TUMOR CLASSIFICATION USING ML : AN APPROACH TO DETECT BREAST CANCER

A PROJECT REPORT

Submitted by :-

Shlok N Srivastava {20BAI10004}
Yash Rai {20BAI10068}
Shrimohan Tripathi {20BAI10088}
Uddhav Davey {20BAI10099}

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

with

SPECIALIZATION IN AI AND ML

VIT®
BHOPAL
www.vitbhopal.ac.in

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY

KOTHRIKALAN, SEHORE
MADHYA PRADESH - 466114

DEC 2021

VIT BHOPAL UNIVERSITY, KOTHRIKALAN, SEHORE MADHYA PRADESH – 466114

BONAFIDE CERTIFICATE

Certified that this project report titled **"Tumor classification using ML: An approach to detect Breast Cancer.'' is** the bonafide work of "**SHLOK N SRIVASTAVA(20BAI10004); YASH KUMAR RAI(20BAI10068); SHRIMOHAN TRIPATHI(20BAI10088); UDDHAV DAVEY(20BAI10099)"** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

PROGRAM CHAIR
Dr. S. Sountharrajan,
Program Chair - B.Tech CSE spl. in AI & ML
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

PROJECT GUIDE
Dr. Anil Kumar Yadav,
Teaching Fellow
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

ACKNOWLEDGEMENT

First and foremost we would like to thank the Lord Almighty for his presence and immense blessings throughout the project work.

We wish to express our heartfelt gratitude to Dr S. Sountharrajan, Program Chair-B.Tech specialization in AI and ML, School of Computer Science and Engineering for much of his valuable support and encouragement in carrying out this work.

We would like to thank our internal guide Dr. Anil Kumar Yadav, for continually guiding and actively participating in our project, giving valuable suggestions to complete the project work.

We would like to thank all the technical and teaching staff of the School of Computer Science and Engineering, who extended directly or indirectly all support.

Last, but not least, we am deeply indebted to our parents who have been the greatest support while we worked day and night for the project to make it a success.

LIST OF IMAGES AND TABLES

Serial No.	TITLE	PAGE NO.
Fig. 1	Percentage of new cancer cases in each continent	8
Fig. 2	Estimated number of deaths in India, females, all ages	9
Fig. 3	Estimated number of new cases in India, females, all ages	9
Fig. 4	Estimated number of new cases in India, females, 70+	10
Fig. 5	Estimated number of new cases in India, females, 25-49	10
Fig. 6	Flow-Chart	14
Fig. 7	Software Architecture Diagram	15
Table 1	Statistical Measure Table	17
Table 2	Statistical Measure Table for separate labels	18
Table 3	Accuracy Table	19
Fig. 8	Working Layout	19
Fig. 9	Output 1	20
Fig. 10	Output 2	20

ABSTRACT

Globally, breast cancer is the most common cancer among women, and the most likely cause of female cancer deaths. High-income countries (HICs) have made the most progress in improving breast cancer outcomes. Between 1990 and 2014, breast cancer death rates dropped by 34% in the US attributable to the combination of improved earlier detection and effective adjuvant therapies. By contrast, breast cancer is an increasingly urgent problem in low- and middle-income countries (LMICs), where historically low incidence rates have been rising by up to 5% per year.

In this study, the diagnosis of breast cancer from mammograms is complemented by using logistic regression. The radiologists can use the results to make a proper judgment as to the presence of breast cancer. The results using logistic regression cross tabulation was to obtain the significant values between the breast cancer factors. The classification table from 570 samples shows the occurrence from prediction and observation samples, producing a percentage of correct classification for mammogram results of 96.49%. The analysis for mammograms screening using parameter estimation is to identify all the factors that were available in the survey. The presence of mass, architectural distortion, skin thickening, and calcification had high odds of getting breast cancer.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	List of Figures and Tables	4
	Abstract	5
1	**CHAPTER-1:** **PROJECT DESCRIPTION AND OUTLINE** 1.1 Introduction 1.2 Motivation for the work 1.3 Problem Statement 1.4 Objective of the work	8
2	**CHAPTER-2:** **RELATED WORK INVESTIGATION** 2.1 Introduction 2.2 Existing Works 2.3 Cons of Existing Work	12
3	**CHAPTER-3:** **REQUIREMENT ARTIFACTS** 3.1 Introduction 3.2 Hardware and Software requirements	13
4	**CHAPTER-4:** **DESIGN METHODOLOGY AND ITS NOVELTY** 4.1 Methodology and goal 4.2 Software Architectural designs	14

