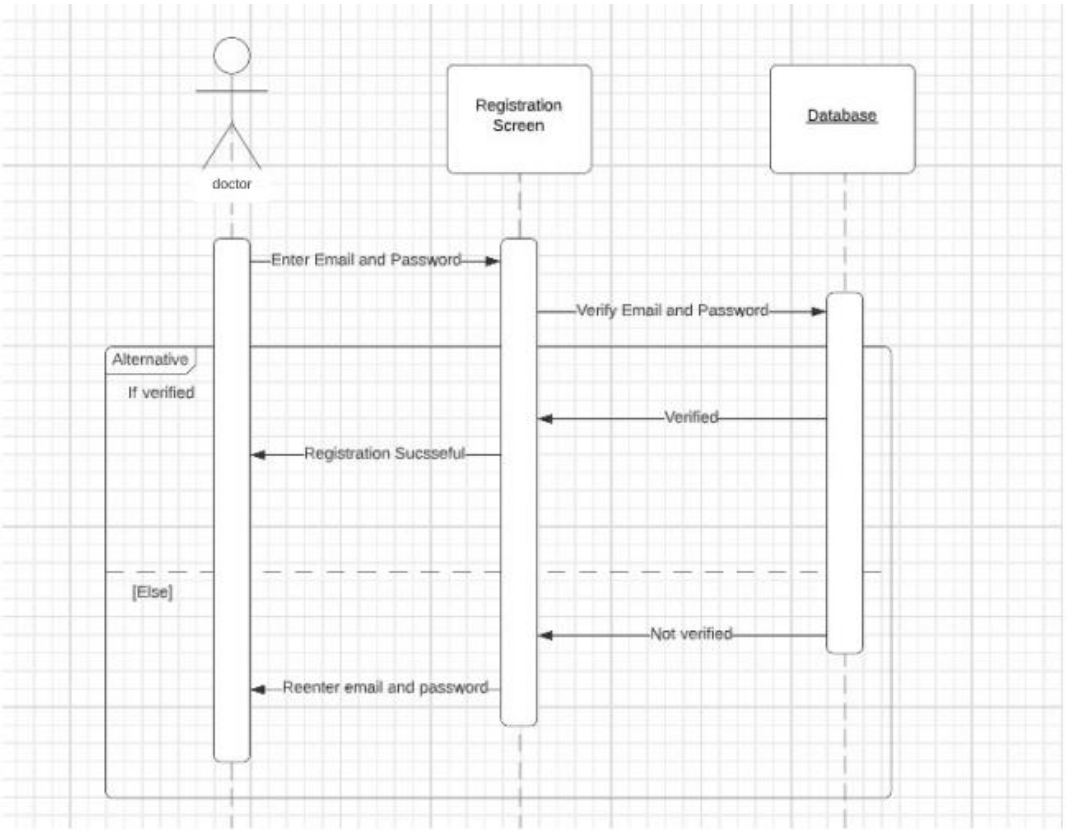
SYSTEM SEQUENCE DIAGRAM

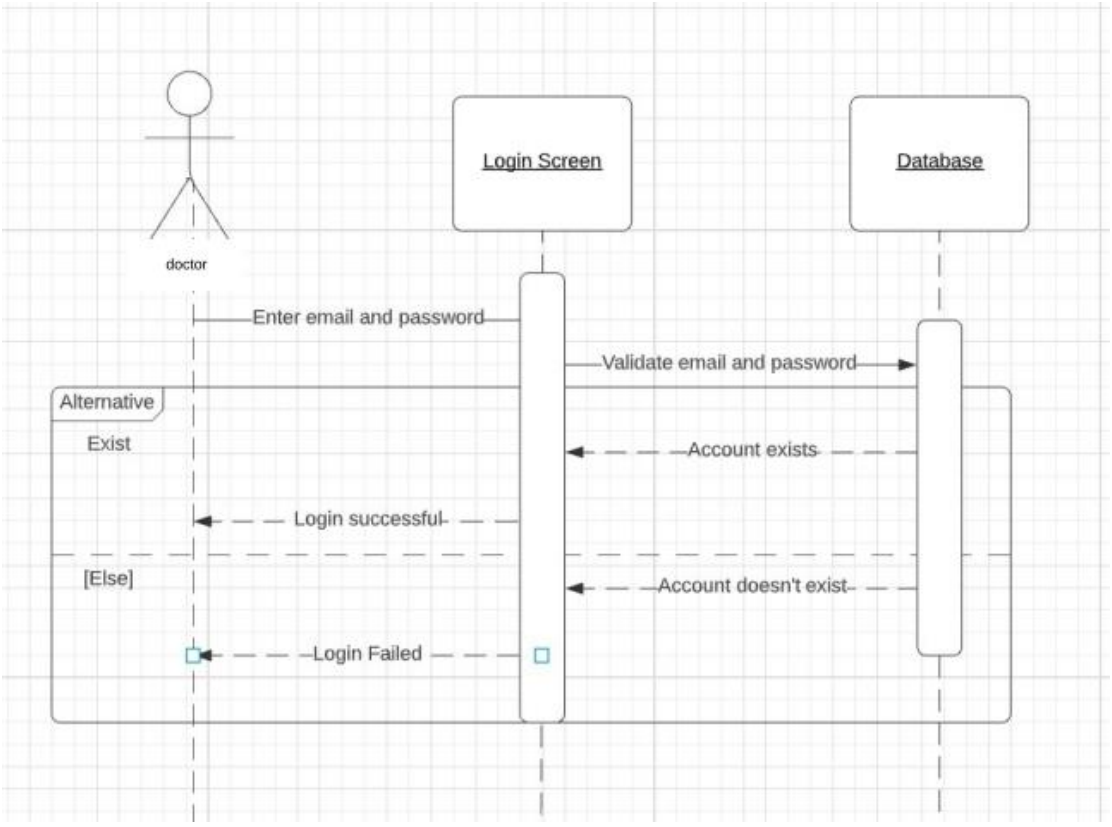
1. patients register and login



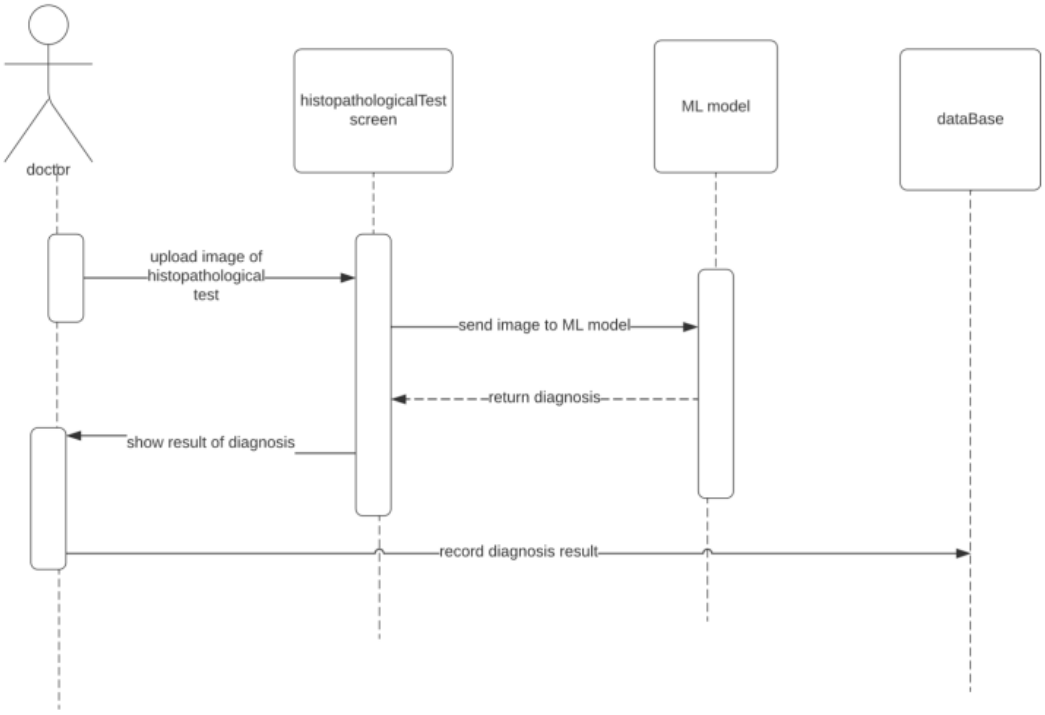


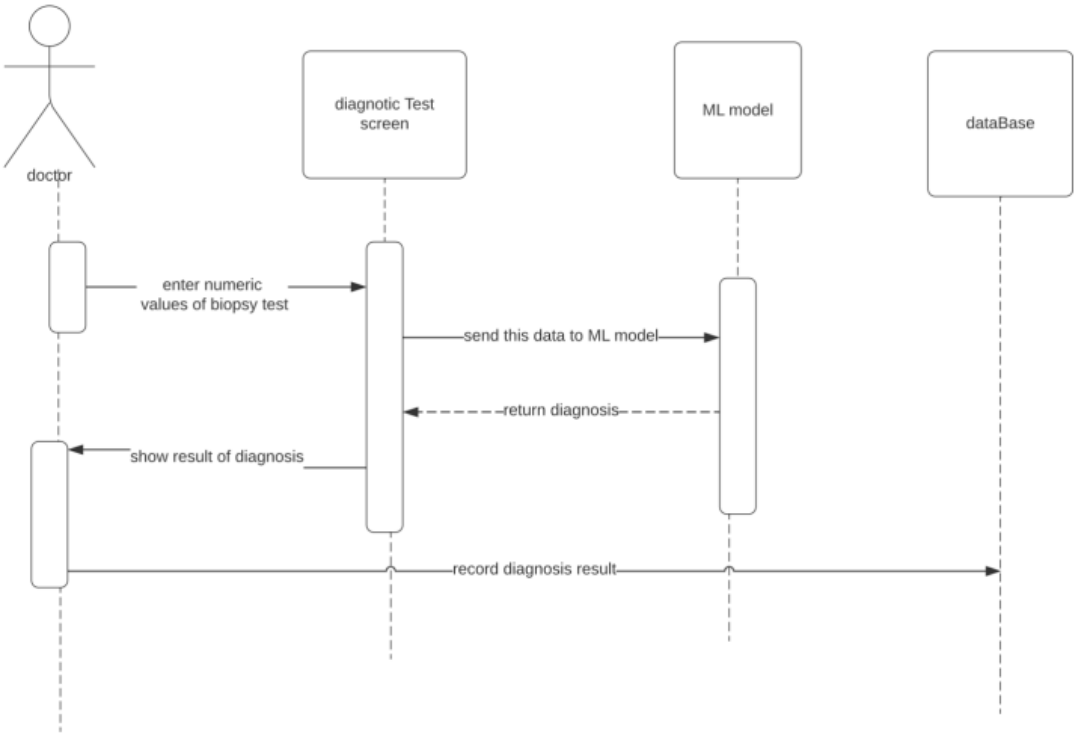
2. doctor register and login



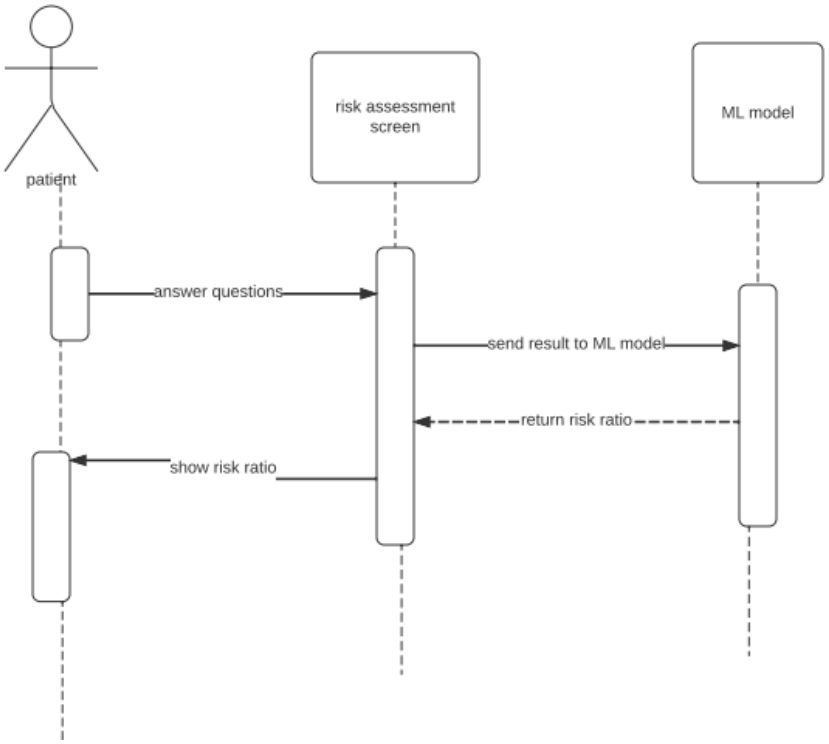


3. doctor screen functions

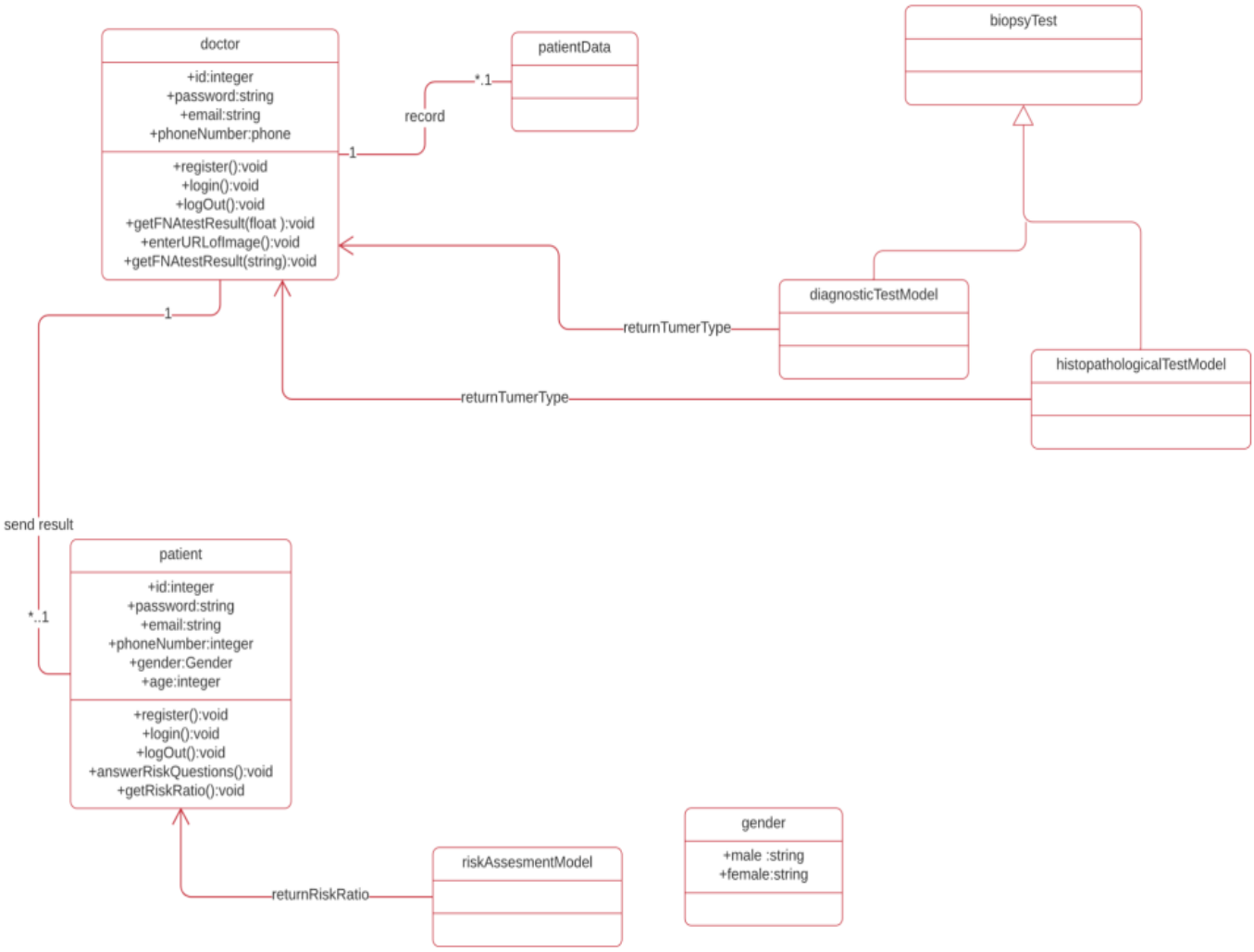




4. patient screen risk assessment



CLASS DIAGRAM



CHAPTER 4: IMPLEMENTATION

MACHINE LEARNING

It is a branch of artificial intelligence (AI) that enables computers to "self-learn" from training data and improve over time, without being explicitly programmed. Machine learning algorithms can detect patterns in data and learn from them, to make their own predictions.

There are four types of machine learning algorithms: supervised, semi-supervised, unsupervised and reinforcement.

1) Supervised learning

In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions.

Under the umbrella of supervised learning fall: Classification, Regression and Forecasting.

2) Semi-supervised learning

Semi-supervised learning is similar to supervised learning, but instead uses both labeled and unlabeled data. Labeled data is essentially information that has meaningful tags so that the algorithm can understand the data, whilst unlabeled data lacks

that information. By using this combination, machine learning algorithms can learn to label unlabeled data.

3) Unsupervised learning

Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analyzing available data. In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly.

Under the umbrella of unsupervised learning, fall: Clustering and Dimension reduction.

4) Reinforcement learning

Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters, and end values. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal.

There are a lot of machine learning algorithms, but we will focus on four of them (supervised Learning algorithms) and we will try to find the best one for breast cancer detection, the algorithms are:

- **Logistic Regression:** focuses on estimating the probability of an event occurring based on the previous data provided. It is

used to cover a binary dependent variable, that is where only two values, 0 and 1, represent outcomes.

- **Decision Trees:** It is a flow-chart-like tree structure that uses a branching method to illustrate every possible outcome of a decision. Each node within the tree represents a test on a specific variable – and each branch is the outcome of that test.

- **Random Forests:** It is an ensemble learning method, combining multiple algorithms to generate better results for classification, regression, and other tasks. Each individual classifier is weak, but when combined with others, can produce excellent results. The algorithm starts with a 'decision tree' (a tree-like graph or model of decisions) and an input is entered at the top. It then travels down the tree, with data being segmented into smaller and smaller sets, based on specific variables.

- **Support Vector Machine Algorithm: A**re supervised learning models that analyze data used for classification and regression analysis. They essentially filter data into categories, which is achieved by providing a set of training examples, each set marked as belonging to one or the other of the two categories. The algorithm then works to build a model that assigns new values to one category or the other.

DEEP LEARNING

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to "learn" from large amounts of data.

While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy.

Deep learning drives many artificial intelligence (AI) applications and services that improve automation, performing analytical and physical tasks without human intervention.

Deep learning algorithms are incredibly complex, and there are different types of neural networks to address specific problems or datasets. For example,

1. Convolutional neural networks (CNNs),

A convolutional neural network, or CNN, is a deep learning neural network designed for processing structured arrays of data such as images. Convolutional neural networks are widely used in computer vision and have become the state of the art for many visual applications such as image classification

