A Mini Project Report on

Breast Cancer Prediction System

T.E. - I.T Engineering

Submitted By

Simran Choudhary	**19104023**
Janhavi Kulkarni	**19104062**
Loveritu Itnare	**19104028**

Under The Guidance Of

Prof. Kiran Deshpande & Ms. Jayshree Jha

DEPARTMENT OF INFORMATION TECHNOLOGY

A.P.SHAH INSTITUTE OF TECHNOLOGY

G.B. Road, Kasarvadavali, Thane (W), Mumbai-400615

UNIVERSITY OF MUMBAI

Academic year : 2021-22

CERTIFICATE

This to certify that the Mini Project report on **Breast Cancer Prediction System** has been submitted by Simran Choudhary (19104023), Janhavi Kulkarni (19104062), Loveritu Itnare (19104028) and who are a Bonafede students of A. P. Shah Institute of Technology, Thane, Mumbai, as a partial fulfilment of the requirement for the degree in **Information Technology**, during the academic year **2021-2022** in the satisfactory manner as per the curriculum laid down by University of Mumbai.

Prof. Kiran Deshpande & Ms. Jayshree Jha
Guide

Prof. Kiran Deshpande
Head Department of Information Technology

Dr. Uttam D.Kolekar
Principal

External Examiner(s)
1.
2.

Place: A.P Shah Institute of Technology, Thane
Date:

TABLE OF CONTENTS

Chapter 1

Introduction:

Cancer is a disease that occurs when there are changes or mutations that take place in genes that help in cell growth. These mutations allow the cells to divide and multiply in a very uncontrolled and chaotic manner. These cells keep increasing and start making replicas which end up becoming more and more abnormal. These abnormal cells later on form a tumor . Tumors, unlike other cells, don't die even though the body doesn't need them. The cancer that develops in the breast cells is called breast cancer. This type of cancer can be seen in the breast ducts or the lobules. Cancer canal so occur in the fatty tissue or the fibrous connective tissue within the breast. These cancer cells become uncontrollable and end up invading other healthy breast tissues and can travel to the lymph nodes under the arms. There are two types of cancers. Malignant and Benign. Malignant cancers are cancerous. These cells keep dividing uncontrollably and start affecting other cells and tissues in the body. They spread to all other parts of the body, and it is hard to cure this type of cancer. Chemotherapy, radiation therapy and immunotherapy are types of treatments that can be given for these types of tumors. Benign cancer is non-cancerous. Unlike malignant, this tumor does not spread to other parts of the body and hence is much less risky that malignant. In many cases, such tumors don't really require any treatment. Breast cancer is most commonly diagnosed in women of ages above 40. But this disease can affect men and woman of any age. It can also occur when there's a family history of breast cancer. Breast Cancer has always had a high mortality rate and according to statistics, it alone accounts for about 25% of all new cancer diagnoses and 15% of all cancer deaths among women worldwide.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modeling.

Some Risk Factors for Breast Cancer . The following are some of the known risk factors for breast cancer. However, most cases of breast cancer cannot be linked to a specific cause. Talk to your doctor about your specific risk.

- **Age**. The chance of getting breast cancer increases as women age. Nearly 80 percent of breast cancers are found in women over the age of 50.
- **Personal history of breast cancer**. A woman who has had breast cancer in one breast is at an increased risk of developing cancer in her other breast.
- **Family history of breast cancer**. A woman has a higher risk of breast cancer if her mother, sister or daughter had breast cancer, especially at a young age (before 40). Having other relatives with breast cancer may also raise the risk.
- **Genetic factors**. Women with certain genetic mutations, including changes to the BRCA1 and BRCA2 genes, are at higher risk of developing breast cancer during their lifetime. Other gene changes may raise breast cancer risk as well.
- **Childbearing and menstrual history**. The older a woman is when she has her first child, the greater her risk of breast cancer. Also, at higher risk are:
    1. Women who menstruate for the first time at an early age (before 12)
    2. Women who go through menopause late (after age 55)
    3. Women who've never had children

1.1. Purpose :

Breast cancer detection is done with the help of mammograms, which are basically X-rays of the breasts. It's a tool which can help detect and diagnose breast cancer. But detection is not easy due to different kinds of uncertainties in using these mammograms. The result of a mammogram are images that can show any calcifications or deposits of calcium in the breasts. These don't always have to be cancerous. These tests can also find cysts which are fluid-filled sacs that are very normal during some women's menstrual cycles — and any cancerous or noncancerous lumps. Mammograms can cost around ₹15,000 - ₹40,000 or more based on the hospital, location, and area of body to be covered. This is very expensive and not many can afford it. It is always best for an early diagnosis so that the treatment process can also be started early on.

1.2. Objectives :

Breast cancer is a disease which we hear about a lot nowadays. It is one of the most widespread diseases. There are around 2000+ new cases of breast cancer in men each year, and about 2,30,000 new cases in women every year. Diagnosis of this disease is

crucial so that woman can get it treated faster. It is best for a correct and early diagnosis. The main objective of this project is to help doctors analyze the huge datasets of cancer data and find patterns with the patient's data and that cancer data available. With this analysis we can predict whether the patient might have breast cancer or not. Machine learning algorithms will help with this analysis of the datasets. These techniques will be used to predict the outcome. The outcome can be either that the cancer is benign or malignant.

