DAYANANDA SAGAR UNIVERSITY

KUDLU GATE, BANGALORE – 560068

Bachelor of Technology
in
COMPUTER SCIENCE AND ENGINEERING

19CS4702- Major Project Phase II Report

(BREAST CANCER PREDICTION MODEL USING ML)

By
Rahul Kumar - ENG19CS0243
Ruchith B M - ENG19CS0264
Sagar M A - ENG19CS0272
Sai Lakshmi Sridhar - ENG19CS0277

Batch no - **84**

Under the supervision of

Dr.Shaila S G

Chairman of Data Science

Department of CSE, DSU

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,
SCHOOL OF ENGINEERING
DAYANANDA SAGAR UNIVERSITY,

(2022-2023)

DAYANANDA SAGAR UNIVERSITY

School of Engineering
Department of Computer Science & Engineering
Kudlu Gate, Bangalore – 560068
Karnataka, India

CERTIFICATE

This is to certify that the Major Project Stage-1 work titled **"BREAST CANCER PREDICTION MODEL USING ML"** is carried out by **Rahul Kumar (ENG19CS0243), Ruchith B M (ENG19CS0264), SAGAR M A (ENG19CS0272), Sai Lakshmi Sridhar (ENG19CS0277)** a bonafide students of Bachelor of Technology in Computer Science and Engineering at the School of Engineering, Dayananda Sagar University, Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering, during the year **2022-2023**.

DR Shaila S G	**Dr. Girisha G S**	**Dr. Udaya Kumar Reddy K R**
Chairman Data Science Dept. of CS&E, School of Engineering Dayananda Sagar University	Chairman CSE School of Engineering Dayananda Sagar University	Dean School of Engineering Dayananda Sagar University
Date:	Date:	Date:

Name of the Examiner **Signature of Examiner**

1.

2.

DECLARATION

We **Rahul Kumar (ENG19CS0243), Ruchith B M (ENG19CS0264), SAGAR M A (ENG19CS0272), Sai Lakshmi Sridhar (ENG19CS0277)** are student's of seventh semester B.Tech in **Computer Science and Engineering**, at School of Engineering, **Dayananda Sagar University**, hereby declare that the Major Project Stage-1 titled **"BREAST CANCER PREDICTION MODEL USING ML"** has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Technology in Computer Science and Engineering** during the academic year **2022-2023**.

Student **Signature**

Name1: Rahul Kumar
USN : ENG19CS0243

Name2: Ruchith B M
USN : ENG19CS0264

Name3: SAGAR M A
USN : ENG19CS0272

Name4: Sai Lakshmi Sridhar
USN : ENG19CS0277

Place : Bangalore
Date :

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to School of Engineering & Technology, Dayananda Sagar University for providing us with a great opportunity to pursue our Bachelor's degree in this institution.

We would like to thank ***Dr. Udaya Kumar Reddy K R, Dean, School of Engineering & Technology, Dayananda Sagar University*** *for his constant encouragement and expert advice.*

It is a matter of immense pleasure to express our sincere thanks to ***Dr. Girisha G S, Department Chairman***, ***Computer Science and Engineering***, ***Dayananda Sagar University,*** *for providing right academic guidance that made our task possible.*

We would like to thank our guide ***Dr.Shaila S G Chairman Data Science***, ***Dept. of Computer Science and Engineering***, ***Dayananda Sagar University***, *for sparing his/her valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.*

We would like to thank our ***Project Coordinator Dr. Meenakshi Malhotra, Dr. Pramod Naik, Associate Professor, Department of Computer Science and Engineering*** *and all the staff members of Computer Science and Engineering for their support.*

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in the Project work.

Signature of Students

USN : ENG19CS0243, ENG19CS0264, ENG19CS0272,ENG19CS0277

Name: Rahul Kumar, Ruchith B M, Sagar M A, Sai Lakshmi Sridhar

ABSTRACT

Women are seriously threatened by breast cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, the requirement of time is to develop the technique which gives minimum error to increase accuracy. Four algorithm Logistic Regression, Random Forest and KNN which predict the breast cancer outcome have been compared in the paper using different datasets. All experiments are executed within a simulation environment and conducted in Google Collab platform.

Sequential forward selection-based feature selection and deep learning Transfer learning Vgg16 model for feature extraction to distinguish between the benign and malignant tumors of breast. The best overall accuracy for breast cancer diagnosis is achieved equal to 99.10% and 99.70% respectively for random forest and logistic regression machines classifier models against two widely used breast cancer benchmark datasets.

This project is carried out to predict the accuracy. The future research can be carried out to predict the other different parameters and breast cancer research can be categorized on basis of other parameters.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