Convolutional neural networks are very good at picking up on patterns in the input image, such as lines, gradients, circles, or even eyes and faces. It is this property that makes convolutional neural networks so powerful for computer vision. Unlike earlier computer vision algorithms, convolutional neural networks can operate directly on a raw image and do not need any preprocessing

2. Recurrent neural network (RNNs)

Are typically used in natural language and speech recognition applications as it leverages sequential or times series data

MACHINE LEARNING AND DEEP LEARNING MODELS IMPLEMENTATION

1. Diagnostic test model (numeric)

1.1. Dataset

- This data was donated by researchers of the University of Wisconsin and includes the measurements from digitized images of fine-needle aspirate (FNA) of a breast mass.
- The breast cancer data includes 569 examples of cancer biopsies, each with 32 features. One feature is an identification number, another is the cancer diagnosis and 30 are numeric-valued laboratory measurements. The diagnosis is coded as "M" to indicate malignant or "B" to indicate benign.

```
# import data
df= pd.read_csv("C:/Users/a/Downloads/data.csv")
df.head(7)
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...
5	843786	M	12.45	15.70	82.57	477.1	0.12780	0.17000	0.1578	0.08089	...
6	844359	M	18.25	19.98	119.60	1040.0	0.09463	0.10900	0.1127	0.07400	...

1.2. feature selection

- After extracting the mentioned features, we check the correlation between the remaining features and the output (diagnosis). We have 4 features that have negative correlation with the output, so we drop them. Lastly, we are left with 8 features that will be used as the input features.

```
In [102]: df.columns

Out[102]: Index(['diagnosis', 'radius_mean', 'texture_mean', 'smoothness_mean',
                 'compactness_mean', 'symmetry_mean', 'radius_se', 'compactness_se',
                 'fractal_dimension_se'],
                dtype='object')
```

1.3. Encode the categorical fields

```
#encode the categorial data fields
from sklearn.preprocessing import LabelEncoder
le= LabelEncoder()
df['diagnosis']=le.fit_transform(df['diagnosis'])

df['diagnosis'].value_counts()
```

```
0    357
1    212
Name: diagnosis, dtype: int64
```

1.4. Preparing the data

```
# prepare

from sklearn.model_selection import train_test_split

x=df.iloc[:,1:8].values
y=df.iloc[:,1].values

# Split into training and test set
x_train, x_test, y_train, y_test = train_test_split( x, y, test_size = 0.25, random_state=16)

#feature scaling

from sklearn.preprocessing import StandardScaler
sc= StandardScaler()
x_train=sc.fit_transform(x_train)
x_test=sc.fit_transform(x_test)
```

1.5. Algorithms

Algorithms	Accuracy Testing Set (%)
Logistic Regression	95.1%
Decision Tree	86.7%
Random Forest	95.8%
SVM	97.2%

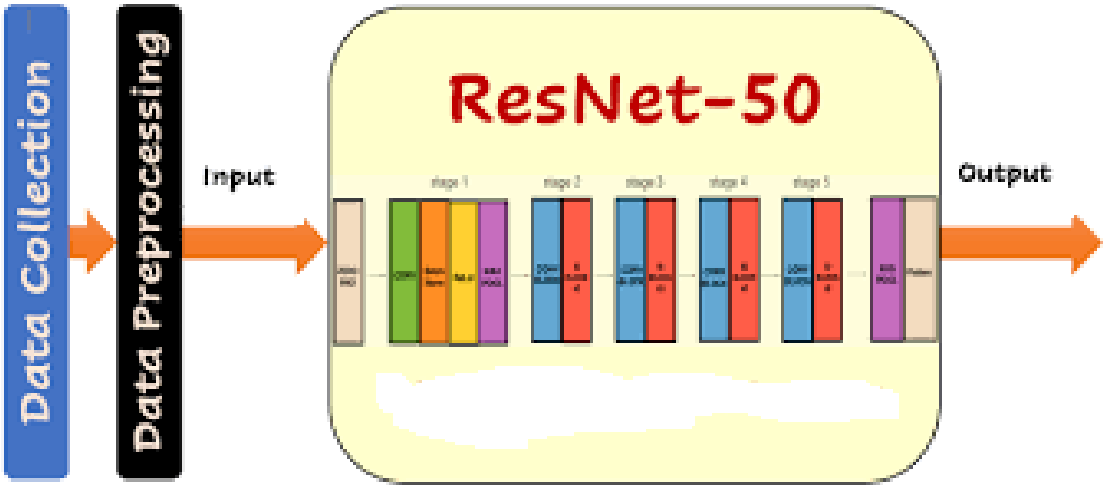
1.6. Conclusion

- After applying 5 different machine learning classifiers algorithms to Wisconsin Breast Cancer Diagnostic dataset (WBCD) and extracting the unneeded features we found that the number of input features we will use are 8 features and 1 output feature which represent the diagnosis (Benign or Malignant) and the algorithm we will use will be the Support Vector Machine (SVM) which gave us the highest accuracy at test size 25% with test accuracy (97.2%) and average precision (97%) predicting 139 cases correctly from 143 case from the test set.

2. Breast Cancer Histopathological Image Classification Model

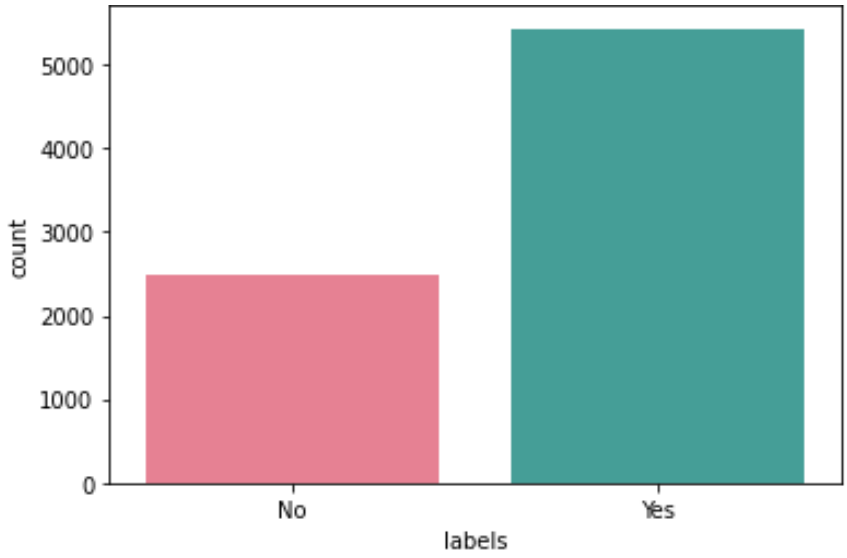
2.1. Methodology

- Transfer learning

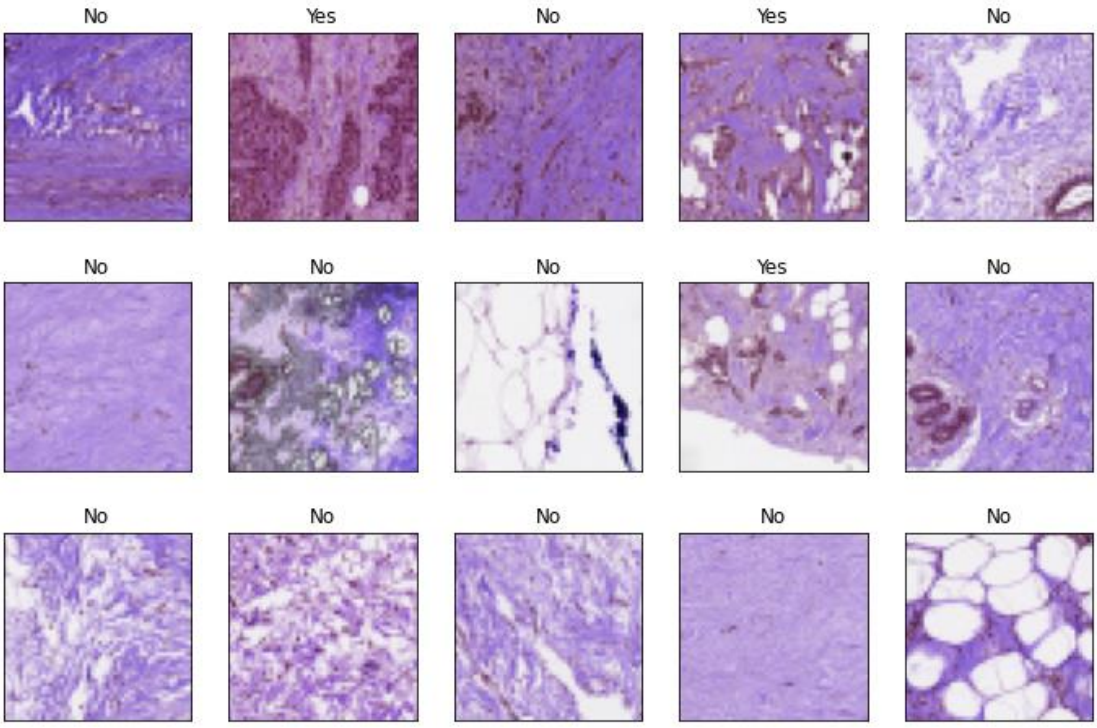


2.2. Dataset Description

- The Breast Cancer Histopathological Image Classification (BreakHis) is composed of 7909 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors, it contains 2,480 benign and 5,429 malignant samples. This database has been built in collaboration with the P&D Laboratory – Pathological Anatomy and Cytopathology, Parana, Brazil.
- The dataset BreaKHis is divided into two main groups: benign tumors and malignant tumors.
- The samples present in dataset were collected by surgical open biopsy method
- In this project we used 7909 pictures from the BreakHis dataset classified as follows:



2.3. Dataset samples



2.4. Data preprocessing

- Data splitting: 85% for training - 5% for testing – 10% for validation
- Data augmentation: To increase the efficiency of the model
- Images resizing: Make all images of size ($224\ x\ 224$)