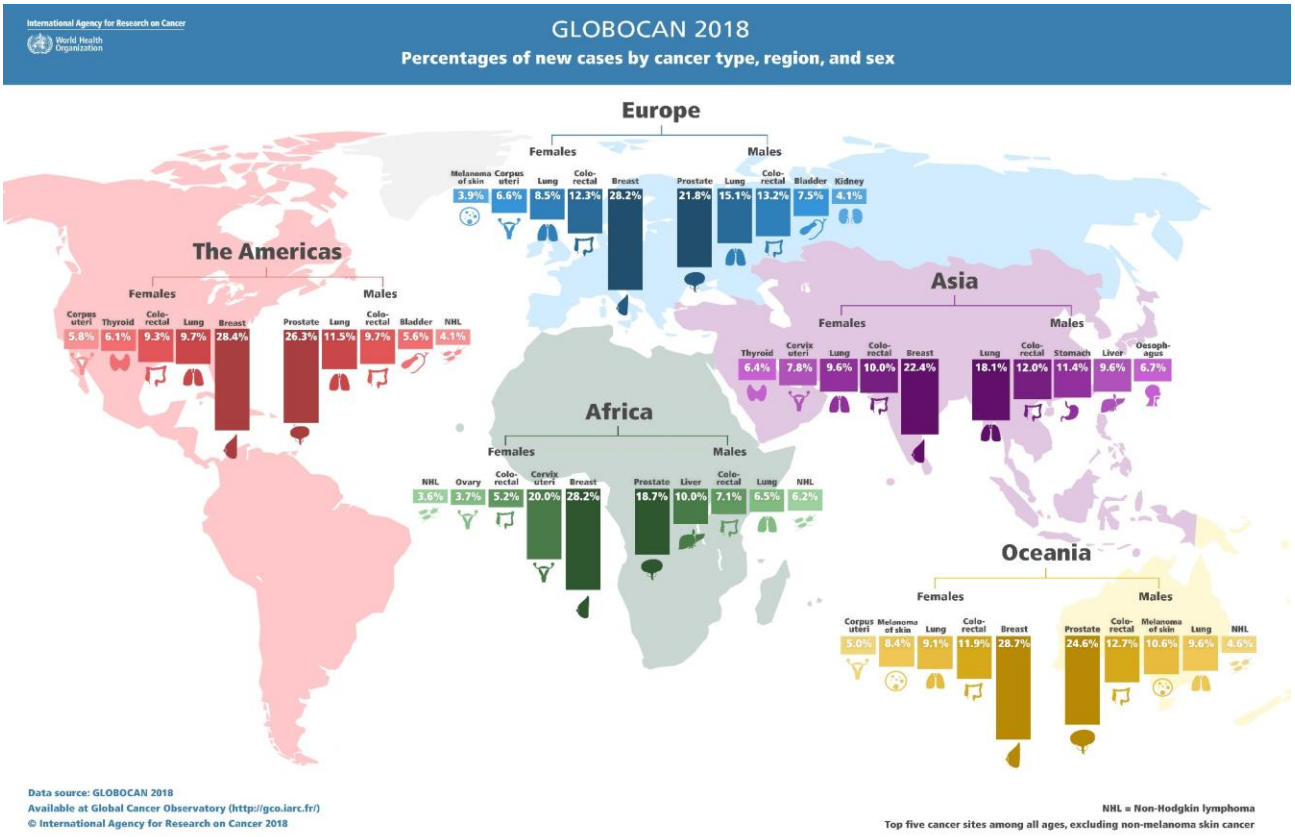
5	**CHAPTER-5:** **TECHNICAL IMPLEMENTATION & ANALYSIS** 5.1 Outline 5.2 Technical coding 5.3 Working Layout 5.4 Tables	16
6	**CHAPTER-6:** **PROJECT OUTCOME AND APPLICABILITY** 6.1 Outline 6.2 Project outcomes 6.3 Project applicability on Real-world applications 6.4 Limitations of the System	20
7	**CHAPTER-7:** **CONCLUSIONS AND RECOMMENDATION** 7.1 Outline 7.3 Future Enhancement	22