Benign cancer is the cancer which doesn't spread where, as malignant cancer cells spread across the body making it very dangerous. This prediction can help doctors prescribe different medical examinations for the patients based on the cancer type. This helps save a lot of time as well as money for the patient.

1.3. Scope :

This project scheme was developed to reduce some amount of work for the physicians and other doctors so that they don't have to conduct many tests on the patients. It also helps minimize the amount of time and money spent by the patients undergo these tests. As everything is digitalized and based on data analysis, it takes less amount of time to get results. Based on the results, further action can be taken. It also helps researchers in the medical as well as IT sector to understand how different algorithm scan predict different outcomes. This scheme normally requires a huge amount of data about different patient history and the cancer details. This data has been collected by many doctors for a long period of time and will be used to do the analysis. This reduces the computational time required to gather all the data necessary.

Chapter 2

2.Problem definition :

In 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer. There are more lost disability-adjusted life years (DALYs) by women to breast cancer globally than any other type of cancer. Breast cancer occurs in every country of the world in women at any age after puberty but with increasing rates in later life. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. The main issue pertaining to its cure is early recognition.

Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue. This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection.

The goal is to classify whether the breast cancer is benign or malignant. To achieve this, we have used machine learning classification methods to fit a function that can predict the discrete class of new input.

Chapter 3

3. Proposed System :

In the proposed system we plan on using existing data of breast cancer patients which has been collected for a number of years and run different machine learning algorithms on them. These algorithms breast cancer or not and it will also tell us if the cancer is malignant or benign. It is done by taking the patient's data and mapping it with the dataset and checking whether there are any patterns found with the data. If a patient has breast cancer, then instead of taking more tests to check whether the cancer is malignant or benign, ML can be used to predict the case based on the huge amount of data on breast cancer. This proposed system helps the patients as it reduces the amount of money they need to spend just for the diagnosis.

Also, if the tumor is benign, then it is not cancerous, and the patient doesn't need to go through any of the other tests. This saves a lot of time as well.

Advantages:

• Reduces costs for medical tests.

• Does not take huge amount of time.

• Accurate.

• Intelligent way of using available data.

3.1. Features and Functionality :

For building this project we have used Wisconsin Breast cancer data which has 569 rows of which 357 are benign and 212 are malignant. The data is prepossessed and scaled. We have trained with Random Forest Classifier gives the best accuracy of 95.0%. To provide an easy to-use interface to doctors We have developed a website that will take the data and display the output with accuracy and time taken to predict.

In machine learning, feature selection is the process of choosing a subset of relevant attributes from various candidate subsets, and it is a prerequisite for model building. Feature selection plays a vital role in creating an effective predictive model. There are several benefits to applying the feature selection methods: it

(a) It is effective and faster in training the machine learning algorithm,

(b) Reduces the complexity of a model and makes it easier to interpret, (c) improves the accuracy of a model if the right subset is chosen, and

(d) Reduces overfitting.

Chapter 4

4. Project outcome :

The outcome of this model is to correctly check and predict whether a patient has breast cancer or not. If yes, the model should also be able to tell if the patient has malignant or benign type of cancer.

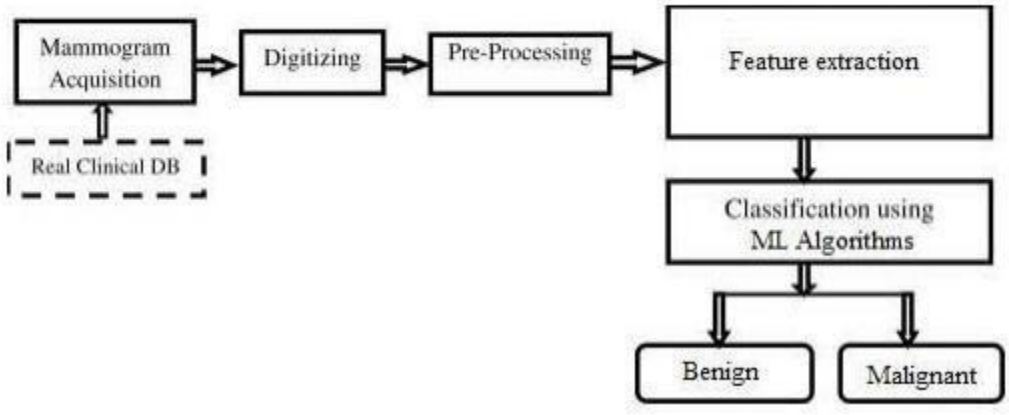

Fig.1: The diagram represents how the data is being processed

Early detection of BC can greatly improve prognosis and survival chances by promoting clinical treatment to patients. We will use the UCI Machine Learning Repository for breast cancer dataset.

n. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector Attribute Information: 1. ID number

2) Diagnosis (M = malignant, B = benign) 3–32) Ten real-valued features are computed for each cell nucleus: 2. radius (mean of distances from center to points on the perimeter)

3. texture (standard deviation of gray-scale values)