```
[ ] print(f"test set shape: {test.shape}")
    print(f"validation set shape: {valid.shape}")

    train set shape: (6761, 2)
    test set shape: (396, 2)
    validation set shape: (752, 2)

[ ]
    train_datagen = ImageDataGenerator(rescale = 1./255.,rotation_range = 40, width_shift_range = 0.2, height_shift_range = 0.2,
                                       shear_range = 0.2, zoom_range = 0.2, horizontal_flip = True, vertical_flip =True)
    test_datagen = ImageDataGenerator(rescale = 1.0/255.)

    train_gen = train_datagen.flow_from_dataframe(dataframe = train_new,
                                                  x_col = 'filepaths', y_col ='labels',
                                                  target_size = (224,224), batch_size = 32,
                                                  class_mode = 'binary', shuffle = True)
    val_gen = train_datagen.flow_from_dataframe(valid,
                                                target_size=(224,224), x_col = 'filepaths', y_col ='labels',
                                                class_mode='binary',
                                                batch_size= 16, shuffle=True)
    test_gen = test_datagen.flow_from_dataframe(test,
                                                target_size = (224,224), x_col = 'filepaths', y_col ='labels',
                                                 class_mode = 'binary',
                                                batch_size = 16, shuffle = False)

    Found 6761 validated image filenames belonging to 2 classes.
    Found 752 validated image filenames belonging to 2 classes.
    Found 396 validated image filenames belonging to 2 classes.
```

2.5. Transfer learning

- Transfer Learning is a research problem in machine learning that focuses on storing knowledge gained from pretrained models.
- ResNet-50 is a convolutional neural network that is 50 layers deep. You can load a pre-trained version of the network trained on more than million images from the ImageNet database. The pre-trained network can classify images of 1000 object categories such as

keyboard, mouse, and many animals. Thus, the network has learned rich feature representations for a wide range of images

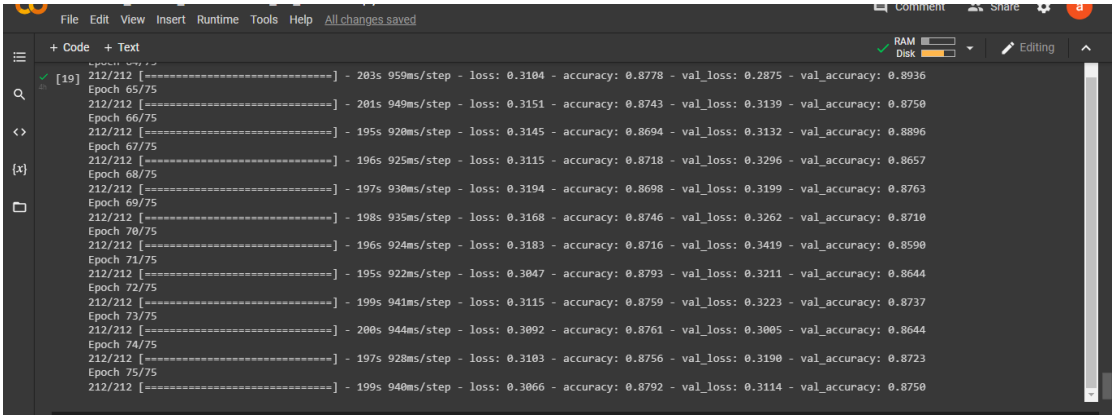
2.6. CNN parameters

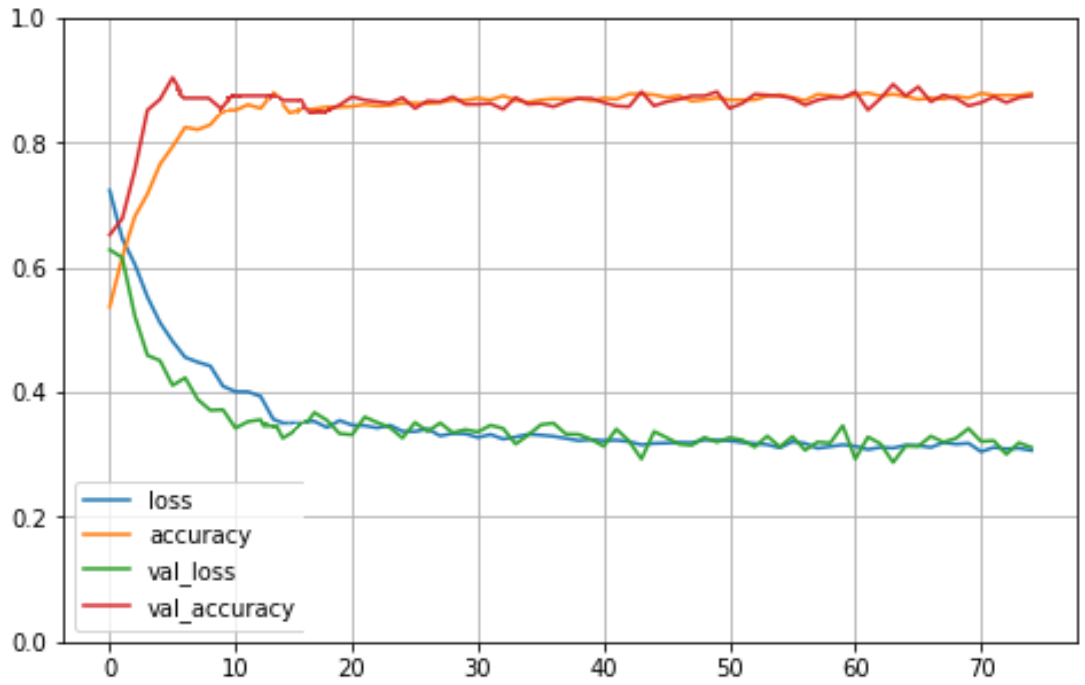
- We removed the input and the output layers of ResNet-50.
- The output is a binary classification (Benign and Malignant), thus we used sigmoid activation function.
- The weights loaded from ImageNet database

```
Downloading data from https://storage.googleapis.com/tensorflow/keras-
94674944/94668760 [==============================] - 1s 0us/step
94683136/94668760 [==============================] - 1s 0us/step
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_2 (InputLayer)        [(None, 224, 224, 3)]     0

 resnet50v2 (Functional)     (None, 7, 7, 2048)        23564800

 global_average_pooling2d (G  (None, 2048)             0
 lobalAveragePooling2D)

 dropout (Dropout)           (None, 2048)              0

 dense (Dense)               (None, 1)                 2049

=================================================================
Total params: 23,566,849
Trainable params: 2,049
Non-trainable params: 23,564,800
_________________________________________________________________
```

2.7 Results and Evaluation

- The average test accuracy obtained from the model was 0.89 and loss 0.26, the average Validation accuracy was 0.91 and validation loss 0.23.
- We chose the learning rate = 0.0001 and performed 75 epochs.





```
from PIL import Image
model_path = "model.h5"
loaded_model = tf.keras.models.load_model(model_path)

# import matplotlib.pyplot as plt
import numpy as np

image = cv2.imread("/content/drive/MyDrive/colab_notebook/10264/0/10264_idx5_x101_y1401_class0.png")

image_fromarray = Image.fromarray(image, 'RGB')
resize_image = image_fromarray.resize((224, 224))
expand_input = np.expand_dims(resize_image,axis=0)
input_data = np.array(expand_input)
input_data = input_data/255

pred = loaded_model.predict(input_data)
if pred >= 0.5:
  print("Yes")
else:
  print("No")
```

```
No
```

```
from PIL import Image
model_path = "model.h5"
loaded_model = tf.keras.models.load_model(model_path)

# import matplotlib.pyplot as plt
import numpy as np

image = cv2.imread("/content/drive/MyDrive/colab_notebook/10264/1/10264_idx5_x1001_y1151_class1.png")

image_fromarray = Image.fromarray(image, 'RGB')
resize_image = image_fromarray.resize((224, 224))
expand_input = np.expand_dims(resize_image,axis=0)
input_data = np.array(expand_input)
input_data = input_data/255

pred = loaded_model.predict(input_data)
if pred >= 0.5:
  print("Yes")
else:
  print("No")
```