Project Description and Outline

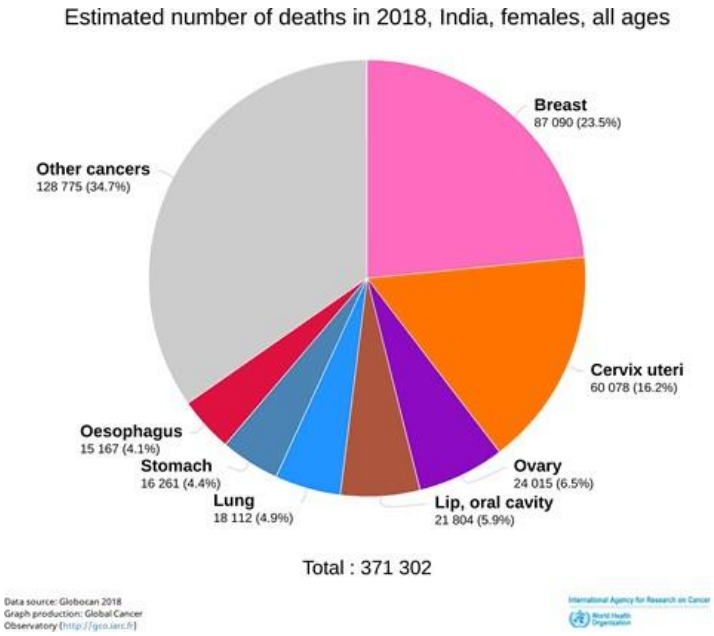
Introduction

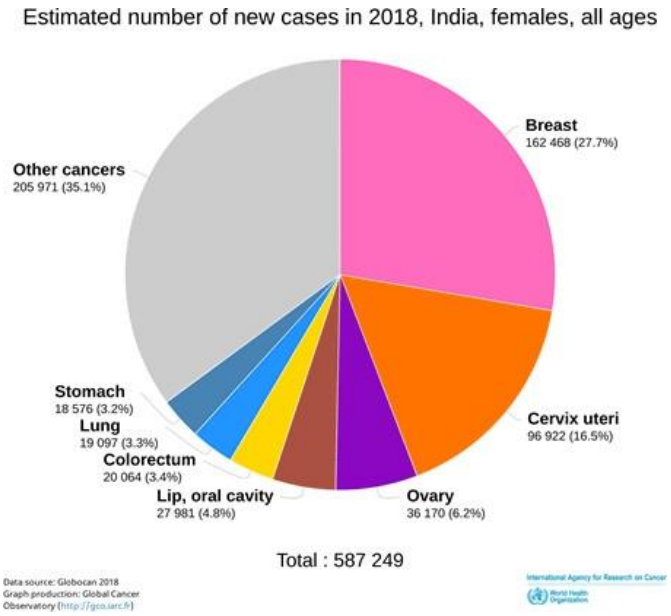
Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. The early diagnosis of BC can improve the prognosis and 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.

Motivation for Work

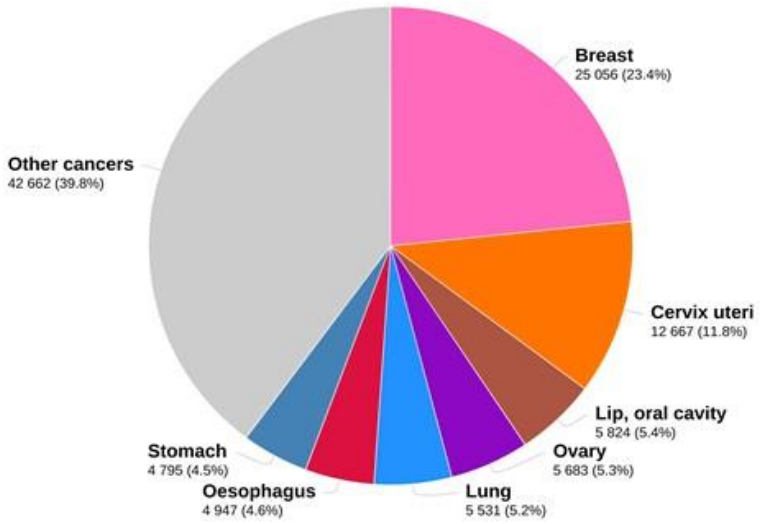


As you can see from the attached image, womens from every continent are suffering from breast cancer. Most cancer cases among womens is because of breast cancer. It ranges from 22.4% of total cancer cases to as high as 28.4% of total cancer cases. It also accounts for the most number of deaths because of cancer. If we are able to predict it at early stages, then it can be treated easily and save a woman from death. This motivated us to take this as a topic for betterment of the society.