4. Perimeter

5. Area

6. Smoothness (local variation in radius lengths)

7. Compactness (perimeter² / area — 1.0)

8. Concavity (severity of concave portions of the contour)

9. Concave points (number of concave portions of the contour)

10. Symmetry

11. Fractal dimension (“coastline approximation” — 1)

The mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

PREPARING THE DATA, We will first go with importing the necessary libraries and import our dataset to colab.research.google.com.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compa
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	
5	843786	M	12.45	15.70	82.57	477.1	0.12780	
6	844359	M	18.25	19.98	119.60	1040.0	0.09463	

perimeter_se	area_se	smoothness_se	compactness_se	concavity_se	concave points_se	symmetry_se	fractal_dimension_se	radius_worst	texture_worst	perimeter_worst	area_worst	sm
8.589	153.40	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193	25.38	17.33	184.60	2019.0	
3.398	74.08	0.005225	0.01308	0.01860	0.01340	0.01389	0.003532	24.99	23.41	158.80	1956.0	
4.585	94.03	0.006150	0.04006	0.03832	0.02058	0.02250	0.004571	23.57	25.53	152.50	1709.0	
3.445	27.23	0.009110	0.07458	0.05661	0.01867	0.05963	0.009208	14.91	26.50	98.87	567.7	
5.438	94.44	0.011490	0.02381	0.05688	0.01885	0.01756	0.005115	22.54	16.67	152.20	1575.0	
2.217	27.19	0.007510	0.03345	0.03672	0.01137	0.02165	0.005082	15.47	23.75	103.40	741.6	
3.960	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002179	22.88	27.66	153.20	1606.0	

concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst	Unnamed: 32
0.7119	0.2654	0.4601	0.11890	NaN
0.2416	0.1860	0.2750	0.08902	NaN
0.4504	0.2430	0.3613	0.08758	NaN
0.6869	0.2575	0.6638	0.17300	NaN
0.4000	0.1625	0.2364	0.07678	NaN
0.5355	0.1741	0.3985	0.12440	NaN
0.3784	0.1932	0.3063	0.08368	NaN

Fig.2: We can examine the data set using the pandas' head() method. df.head(7) {first 7 rows of the data}

We can find the dimensions of the data set using the panda dataset 'shape' attribute.
df.shape

(569, 33) We observe that the data set contain 569 rows and 33 columns. 'Diagnosis' is the column which we are going to predict , which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labeled as B (benign) and 212 as M (malignant). Each row represents a patient and 33 features on the 569 patients. The last column Unnamed 32 has NaN values so we need to remove that column with empty values. So we count the number of empty colums and drop the colums with empty values.