```
Yes
```

3. Risk assessment Model

3.1. What is risk assessment tool?

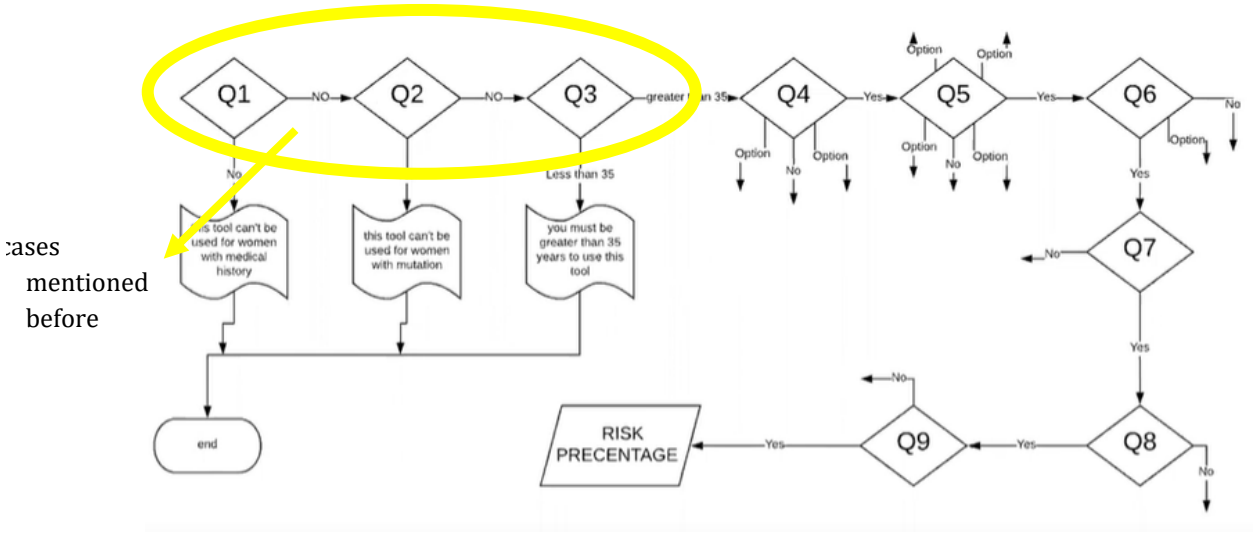
- Based on a statistical model known as the Gail Model
- The Gail model is based on the data from the breast cancer screening study that involved more than 280,000 females of different ages and ethnicity. The tool estimates patients' risk based on the following personal and family information:
    - Age
    - Age at the start of menstruation
    - Age at first live birth of a child
    - Number of first-degree relative (mother, sisters, daughters) with breast cancer
    - Number of previous breast biopsies (whether positive or negative)
    - Presence of atypical hyperplasia in a biopsy
- Combination of question, every question has a weight according to its answer
- Used to estimate the likelihood of a woman developing breast cancer

3.2. Cases that the tool cannot deal with:

- Women carrying brca1 or brca2

- Women under 35 years
- Women with previous medical history (radiation exposure)

3.3. Flowchart



cases mentioned before

3.4. Problems

- All the research papers worked on private datasets.

3.5. Creating a dataset (reverse engineering)

- Using Brexa app (online risk assessment tool).
- By building a truth table with some of the possible combinations of answers.
- Try all the answers in the truth table on the app manually to find the risk values.

3.6. Generating possible combinations

- Using excel tables

Q4	Q5	Q6	Q7	Q8	Q9
unknown	unknown	no	yes	one	yes
7to11	nobirths	one	no	greaterone	no
12to13	less20	graterone			
greater14	20to24				
	25to29				
	less35				

VIEW LIST OF COMBINATIONS BELOW

LIST
unknown unknown no yes one yes
unknown unknown no yes one no
unknown unknown no yes greaterone yes
unknown unknown no yes greaterone no
unknown unknown no no one yes
unknown unknown no no one no
unknown unknown no no greaterone yes
unknown unknown no no greaterone no
unknown unknown one yes one yes
unknown unknown one yes one no
unknown unknown one yes greaterone yes
unknown unknown one yes greaterone no
unknown unknown one no one yes
unknown unknown one no one no
unknown unknown one no greaterone yes
unknown unknown one no greaterone no
unknown unknown graterone yes one yes
unknown unknown graterone yes one no
unknown unknown graterone yes greaterone yes
unknown unknown graterone yes greaterone no
unknown unknown graterone no one yes
unknown unknown graterone no one no
unknown unknown graterone no greaterone yes
unknown unknown graterone no greaterone no
unknown nobirths no yes one yes

3.7. Building the dataset

- After trying the combinations manually on Brexa app

4	5	6	7	8	9	sult
unkown	unkown	no	no			0.3624
unkown	unkown	no	yes	one	no	0.0617
unkown	unkown	no	yes	one	yes	0.0617
unkown	unkown	no	yes	graeter one	no	0.1046
unkown	unkown	no	yes	graeter one	yes	0.1064
unkown	unkown	one	no			0.9469
unkown	unkown	one	yes	one	no	0.1602
unkown	unkown	one	yes	one	yes	0.1602
unkown	unkown	one	yes	graeter one	no	0.2704
unkown	unkown	one	yes	graeter one	yes	0.2704
unkown	unkown	greater one	no			0.245
unkown	unkown	greater one	yes	one	no	0.4132
unkown	unkown	greater one	yes	one	yes	0.4132
unkown	unkown	greater one	yes	graeter one	no	0.6899
unkown	unkown	greater one	yes	graeter one	yes	0.6899

3.8. Dataset

- The first column is the risk ratio (output).
- From column 2 to column 6 are the multiple choices questions.

```
# import data
df= pd.read_excel("F:/college/Grad/risk model/data.xlsx")
df.head(7)
```

	result	Q9	Q8	Q7	Q6	Q5	Q4
0	0.3624	NaN	NaN	no	no	unkown	unkown
1	0.0617	no	one	yes	no	unkown	unkown
2	0.0617	yes	one	yes	no	unkown	unkown
3	0.1046	no	graeter one	yes	no	unkown	unkown
4	0.1064	yes	graeter one	yes	no	unkown	unkown
5	0.9469	NaN	NaN	no	one	unkown	unkown
6	0.1602	no	one	yes	one	unkown	unkown

3.9. Dataset preparation

- Label encoding

```
#label encoding
from sklearn import preprocessing
label_encoder= preprocessing.LabelEncoder()
df['Q4']= label_encoder.fit_transform(df['Q4'])
df['Q5']= label_encoder.fit_transform(df['Q5'])
df['Q6']= label_encoder.fit_transform(df['Q6'])
df['Q7']= label_encoder.fit_transform(df['Q7'])
df['Q8']= label_encoder.fit_transform(df['Q8'].astype(str))
df['Q9']= label_encoder.fit_transform(df['Q9'].astype(str))
```

- Dataset in numeric form

```
df.head(7)
```

	result	Q9	Q8	Q7	Q6	Q5	Q4
0	0.3624	0	1	0	1	0	0
1	0.0617	1	2	1	1	0	0
2	0.0617	2	2	1	1	0	0
3	0.1046	1	0	1	1	0	0
4	0.1064	2	0	1	1	0	0
5	0.9469	0	1	0	2	0	0
6	0.1602	1	2	1	2	0	0

3.10. Splitting the factors and results

```
#splitting the factors and the result
x=df.drop(columns='result', axis=1)
y=df['result']
```

```
print(y)
```

```
0      0.6176
1      0.6176
2      0.1460
3      0.1460
4      0.0362
        ...
243    0.2549
244    0.2549
245    0.4289
246    0.4289
247    0.1509
Name: result, Length: 248, dtype: float64
```

3.11. Splitting the data (training and testing)

```
#splitting the data
x_train, x_test, y_train, y_test = train_test_split( x, y, test_size = 0.2, random_state=5)
```

```
print(x.shape,x_train.shape,x_test.shape)
```

```
(248, 6) (198, 6) (50, 6)
```

3.12. Regression algorithms comparison

```
#linear regression
# importing r2_score module
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
# predicting the accuracy score
score=r2_score(y_test,y_prediction)
print("test variance score is: %0.2f"% score)
print("mean_sqrd_error is: %0.2f"% mean_squared_error(y_test,y_prediction))
```

```
test variance score is: 0.39
mean_sqrd_error is: 0.02
```

```
#decision tree
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
# predicting the accuracy score
score=r2_score(y_test,y_prediction)
print("test variance score is: %0.2f"% score)
print("mean_sqrd_error is: %0.2f"% mean_squared_error(y_test,y_prediction))
```

```
test variance score is: 0.59
mean_sqrd_error is: 0.01
```

```
#svm
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
# predicting the accuracy score
score=r2_score(y_test,y_prediction)
print("test variance score is: %0.2f"% score)
print("mean_sqrd_error is: %0.2f"% mean_squared_error(y_test,y_prediction))
```

```
test variance score is: 0.69
mean_sqrd_error is: 0.01
```

```
#random forest
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
# predicting the accuracy score
score=r2_score(y_test,y_prediction)
print("test variance score is: %0.2f"% score)
print("mean_sqrd_error is: %0.2f"% mean_squared_error(y_test,y_prediction))
```