Estimated number of new cases in 2018, India, females, ages 70+

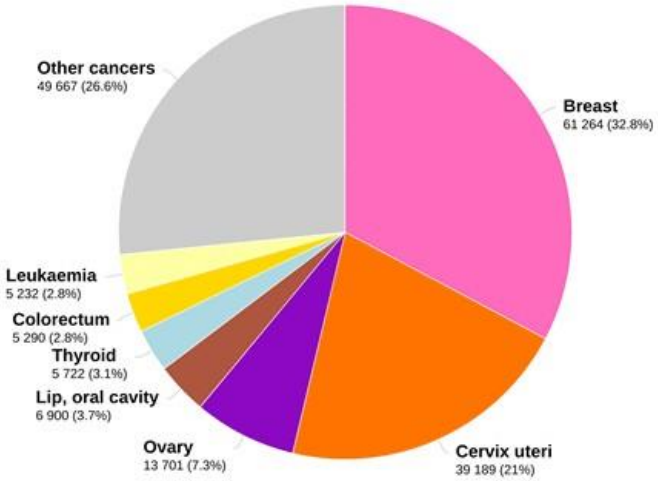


Total : 107 165

Data source: Globocan 2018
Graph production: Global Cancer Observatory (http://gco.iarc.fr)

International Agency for Research on Cancer
World Health Organization

Estimated number of new cases in 2018, India, females, ages 25-49



Total : 186 965

Data source: Globocan 2018
Graph production: Global Cancer Observatory (http://gco.iarc.fr)

International Agency for Research on Cancer
World Health Organization

Problem Statement

Breast Cancer is one of the leading cancers developed in many countries including India. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased dramatically in the last few decades. 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 hyperparameter selection. The goal is to classify whether 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.

Objective of the Work

The objective of our project is to create a program with the help of python and its libraries that can be used by medical professionals, after doing needle biopsy, for detecting breast cancer even at early stages. This will help them by allowing its treatment at early stages which in turn can save lives of millions of people.

Related Work Investigation

Introduction

Here, we will tell you some of the different types of already existing projects which are similar{not-exact} to our project. Then we will also tell you how our project is different from theirs.

Existing Work

These are some projects which we got to know through journals and the internet. Here, I am attaching links to some of those available on the internet:

- https://github.com/mrdvince/breast_cancer_detection
- https://github.com/Aftaab99/Cancer-diagnosis-and-early-detection
- https://github.com/gscdit/Breast-Cancer-Detection

Cons/Limitation of Existing Work

The existing work which we came through consists of machine learning technologies like NN which are computationally expensive. While in our project we are using a logistic regression model which makes our model fast working and efficient. Its low computational cost also makes it easy to install in any system and thus can be easily installed in medical machines.

Requirement Artifacts

Introduction

Here, we are going to discuss the minimum hardware and software requirements for our project.

Hardware and Software Requirements

Hardware Requirements:
Intel Core i3 Processor or more
2GB of RAM or more
Windows 7 or Mac OS X 10.11 or later
500 MB or more disk space
Input and Output machines like keyboard, display

Software Requirements:
Python IDE
Sklearn.linear_model Library
Sklearn.model_selection Library
Sklearn Library
Pandas Library
Numpy Library