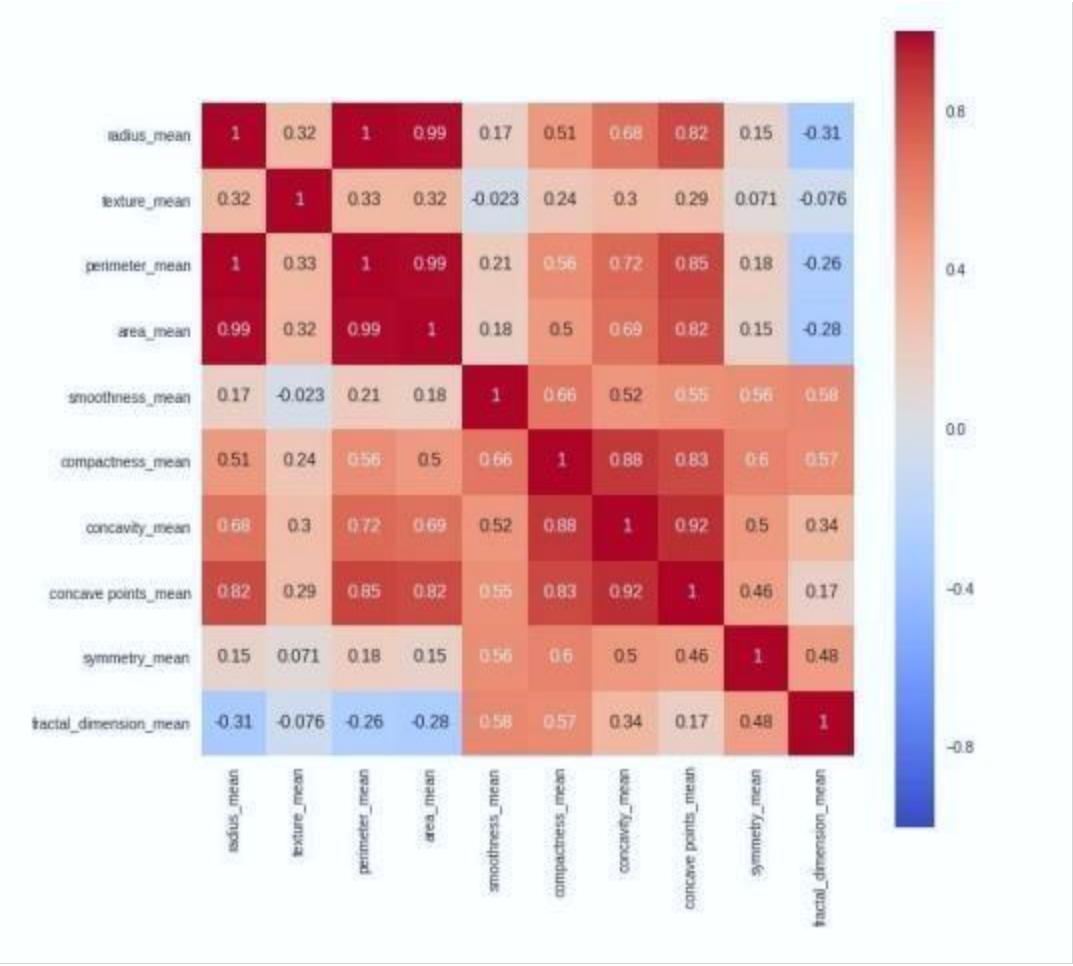

Fig.3 Heat Map

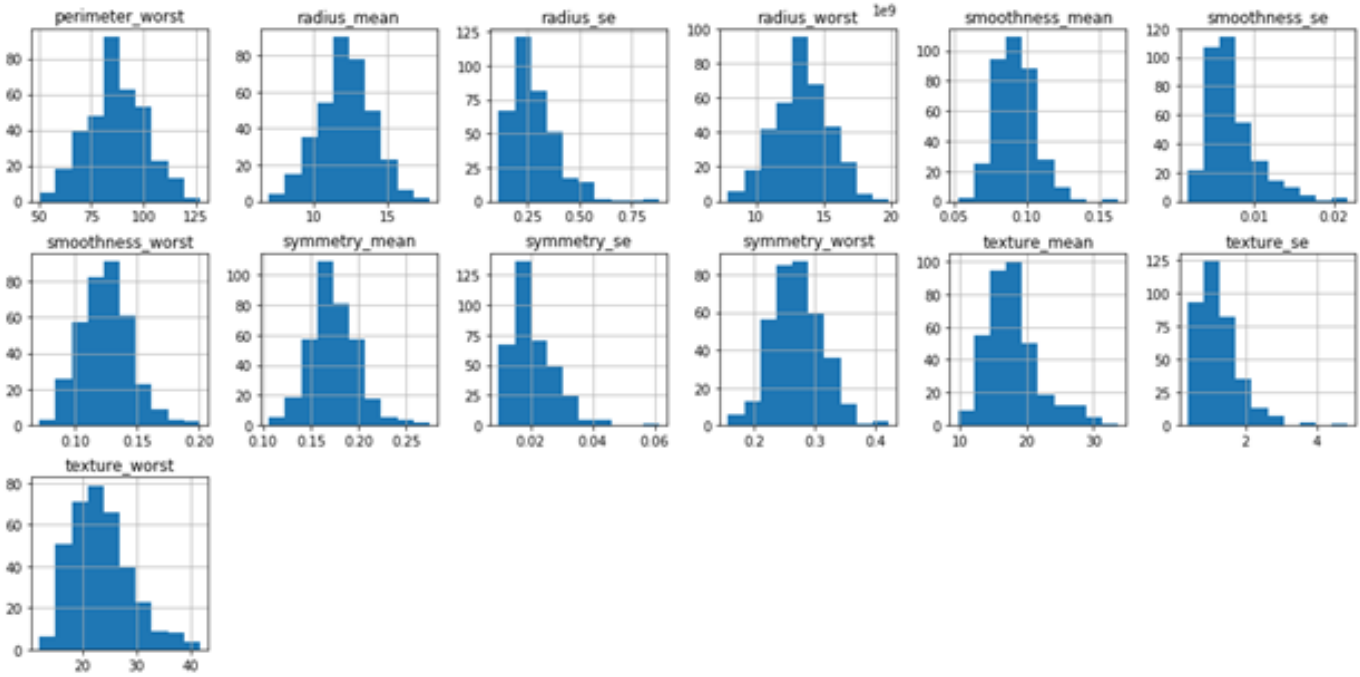

Fig.4: Visualization of dataset

We can also see how the malignant or benign tumors cells can have (or not) different values for the features plotting the distribution of each type of diagnosis for each of the mean features.