```
test variance score is: 0.88
mean_sqrd_error is: 0.00
```

3 Conclusion

Algorithm	variance	Mean sqrd error
Linear regression	0.39	0.02
Decision tree	0.57	0.01

Support vector machine	0.69	0.01
Random forest	0.88	0.00

- The final coefficients

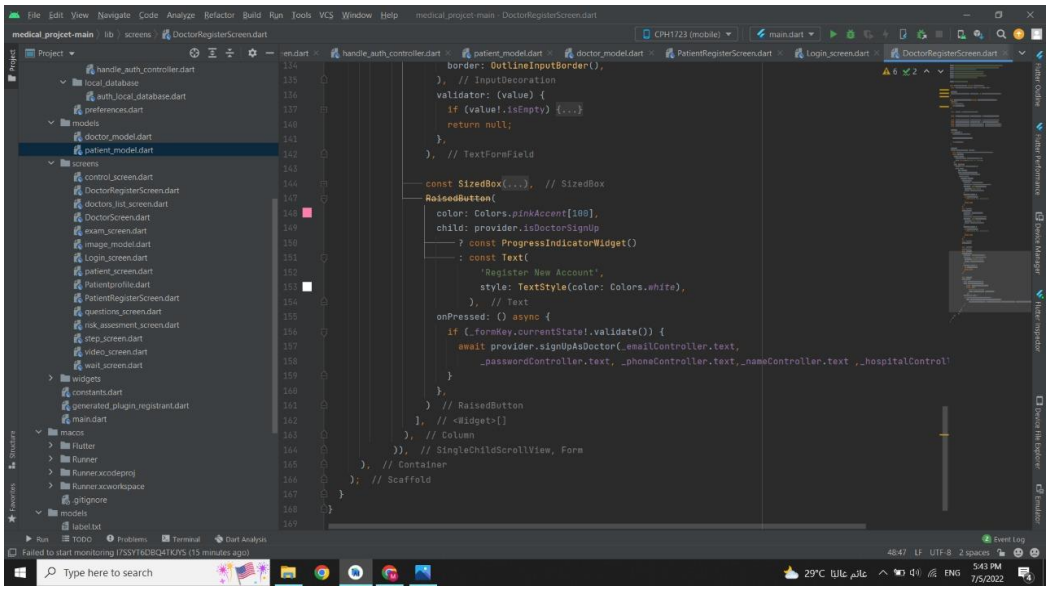
C1	C2	C3	C4	C5	C6
0.04900	0.08818	0.46331	0.0266	0.28366	0.11584

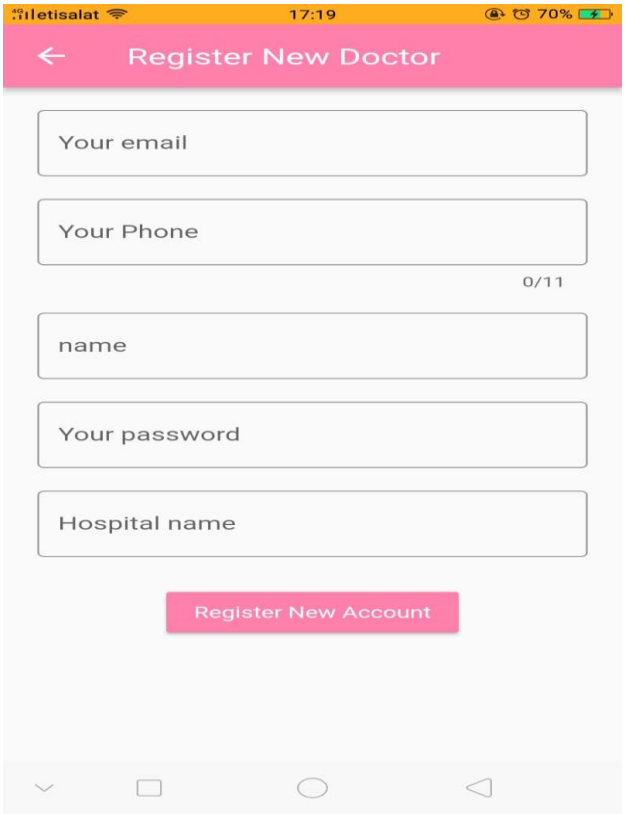
- The equation of decision tree regressor:

[(c1*Q4) +(c2*Q5) +(c3*Q6) +(c4*Q7) +(c5*Q9) +(c6*Q9)]

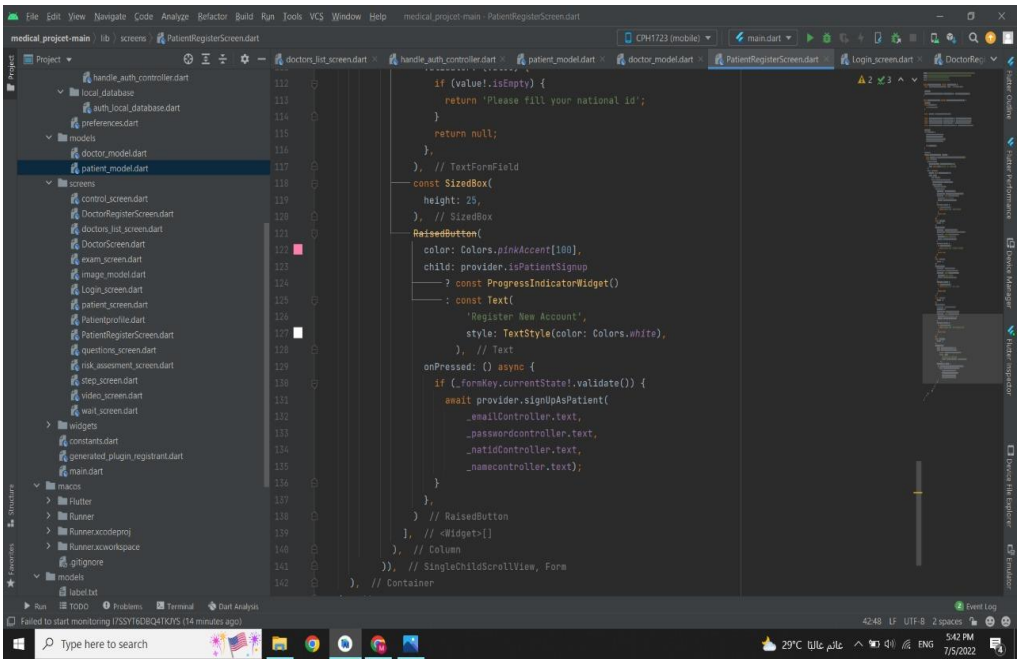
MOBILE APPLICATION UI

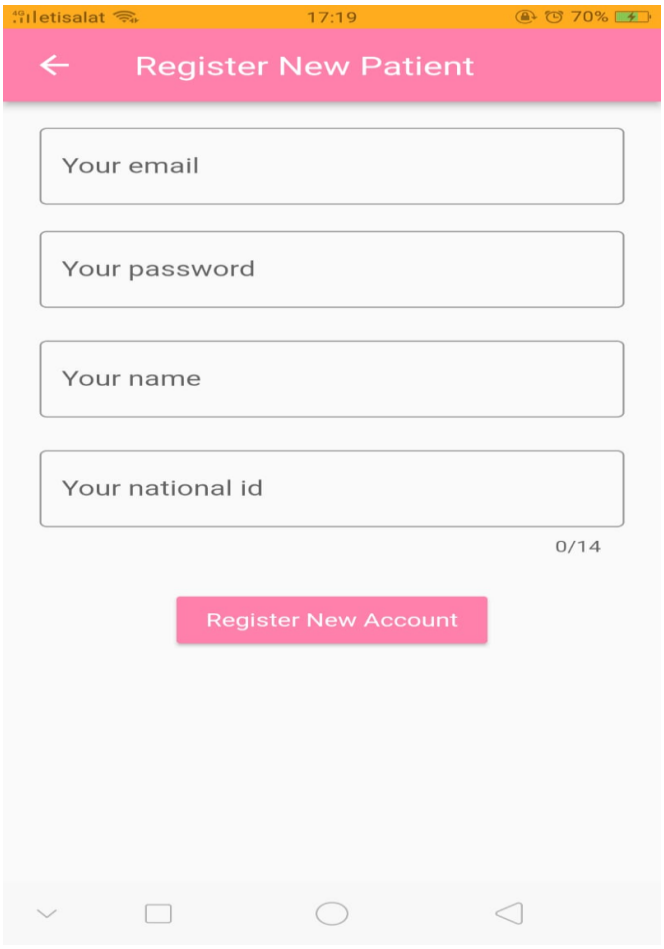
1. Doctor register screen



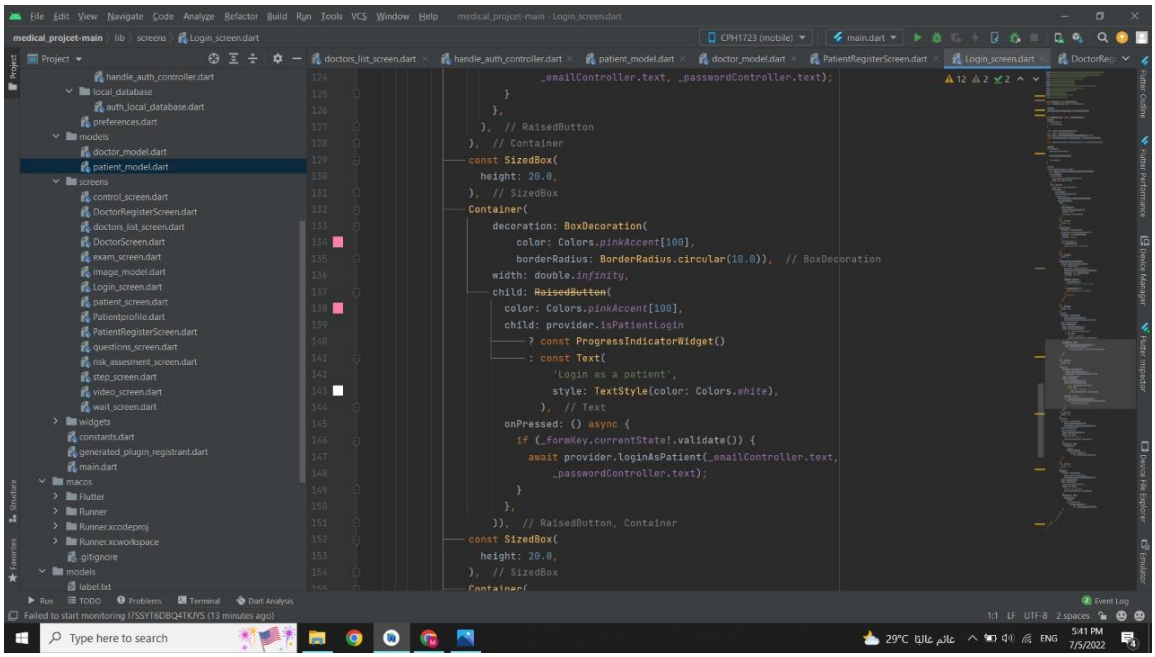


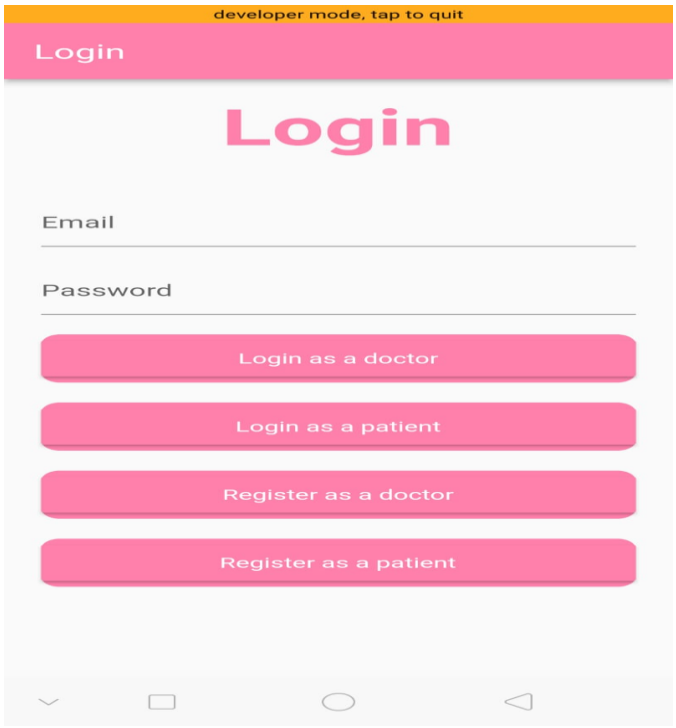
2. patients register screen



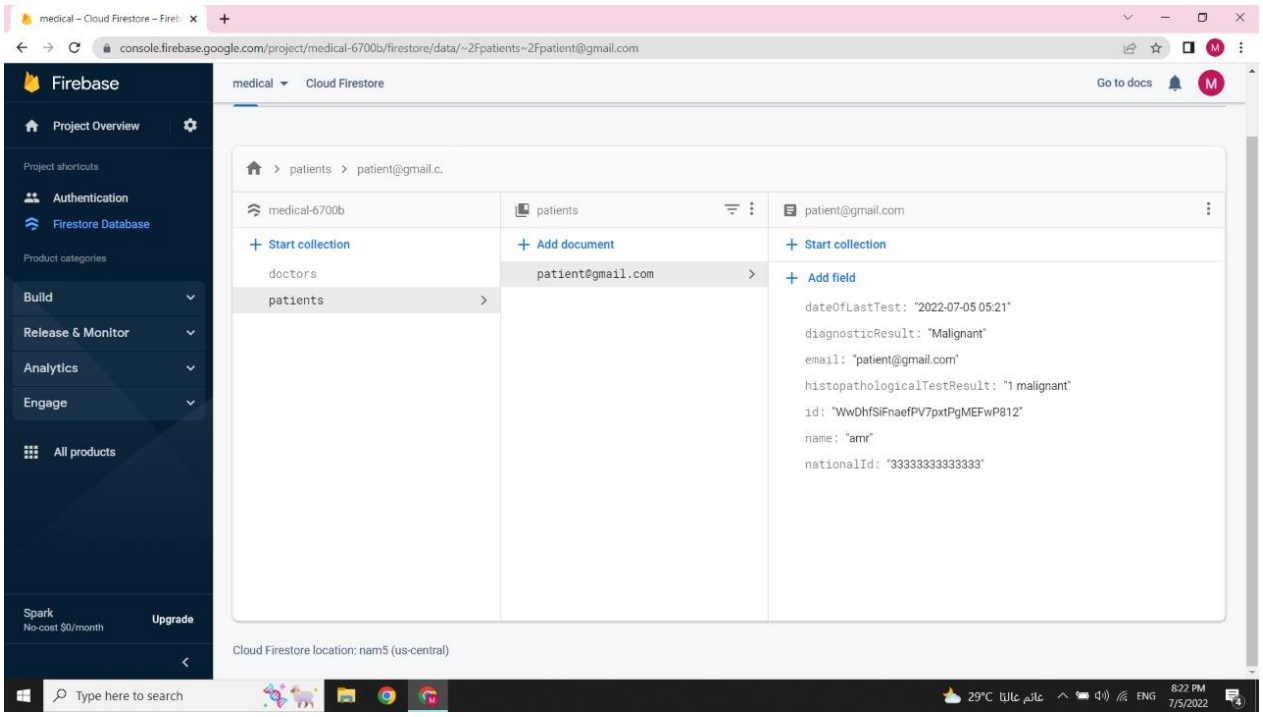


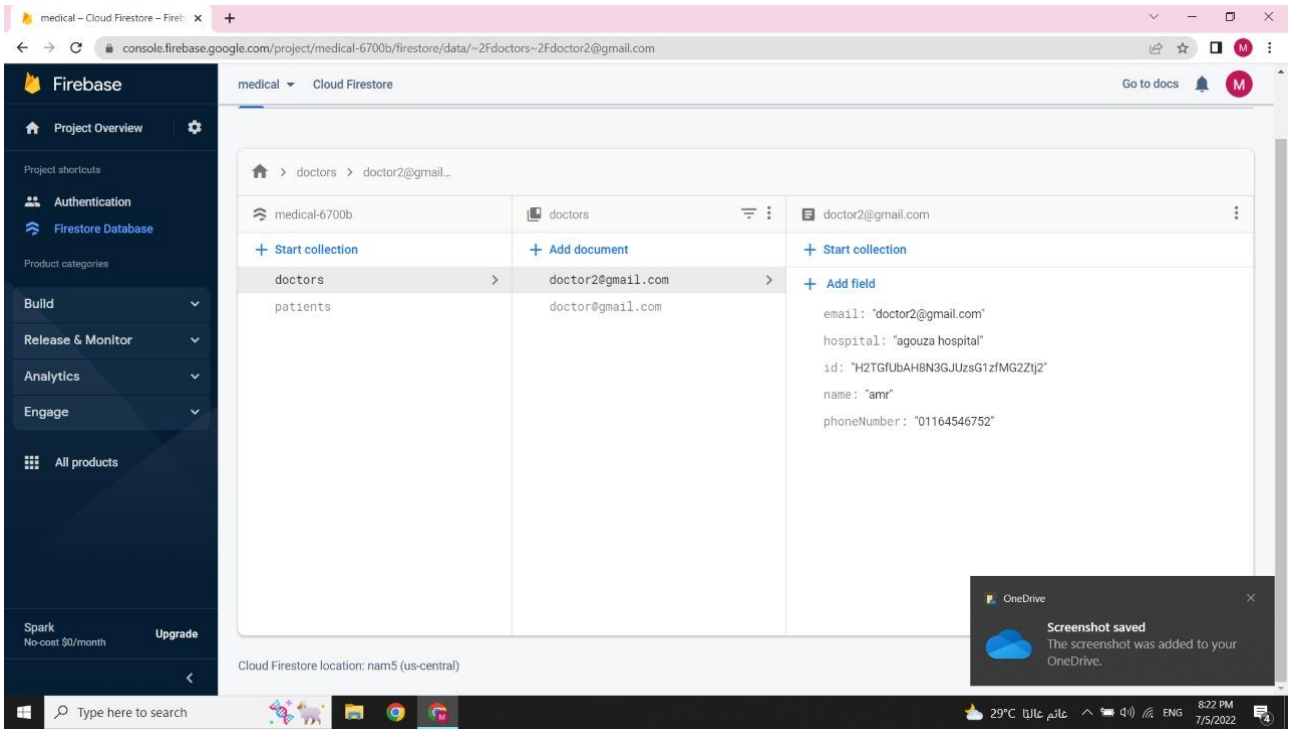
3. login screen



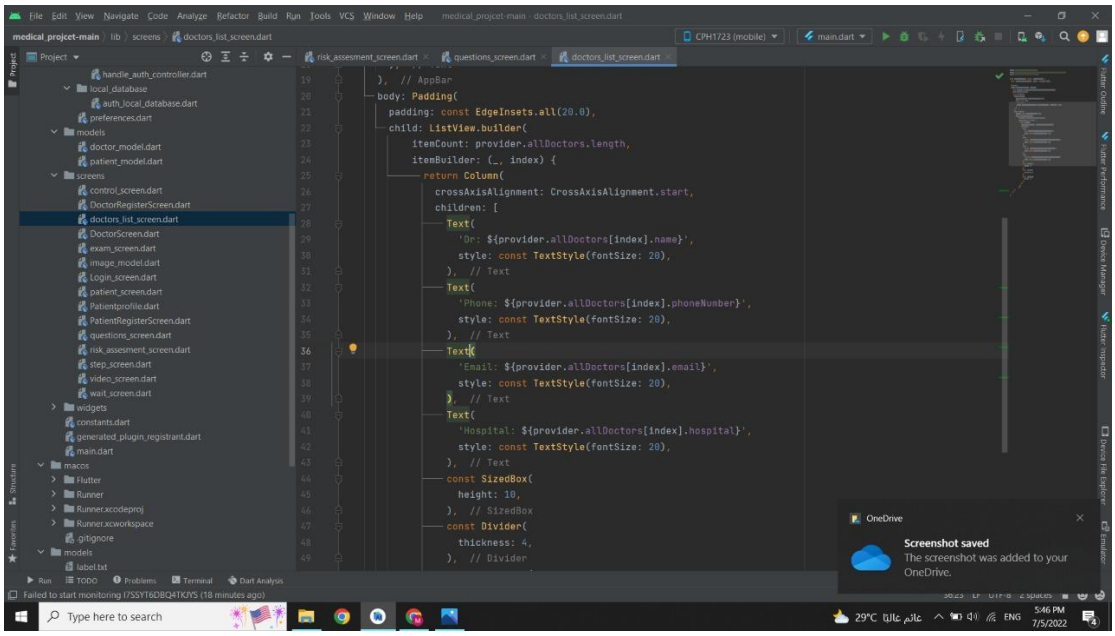


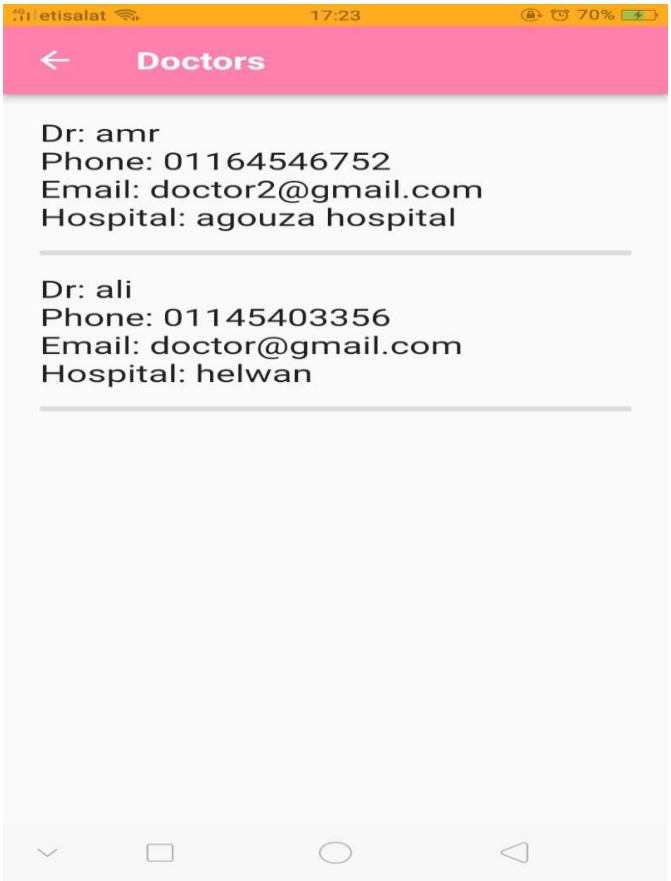
4. firebase record



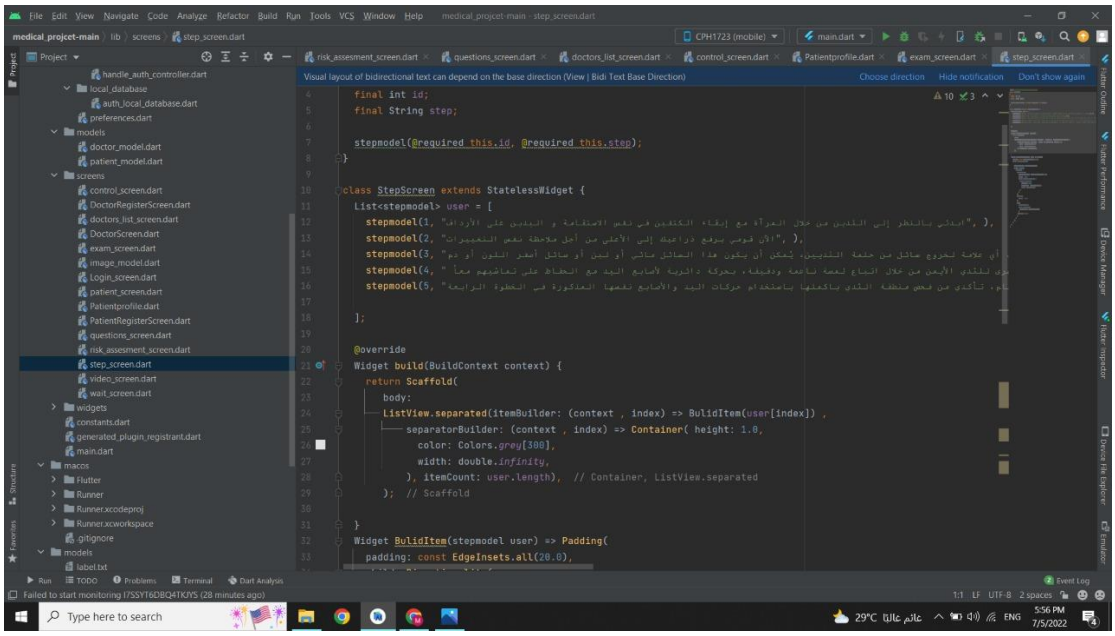


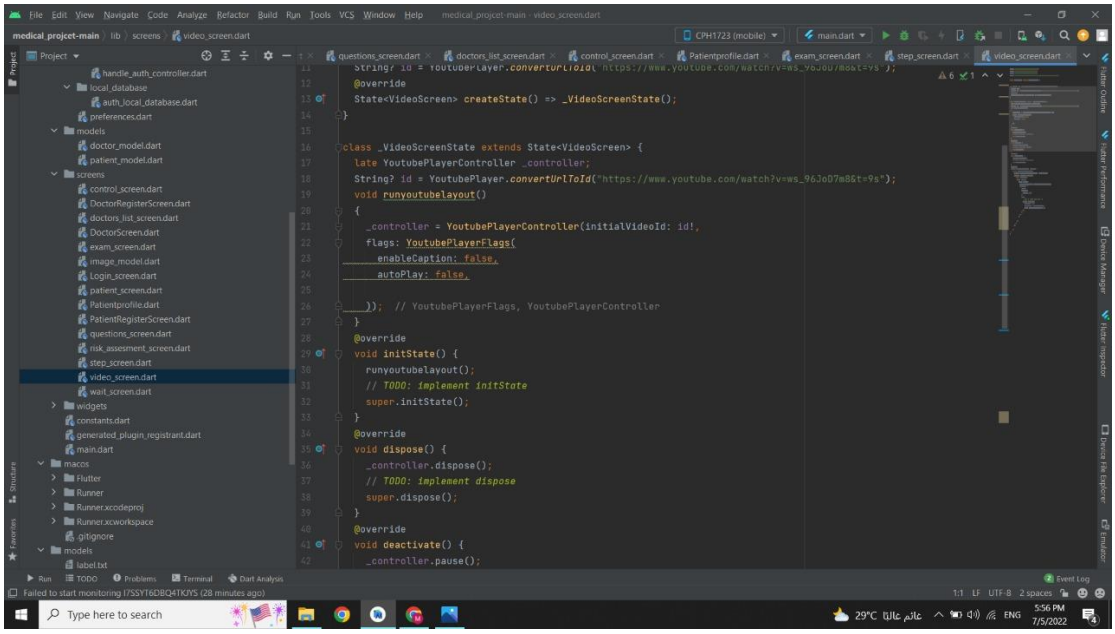
5. doctors list screen



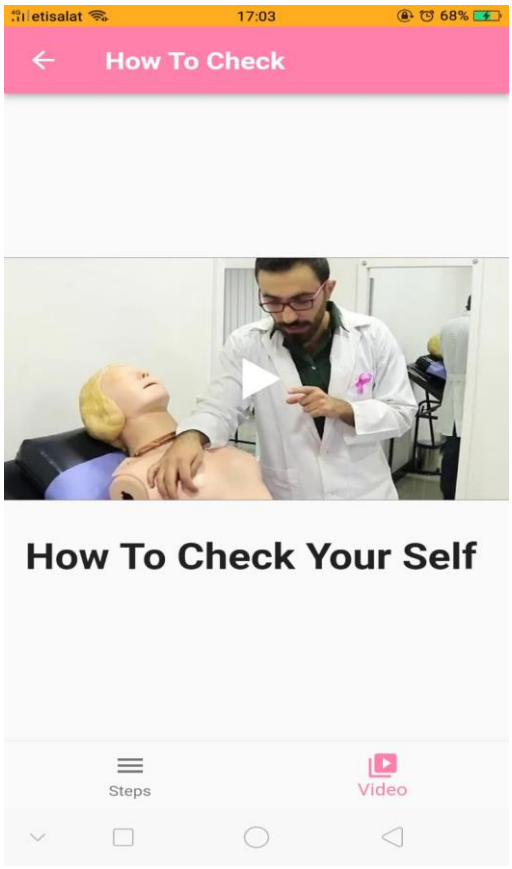


6. self-examination screen

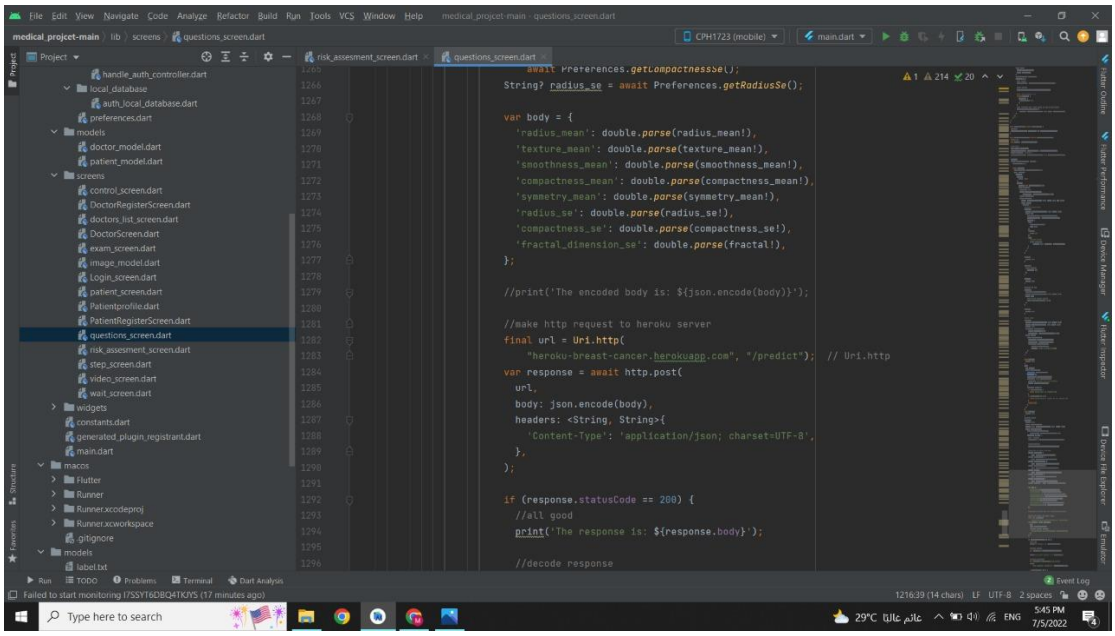


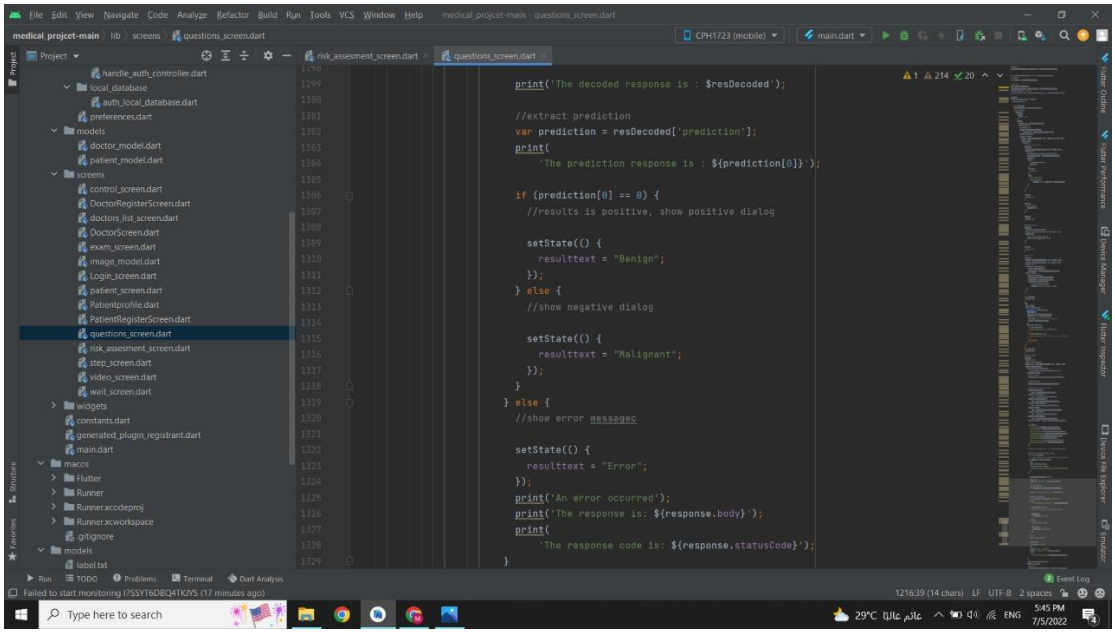


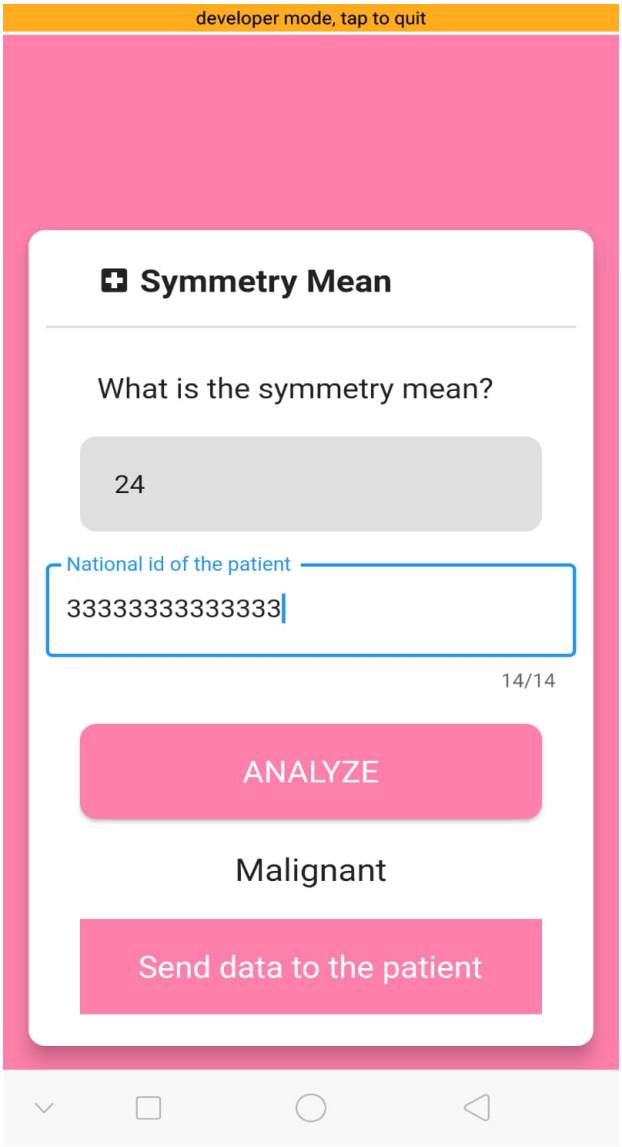

developer mode, tap to quit
How To Check
1
ابدئي بالنظر إلى الثدين من خلال المرآة مع إبقاء الكتفين في نفس الاستقامة و اليدين على الأرداف
2
الآن قومي برفع ذراعيك إلى الأعلى من أجل ملاحظة نفس التغييرات
3
أثناء النظر في المرآة، قومي بملاحظة ما إذا كان هناك أي علامة لخروج سائل من حلمة الثديين، يُمكن أن يكون هذا السائل مائي أو لبن أو سائل أصفر اللون أو دم
4
Steps
Video


etisalat
17:03
68%
How To Check
How To Check Your Self
Steps
Video

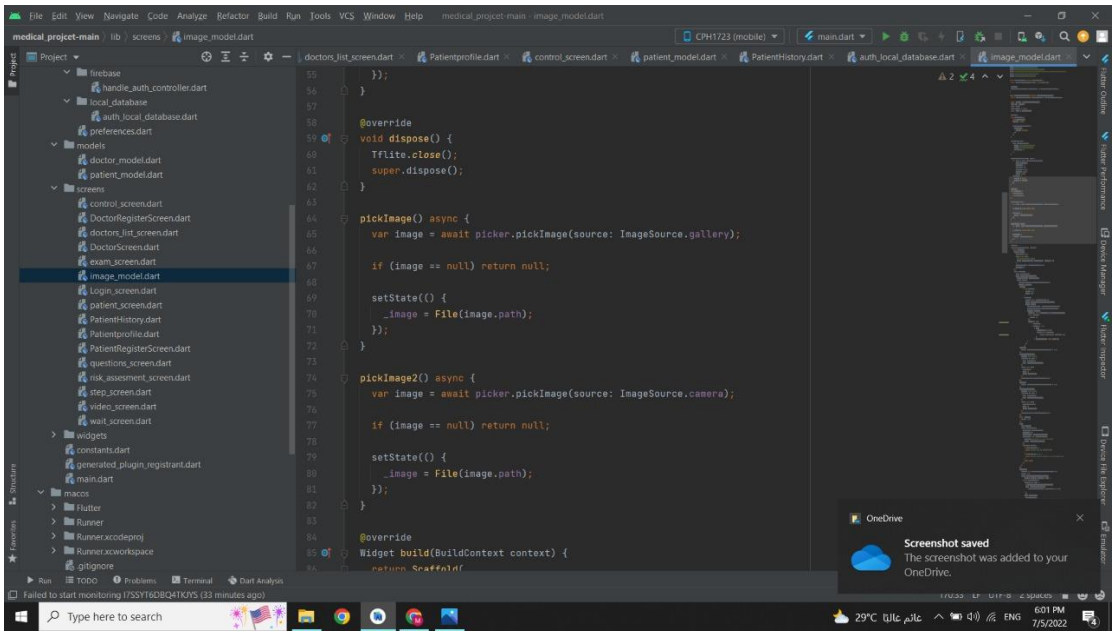
7. diagnostic test

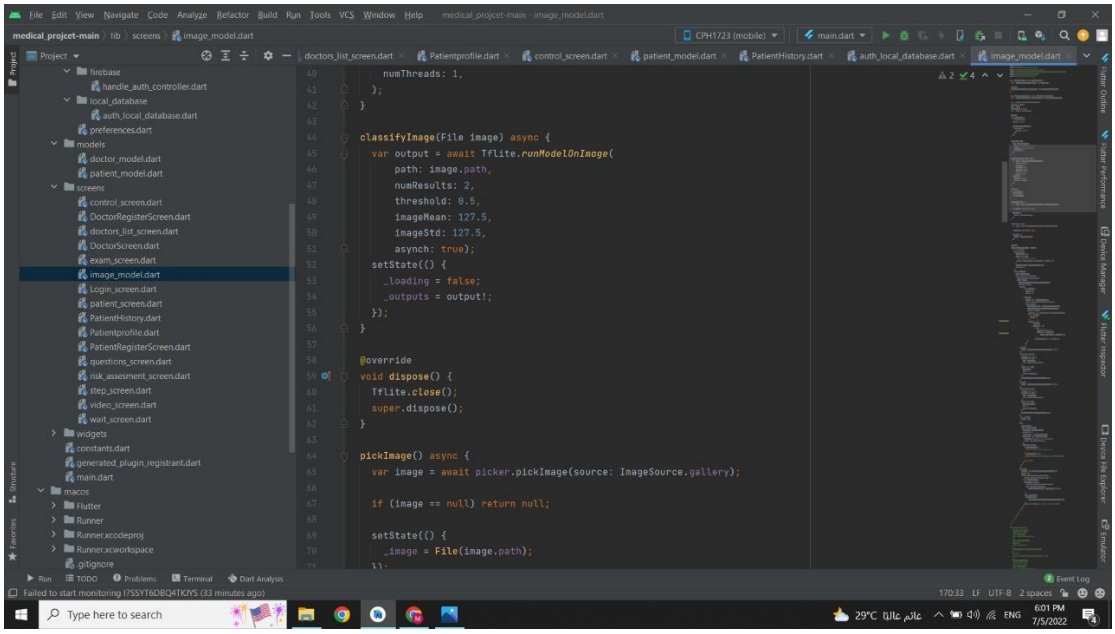


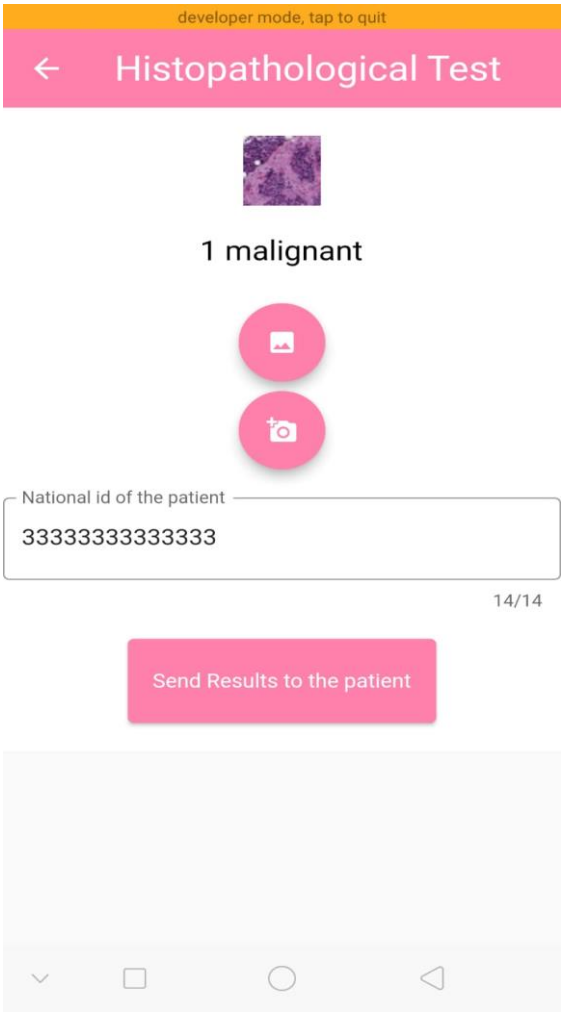



developer mode, tap to quit
Symmetry Mean
What is the symmetry mean?
24
National id of the patient
33333333333333
14/14
ANALYZE
Malignant
Send data to the patient

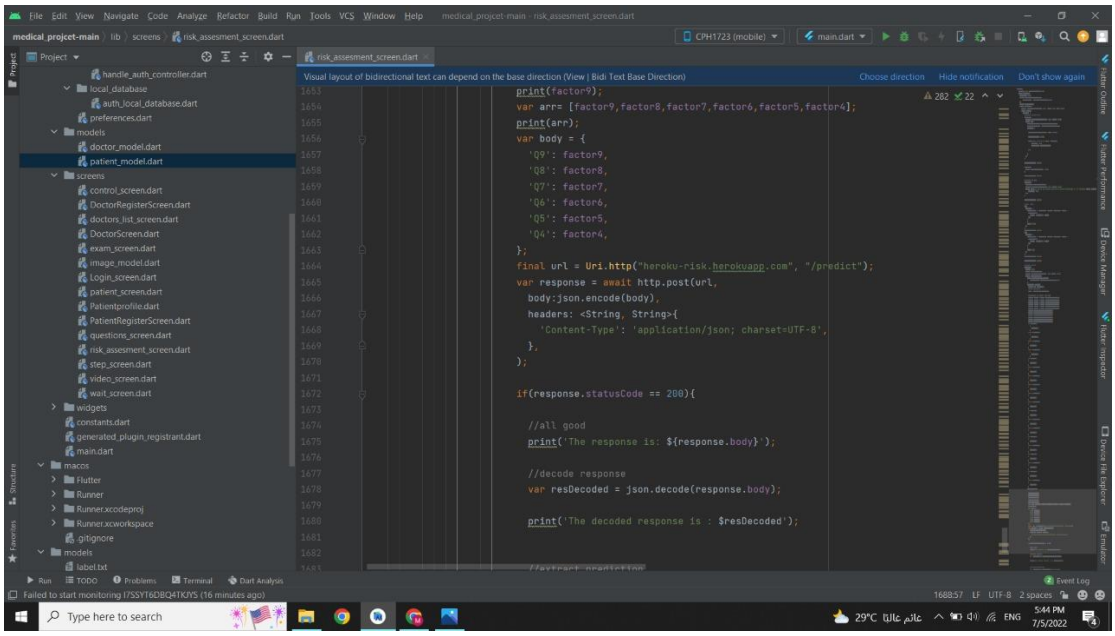
8. histopathological test

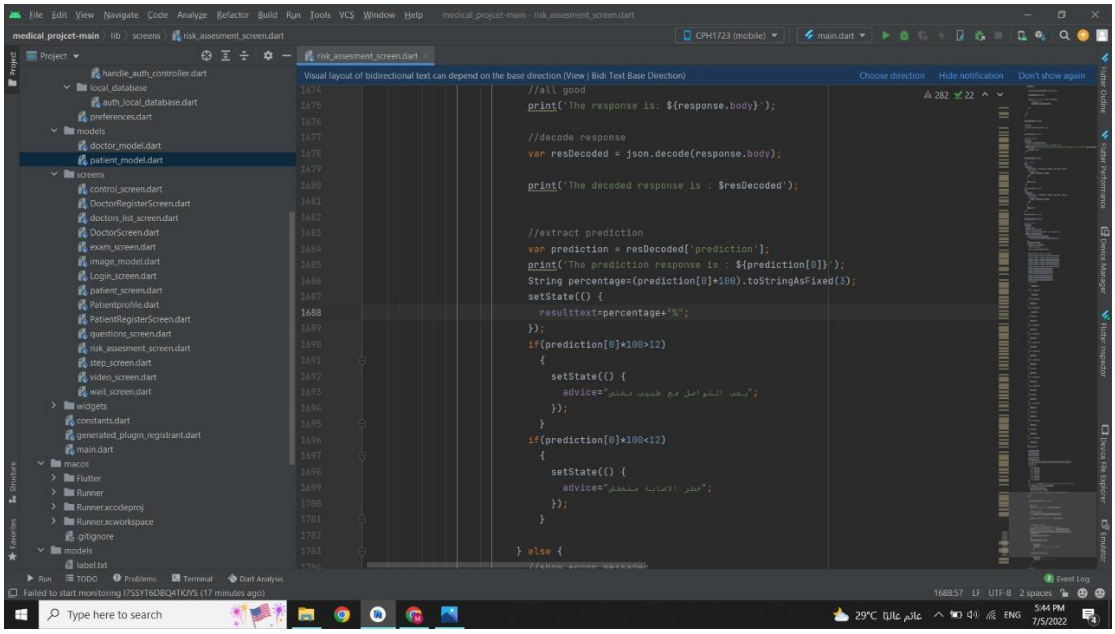


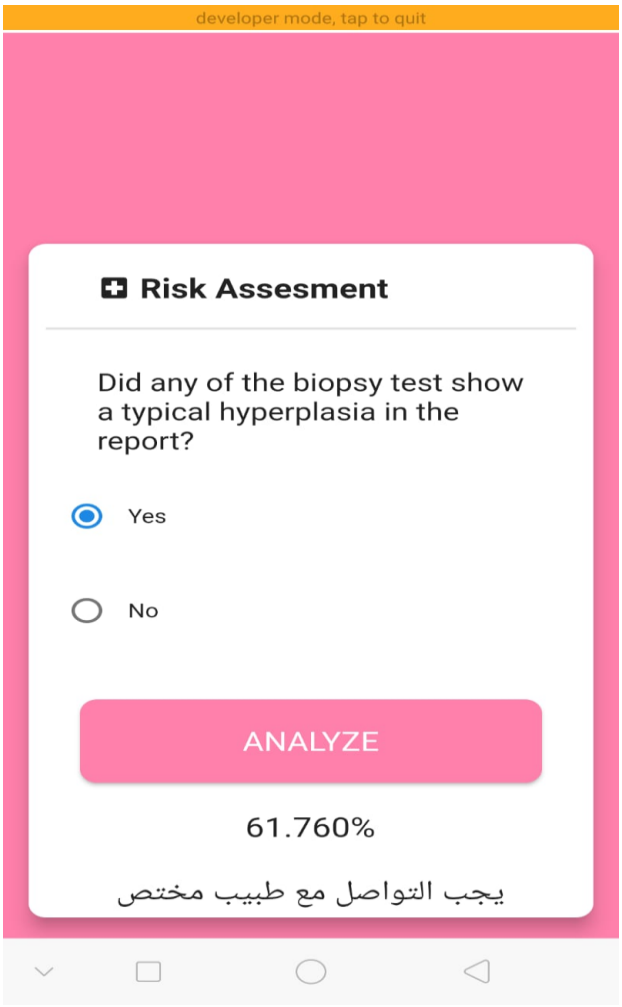



developer mode, tap to quit
Histopathological Test
1 malignant
National id of the patient
33333333333333
14/14
Send Results to the patient

9. Risk assessment tool






developer mode, tap to quit
Risk Assesment
Did any of the biopsy test show a typical hyperplasia in the report?
Yes
No
ANALYZE
61.760%
يجب التواصل مع طبيب مختص

CHAPTER 5: TESTING AND VALIDATION

Software testing is the process of evaluating and verifying that a software product or application does what it is supposed to do. The benefits of testing include preventing bugs, reducing development costs, and improving performance.

There are many different types of software tests, each with specific objectives and strategies:

• **Acceptance testing:** Verifying whether the whole system works as intended.

• **Integration testing:** Ensuring that software components or functions operate together.

• **Unit testing:** Validating that each software unit performs as expected. A unit is the smallest testable component of an application.

• **Functional testing:** Checking functions by emulating business scenarios, based on functional requirements. Black-box testing is a common way to verify functions.

REQUIREMENTS TESTING

1. **sign up/ login as a doctor**