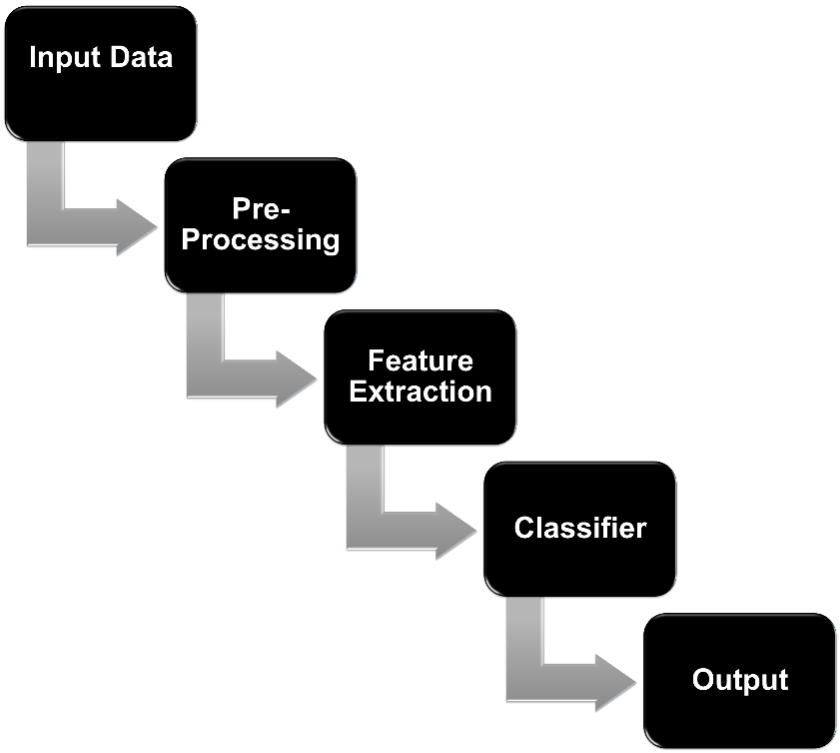
Design Methodology and its Novelty

Methodology

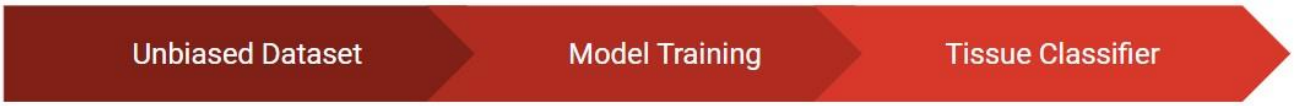
We initially planned to use a decision tree algorithm for our implementation using python. It is so because it is an efficient algorithm which predicts data based on some decisions taken (the comparisons made). A root node is built and then split happens for different nodes and different decisions are being taken and considered before the outcome is predicted. The root node is dependent upon the information gain and entropy in the dataset which is calculated.

Another algorithm that could be used here is a random forest classifier in which it fits a tree for each one of those new data frames and predicts by averaging all the trees in the forest. The pros of this is that this algorithm works well with the large dataset and also works great with high dimensional data.

But, because of high time complexity and computational costs of above stated methods we moved towards the use of logistic regression models. The advantages of using this model is that it is easy to understand and works with both classification and regression problems. It is also less time consuming and thus computational cost is less. It is the best algorithm when there is linear relation among variables.



Software Architectural Design



- Here, we input many datasets as raw data for training of our model.
- We have to take different types of data, some which are benign and some which are malignant, so that our data is not biased.

- Here, we train the model by giving dataset so that it can identify the provided datasets accurately.
- Now, the model will learn from the given datasets.

- After learning, we provide the model with different datasets which were not feeded earlier.
- Now, the model will classify the datasets as benign or malignant.
- If the model classifies accurately, then we say that it is efficient.

Technical Implementation and Analysis

Outline

Here in this section, we are going to tell you about the technical part of our project, i.e. the coding part, and how we can interpret more data from the output.

Also in this section, we will show you the working architecture of the program which will help you better understand our project.

Technical Coding

```
"""
Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1OIT5b6c5N06cf6oeDDXNkzAbKY7UW49_
"""
# Importing Libraries
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn import metrics

# Reading Data-File
mydf = pd.read_csv("data.csv")

# statistical measures about the data
mydf.describe()

# Defining Dependent and Independent Variables
X = mydf.iloc[:, 2:32].values  # Independent Variable
Y = mydf.iloc[:, 1]  # Dependent Variable

# Splitting in training and testing data
X_train, X_test, Y_train, Y_test = train_test_split(
    X, Y, test_size=0.1, random_state=1)

# Training the data
model = LogisticRegression(max_iter=3000)
model.fit(X_train, Y_train)
```

```
print("Model Created Successfully")

# Predicting Values
Y_pred_test = model.predict(X_test)
Y_pred_train = model.predict(X_train)

# Showing Accuracy
print("Accuracy for Testing data=",
      metrics.accuracy_score(Y_test, Y_pred_test)*100, '%')
print("Accuracy for Training data=", metrics.accuracy_score(
    Y_train, Y_pred_train)*100, '%')

# Predicting for custom Input
X_new = (7.76, 24.54, 47.92, 181, 0.05263, 0.04362, 0, 0, 0.1587, 0.05884,
0.3857, 1.428, 2.548, 19.15, 0.007189,
         0.00466, 0, 0, 0.02676, 0.002783, 9.456, 30.37, 59.16, 268.6,
0.08996, 0.06444, 0, 0, 0.2871, 0.07039)
input_data_as_numpy_array = np.asarray(X_new)
input_data_reshaped = input_data_as_numpy_array.reshape(1, -1)
Y_new = model.predict(input_data_reshaped)
print(Y_new)

# 'B' --> Benign --> Non-Cancerous
# 'M' --> Malignant --> Cancerous
```