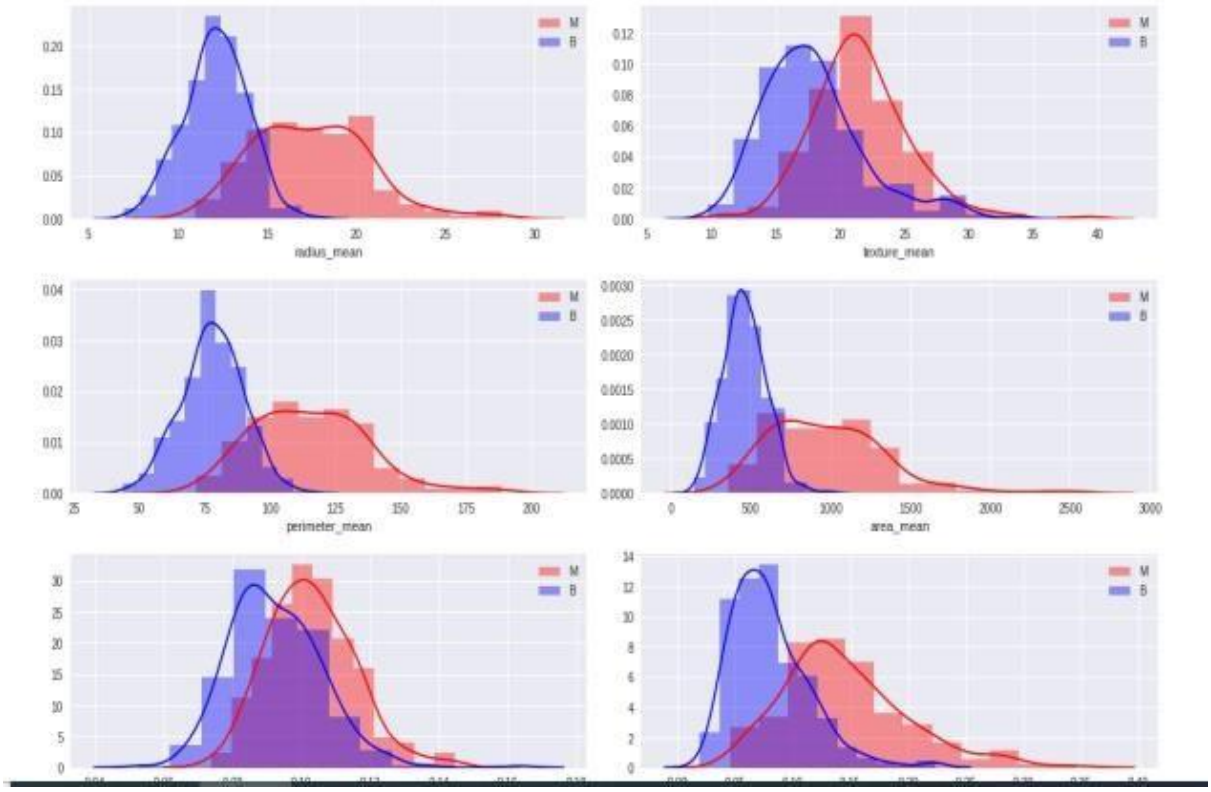

Fig.5 Statistical representation of dataset

```
[[86  4]
 [ 4 49]]
Model[0] Testing Accuracy = "0.9440559440559441!"

[[89  1]
 [ 5 48]]
Model[1] Testing Accuracy = "0.958041958041958!"

[[87  3]
 [ 2 51]]
Model[2] Testing Accuracy = "0.965034965034965!"

[[85  5]
 [ 6 47]]
Model[3] Testing Accuracy = "0.9230769230769231!"

[[84  6]
 [ 1 52]]
Model[4] Testing Accuracy = "0.951048951048951!"
```

Here the matrices are of form [TP FP] [FN TP] where TP is true positive: A true positive is an outcome where the model correctly predicts the positive class TN is true negative: A true negative is an outcome where the model correctly predicts the negative class. FN is false negative: A false negative is an outcome where the model incorrectly predicts the negative class. FP is false positive: A false positive is an outcome where the model incorrectly predicts the positive class. Based on the test data we can see that Model 5 ie Random forest classifier has 96.5% accuracy on the test data so we can use it to predict the actual outcome whether a patient has cancer or not.

Outcome through project:

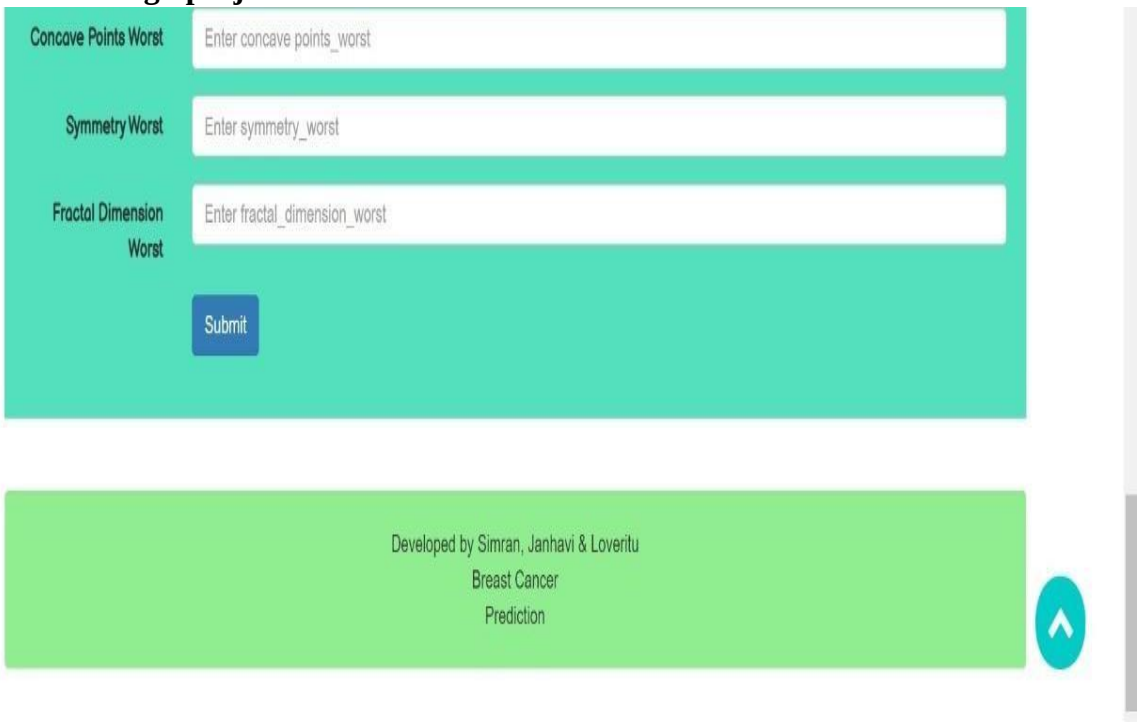

Fig.6: Different Value Insertion for detection

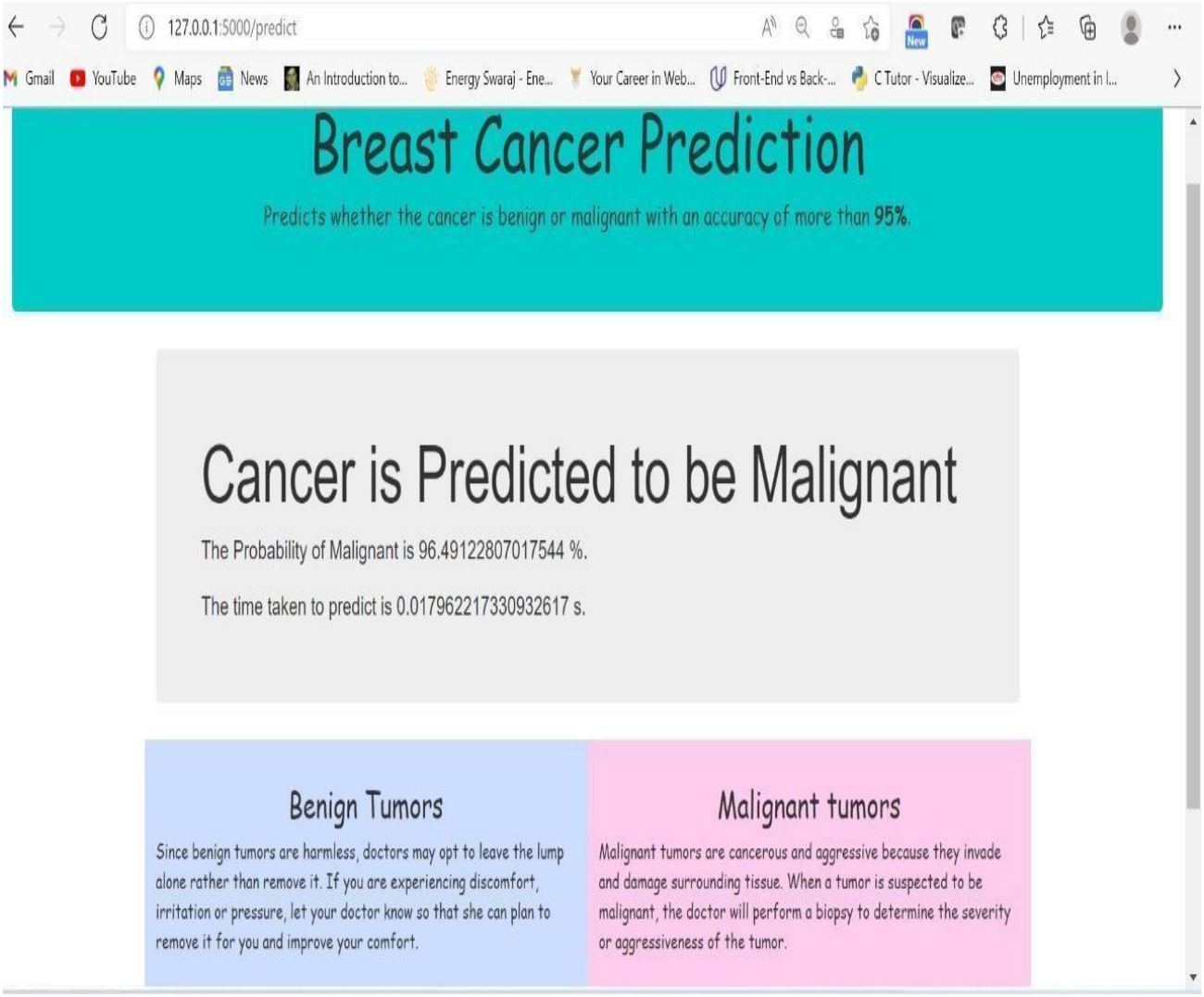

Fig.7 Outcome: The program computes the end result and provides it to the user as shown.

Chapter 5

5. Software Requirements:

Python: language.

NumPy: library for numerical calculations

Pandas: library for data manipulation and analysis

SkLearn: library which features various classification, regression and clustering algorithms Flask: microframework for building web applications using Python.