Testcase Id	Test Objective	Preconditions	Steps	Test data	Expected Result	Actual Result
TC_01	Successful doctor Login to The System	Valid Email and password	1.Enter Email 2.Enter Password 3.Click Login Button	Email = doctor@gmail.com Password =123456	Login Successfully	Login Successfully
TC_02	Make sure that message will appear to tell you that you should enter a valid email or password when you enter email not in the database.	Invalid email or password	1.Enter Email 2.Enter Password 3.Click Login	Email = doctor123@gmail.com Password =123456	Message will appear to tell you that this email or password is not valid, there is no such user in the database with this email and return to log in page again	Massage appeared
TC_03	make sure that the login can't be done while leaving	no email or password	click login without entring		message will appear to tell the	message appeared

	the e-mail and password cells empty		any data		user to enter valid email and password	
TC_04	Make a successful sign up.	valid mail, password, phone number, specialize or degree, clinic or hospital address	1. enter email 2.enter password 3.enter phone number 4.enter degree or specialize 4.enter clinic or hospital address 5.click sign up	email= doctor@gmail.com password=123456 phone number= 01123456789 specialize=surgery address= horryast	Signed up successfully , home page will open	Signed up successfully , home page opened

2. **sign up/ login as a patient**

Testcase Id	Test Objective	Preconditions	Steps	Test data	Expected Result	Actual Result
TC_05	Successful doctor Login to The System	Valid Email and password	1.Enter Email 2.Enter Password 3.Click Login Button	Email = patient@gmail.com Password =123456	Login Successfull y	Login Successfully

TC_06	Make sure that message will appear to tell you that you should enter a valid email or password when you enter email not in the database.	Invalid email or password	1.Enter Email 2.Enter Password 3.Click Login	Email = patient123@gmail.com Password =123456	Message will appear to tell you that this email or password is not valid, there is no such user in the database with this email and return to log in page again	Massage appeared
TC_07	make sure that the login can't be done while leaving the e-mail and password cells empty	no email or password	click login without entring any data		message will appear to tell the user to enter valid email and password	message appeared
TC_08	Make a successful sign up.	valid mail, password,14 digits national id	1. enter email 2.enter password 3.enter national id	email = patient@gmail.com password =123456 natio	Signed up successfully , home page will open	Signed up successfully , home page opened

				nal id= 4444 4444 4444 44		
TC_09	make sure the patient is registering with a correct national id	valid mail, password,less or more than 14 digits national id	1. enter email 2.enter password 3.enter national id	email = patient@gmail.com password =123456 national id= 4444 4444	message will appear to warn the user to enter 14 digits in the national id cell	message appeared

3. Check breast cancer risk

Testcase Id	Test Objective	Preconditions	Steps	Test data	Expected Result	Actual Result
TC_10	make sure that only the user with the right conditions can take the test	Valid Email and password to login as a patient	1.Login as a patient 2.choose the"more than 35" answer for the		the rest of the questions will appear and the user can complete the test	the rest of the questions appeared