Statistical Measure Table

Feature	Mean Value	Minimum Value	Maximum Value
Mean Radius	14.12729173989456	6.981	28.11
Mean Texture	19.28964850615117	9.71	39.28
Mean Perimeter	91.96903339191566	43.79	188.5
Mean Area	654.8891036906857	143.5	2501.0

Mean Smoothness	0.096360281195079	0.05263	0.1634
Mean Compactness	0.104340984182776	0.01938	0.3454
Mean Concavity	0.088799315817223	0.0	0.4268
Mean Concave Points	0.048919145869947	0.0	0.2012
Mean Symmetry	0.181161862917399	0.106	0.304
Mean Fractal Dimension	0.062797609841827	0.04996	0.09744

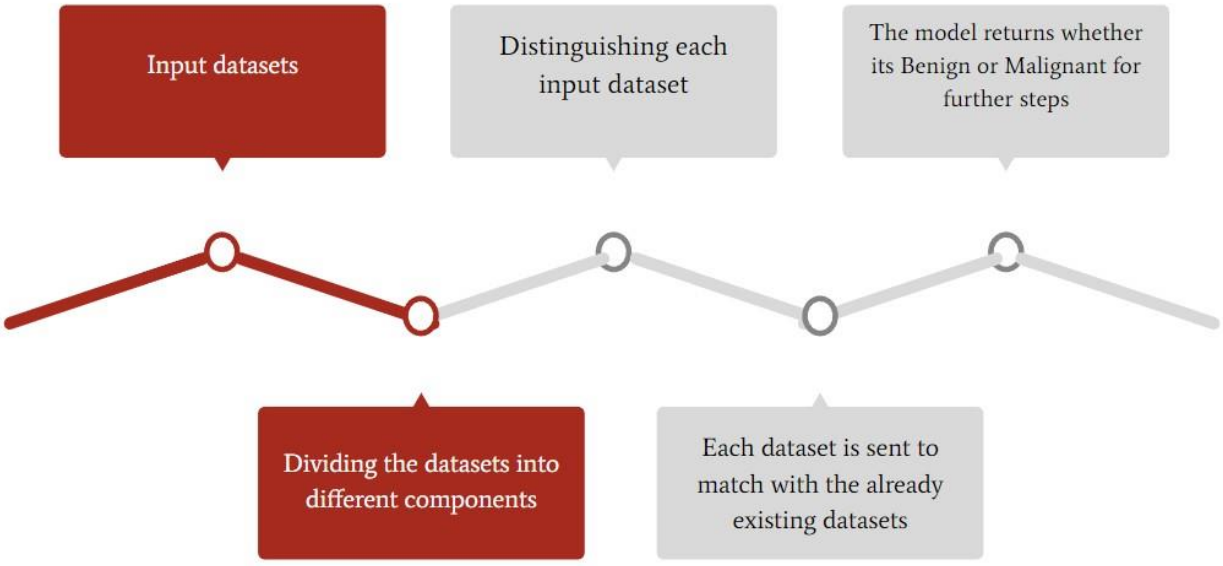
Statistical Measure Table for Different Labels

Feature	Label = 0 {Malignant}	Label = 1 {Benign}
Mean Radius	17.462830	12.146524
Mean Texture	21.604906	17.914762
Mean Perimeter	115.365377	78.075406
Mean Area	978.376415	462.790196
Mean Smoothness	0.102898	0.092478
Mean Compactness	0.145188	0.080085
Mean Concavity	0.160775	0.046058
Mean Concave Points	0.087990	0.025717
Mean Symmetry	0.192909	0.174186
Mean Fractal Dimension	0.062680	0.062867