Chapter 6

6. Project Design:

DESIGN GOALS Under our model, the goal of our project is to create a design to achieve the following:

ACCURACY Only accurate out comes can help make this model a good one. It can be reliable only when all the outcomes are correct and can be trusted. As this data is required for healthcare purposes, it is important that no errors occur.

EFFICIENCY

The model should be efficient as there is no requirement of manual data entry work or any work by doctor. It takes less time to predict outcomes after all the ML algorithms have been used on the data.

FLOW DIAGRAM

z

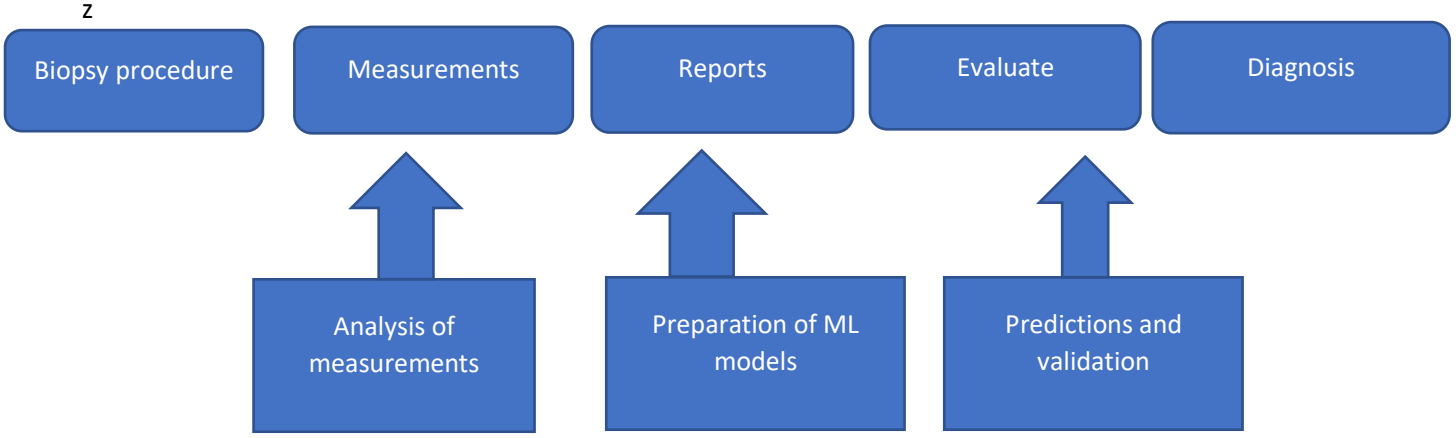

Fig.8: Data Flow Diagram

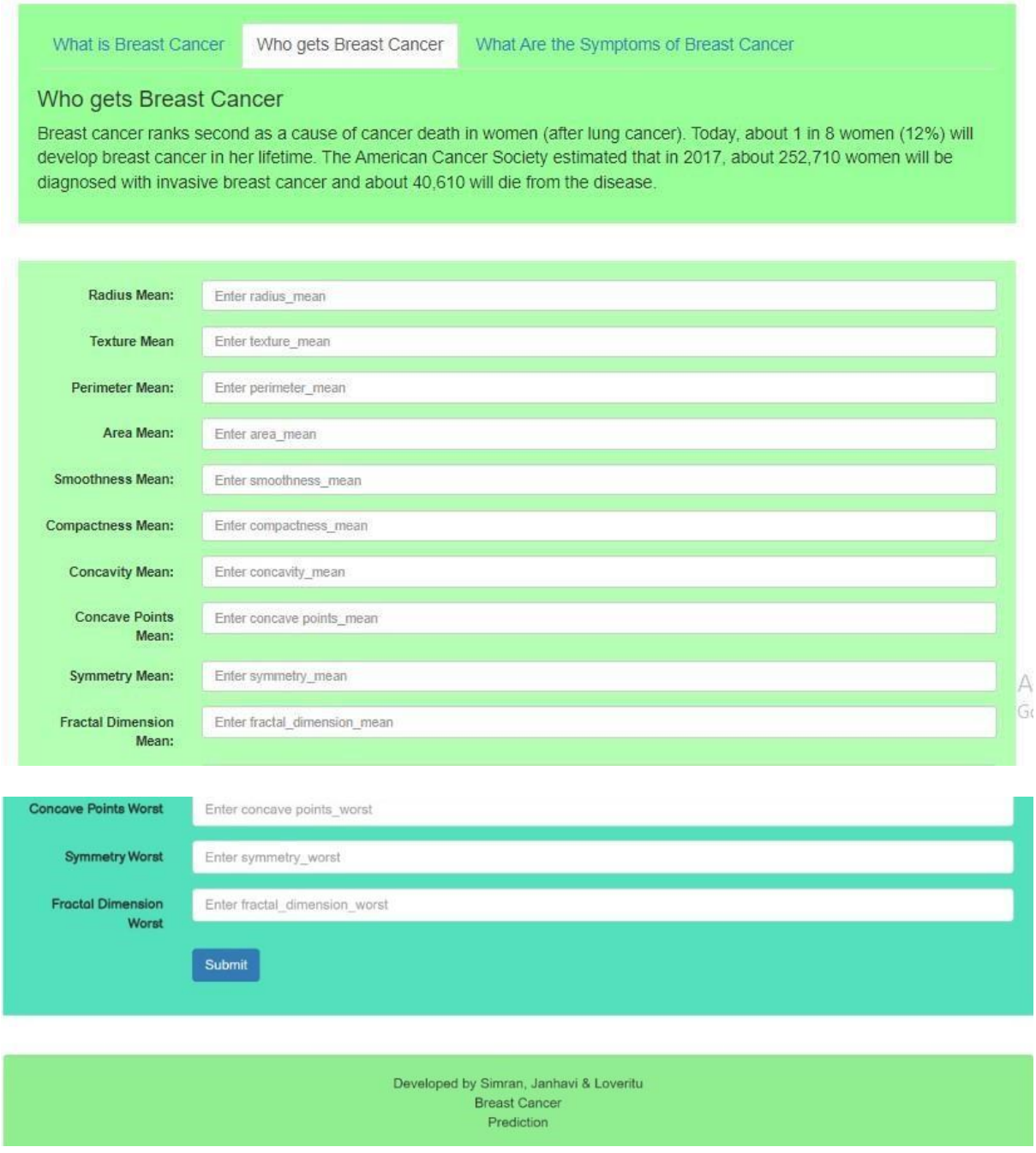

Fig.11:Here in this part of the program we take input from the user for prediction

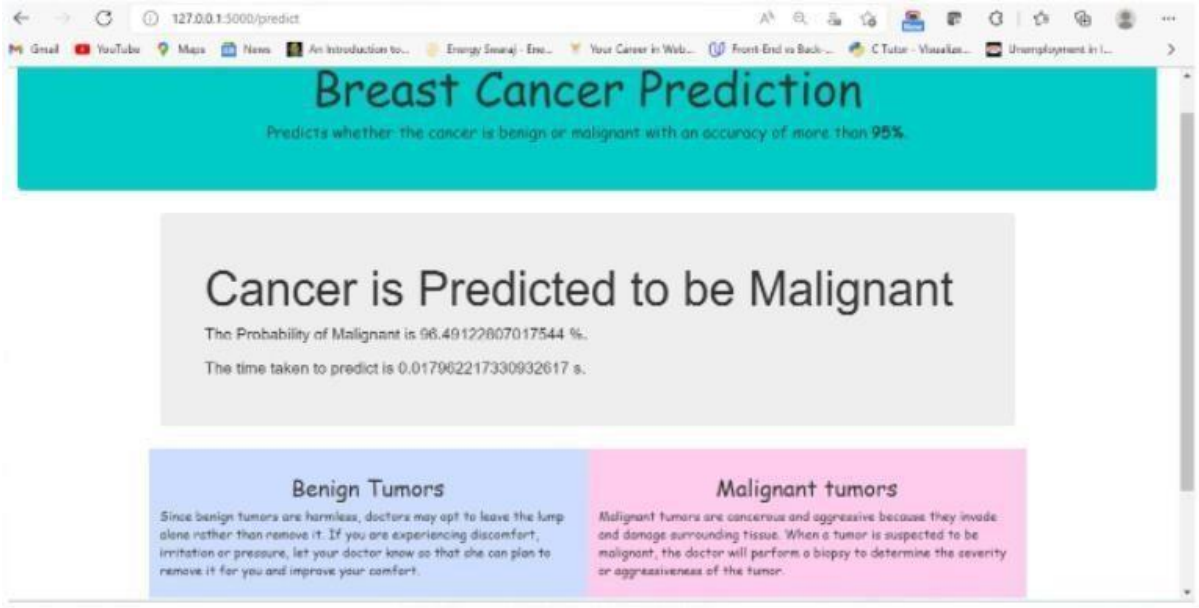

Fig.12: Outcome: The program computes the end result and provides it to the user as shown.

Chapter 7

7. Project Scheduling:

1	**Janhavi Kulkarni**	2nd week of feb	Front end using Python for the website and connections . Report of the project.
2	**Simran Choudhary**	2nd week of feb	Backend database using flask and connections . PPT of the project.
3	**Loveritu Itnare**	By the end of march month	Finding dataset and parts of ml algorithms

Chapter 8

Conclusion:

In this project in python, we learned to build a breast cancer tumor predictor on the wisconsin dataset and created graphs and results for the same. It has been observed that a good dataset provides better accuracy. Selection of appropriate algorithms with good home dataset will lead to the development of prediction systems. These systems can assist in proper treatment methods for a patient diagnosed with breast cancer. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

REFERENCES:

1. https://www.who.int/cancer/prevention/dia gnosis-screening/breast-cancer/en/

2. https://towardsdatascience.com/buildinga-simple-machine-learning-model-onbreastcancerdata-eca4b3b99fa3

3. Original data Set: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

4. Confusion Matrix: https://tatwan.github.io/How-To-Plot-AConfusion-Matrix- In-Python/

5. https://seaborn.pydata.org/tutorial/axis_gri ds.html

6. https://seaborn.pydata.org/generated/seabo rn.pairplot.html

7. https://seaborn.pydata.org/generated/seabo rn.heatmap.html