			first question 3. choose no for the next 2 questions			
TC_11	make sure that only the user with the right conditions can take the test	Valid Email and password to login as a patient	1.Login as a patient 2.choose the"less than 35" answer for the first question 3.choose yes for the next 2 questions		Message will appear to tell the user that this test can't deal with their conditions	Massage appeared

4. tutorial for self-examination

Testcase Id	Test Objective	Preconditions	Steps	Test data	Expected Result	Actual Result
TC_12	make sure that users can	Valid Email and password	1.Login as a		the self examinat	the self exami

	use the feature of self-examination tutorial	to login as a patient	patient 2.click on the self-examination icon		ion tutorial will appear and the user will have to choose between written or video	nation tutorial appeared

5. contact a doctor

Testcase Id	Test Objective	Preconditions	Steps	Test data	Expected Result	Actual Result
TC_13	make sure that users can use the feature of contacting a doctor	Valid Email and password to login as a patient	1.Login as a patient 2.click on the doctor icon		the phone numbers and names of the registered doctor will appear to the user	the phone numbers and names of the registered doctor appeared

6. Breast cancer detection using fine needle aspirate test

Testcase Id	Test Objective	Preconditions	Steps	Test data	Expected	Actual Result

					Result	
TC_14	**make sure that users can use the feature of Breast cancer detection using fine needle aspirate test**	**Valid Email and password to login as a doctor**	**1.Login as a doctor** **2.enter the required values in their cells**		**the diagnosis will appear**	**the diagnosis appeared**

7. Breast cancer detection using histopathological images

Testcase Id	Test Objective	Preconditions	Steps	Test data	Expected Result	Actual Result
TC_15	**make sure that users can use the feature of Breast cancer detection using histopathological images**	**Valid Email and password to login as a doctor**	**1.Login as a doctor** **2.enter the url of the SOB test image**		**the diagnosis will appear**	**the diagnosis appeared**

CHAPTER 6: DEVELOPMENT AND TOOLS TECHNOLOGIES

FRAMEWORK

- **Flutter**

We built our application using **Flutter** open-source framework by Google because it's so powerful and flexible in making cross-platform mobile applications that runs on both android and iOS OS.

LANGUAGES

- **Dart**

Dart is the programming language designed for client development such as for web and mobile applications, **Dart** is used to code Flutter apps

- **Python**

We used the **Python** programming language to prepare our data and build our machine learning models because it is flexible and practical in dealing with Data.

MACHINE LEARNING MODELS DEPLOYMENT

- **Flask**

We will deploy our machine learning models on the mobile application by creating a local host using **Flask**

- **Heroku**

Deploy our Flask app on **Heroku** to deploy the application on a web server

ONLINE TOOLS

- **Google colab**

Google Collaboratory or “Colab” For short, is a product from Google Research. Colab allows users to write and execute arbitrary Python Code through the browser, and is especially well suited to machine Learning, data analysis and education. Colab requires

no setup to use, while providing free access to computing resources