Accuracy Table

Data	Accuracy
Training	96.491228070175 %
Testing	95.5078125 %

Working Layouts



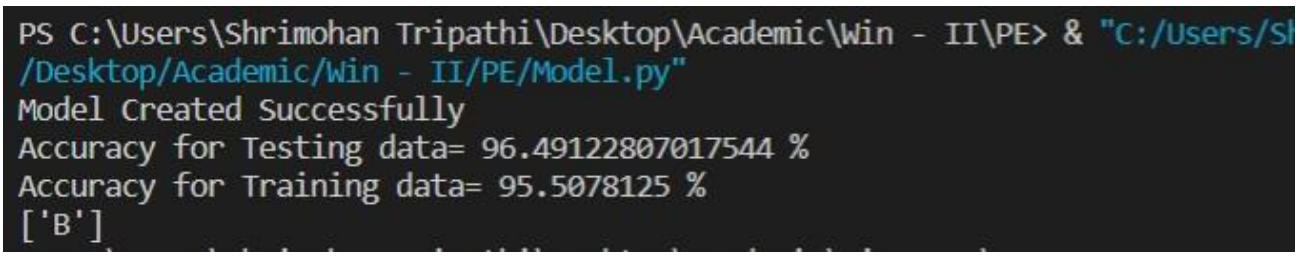
Project Outcome and Applicability

Outline

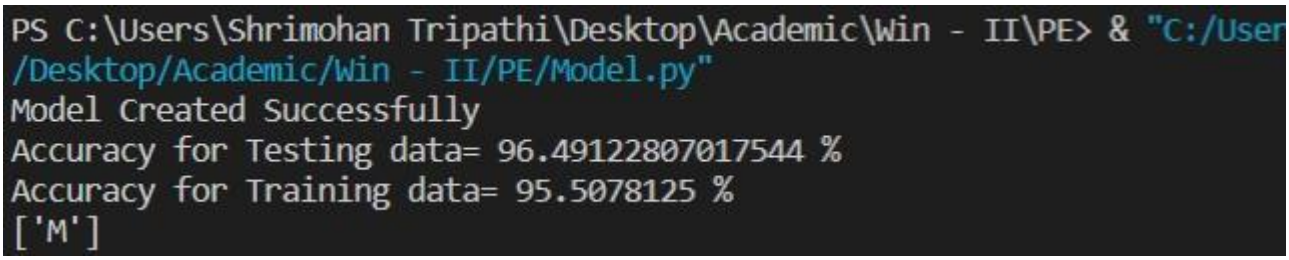
Here, we are discussing the results we got from our prediction model after implementation of the code and also the real life use of our project which can help to detect and control the spread of the malignant/cancerous cells.

Project Outcomes

Output for case 1:



Output for case 2:



Project Applicability on Real-World Application

The biggest and most important application of our project is its use in hospitals where the patient and doctors can get a clear picture regarding breast cancer detection after they have successfully conducted the required medical tests that are required by the model.

Limitation(s) of the System

We tested our model and we found that our model cannot successfully predict the stage of the cancer which the female is suffering from.

Our model also cannot predict the attack of breast cancer which can occur again to the individual after some years In her life.

Conclusion

Outline

Logistic regression analysis was performed using the variables from the mammogram results which are mass, architectural distortion, skin thickening, and calcification. A patient with mass detected on mammogram screening has a probability of five times higher in getting breast cancer. Patients with architectural distortion or skin thickening have a high probability of being afflicted with breast cancer. Also for patients with calcification detected, the probability of getting breast cancer is 18 times higher. Thus, a patient having any of the symptoms or a combination of these symptoms has greater probability of getting breast cancer. The study can assist radiologists to correctly diagnose breast cancer from using mammograms and referring to the patients' history.

Future Enhancements

There is a lot of scope for the enhancement and betterment of this project. In the future, we can add more features in this project like "Stage Detection" which help us to determine the stage of breast cancer. These are just some of the enhancements which someone can do in the future. Afterall the scope of enhancement in anything or any work is endless.

References

- Google
  - Research Paper 1
  - Research Paper 2
- GitHub
- Youtube
- StackOverFlow