including GPUs and Rams up to 12 GB and up to 12 hours which helps in training huge deep learning models.

- **Jupyter notebook**

The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text.

LIBRARIES

- **Data Analysis**

- **Data Visualization**

- **Data Splitting**

- **Working with arrays**

- **Deep learning framework**

- **Database**

MODELS

We used Machine Learning Technology to build and train our rating, prediction and image processing models. For the rating model we used multiple machine learning algorithms to choose the best algorithm for breast cancer detection those algorithms were:(Support vector machine, Decision tree, Random Forest, logistic regression) and for the prediction model we used regression algorithms and for image processing we are going to use CNN, NumPy and tensor flow Libraries.

CHAPTER 7: REFERENCES

1. Gunjan Rawa,Rakesh Rawal,Hirav Shah,Kamlesh Patel, 2019,"*A Comparative Study Between Artificial Neural Networks and Conventional Classifiers for Predicting Diagnosis of Breast Cancer*". In: Kumar A., Paprzycki M., Gunjan V. (eds) ICDSMLA . Lecture Notes in Electrical Engineering, vol 601. Springer, Singapore. https://doi.org/10.1007/978-981-15-1420-3_28

2. Daisuke Kimura,2018,"*Machine Learning Methods for Histopathological Image Analysis"* https://doi.org/10.1016/j.csbj.2018.01.001

3. Abien Fred M. Agarap,2017,"*On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset*" https://arxiv.org/abs/1711.07831

4. Kader potdar,September 2016,"*A Comparative Study of Machine Learning Algorithms applied to Predictive Breast Cancer Data*" https://www.researchgate.net/publication/308725638

5. MohammedAmineNajia,2021,"*Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis"* https://doi.org/10.1016/j.procs.2021.07.062

6. Fabio Spanhol,july 2016,"*Breast Cancer Histopathological Image Classification using Convolutional Neural Networks"* https://www.researchgate.net/publication/304158394

7. Jitendra Maan, Harsh Maan, " *Breast Cancer Detection using Histopathological images"*

https://www.researchgate.net/publication/358603508_Breast_Cancer_Detection_using_Histopathological_Images

8. Yassir Benhammoua, Boujemˆaa Achchab, Francisco Herreraa, SihamTabika, *"BreakHis based Breast Cancer Automatic Diagnosisusing Deep Learning: Taxonomy, Survey and Insights".*

https://www.researchgate.net/publication/335972754_BreakHis_based_Breast_Cancer_Automatic_Diagnosis_using_Deep_Learning_Taxonomy_Survey_and_Insights

9. Gigi F. Stark*," Predicting breast cancer risk using personal health data and machine learning models"*

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6934281/#!po=57.3864

10. Jeungok Choi, *"Development of a Breast Cancer Risk Assessment Model Using a Machine Learning Approach"*

https://www.himss.org/resources/development-breast-cancer-risk-assessment-model-using-machine-learning-approach?utm_campaign=general&utm_source=google&utm_medium=cpc&utm_term=_&adgroupid=134509372449&gclid=Cj0KCQjw5-WRBhCKARIsAAId9FkC2-6cBudEaaDt3MvKURD4dfV8wO6p-1K8tOQtcId3gkx52M7_5VwaAilwEALw_wcB

11. https://www.coursera.org/learn/neural-networks-deep-learning?specialization=deep-learning

12. https://www.coursera.org/learn/convolutional-neural-networks?specialization=deep-learning

13. https://www.udemy.com/course/complete-flutter-arabic/

14. https://www.udemy.com/course/flutter-deeplearning-course/.