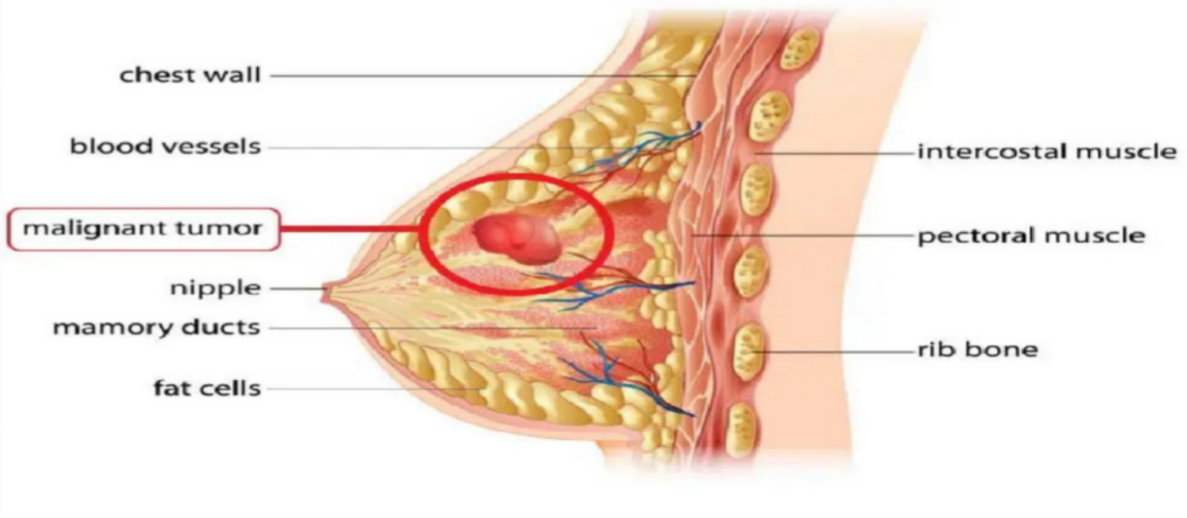
INTRODUCTION

CHAPTER 1 INTRODUCTION

Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumor that can often be seen on an x-ray or felt as a lump. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body. The cause of Breast Cancer includes changes and mutations in DNA. The side effects of Breast Cancer are – Fatigue, Headaches, Pain and numbness (peripheral neuropathy), Bone loss and osteoporosis. There are many algorithms for classification and prediction of breast cancer outcomes. This project gives a comparison between the performance of four classifiers: Ensemble Bagging, Logistic Regression, Random Forest which are among the most influential data mining algorithms.

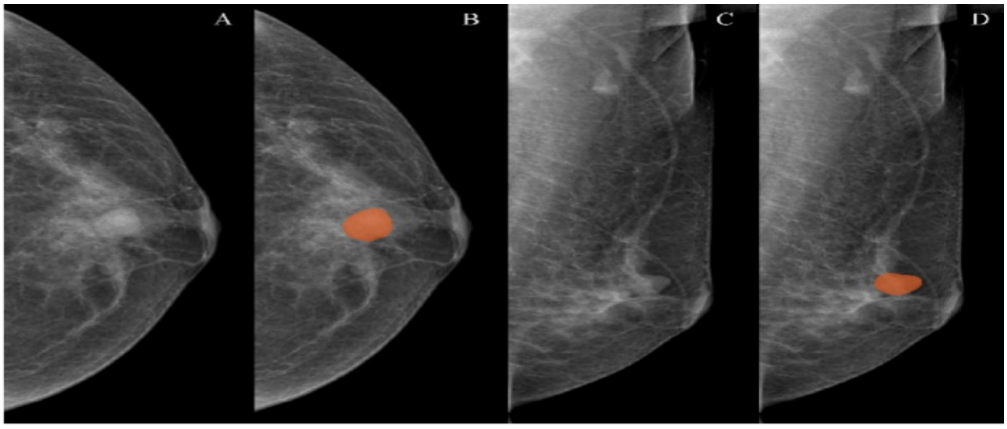
The goal of this project is to identify and classify Malignant and Benign patients and intending how to parametrize our classification techniques hence to achieve high accuracy. We are looking into many datasets and Extracting parameters from Mammography images and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy. 10-fold cross validation test which is a Machine Learning Technique is used in Google collab to evaluate the data and analyze data in terms of effectiveness and efficiency.

1.1 FIGURE



MALIGNANT IMAGE

Figure 1.1: The four factors on which Customer Segmentation Happens.



MALIGNANT IMAGE VIA X-RAY

1.2 SCOPE

As of 2022, on average, 1 in 8 U.S women (approx. 12%) would develop invasive breast cancer at some point during her life. 5-year survival rate for breast cancer is 100% with early detection and 15% with late detection (UK Cancer research) Machine learning (ML) techniques play a key role in healthcare in recent years. In the case of breast cancer, machine learning techniques can be used to distinguish between malignant and benign tumors for enabling early detection.

Most ML based applications focus on large data sets citing ML's ability to handle big data. However, from a user's perspective most users have access to publicly available small data sets. Thus, it is interesting to analyze if the traditional noncomplex basic ML algorithms can achieve high accuracy classifications using small datasets**.**

1.3 NOVELTY OF THE IDEA

Implementing multiple models and comparing with best accuracy and percentage for deployment purposes. Predicting the disease using machine learning algorithms will be very useful for millions of patients to cure it in early stage.

CHAPTER 2
PROBLEM DEFINITION

CHAPTER 2

2.1 PROBLEM DEFINITION

Breast cancer is one of the main causes of cancer death worldwide.Computer-aided diagnosis systems showed the potential for improving diagnostic accuracy. But early detection and prevention can significantly reduce the chances of death. It is important detect breast cancer as early as possible.The main issue pertaining to its cure is early recognition using several other methods.Implementing Machine learning model to Breast cancer will save lives of manypatients. Predicting the cancer in early stages can be cured.

2.2 OBJECTIVES

In vision of the problem statement described in the introduction section, a classification model is proposed with boosted accuracy to predict the breast cancer patient. The framework is composed of the following important phases:

- Dataset Selection
- Data Preprocessing ·
- Learning by Classifier (Training) i.e. Random Forest,
- Linear Regression and Decision Tree·
- Achieving trained model with highest accuracy
- Using trained model for prediction

CHAPTER 3
LITERATURE SURVEY

CHAPTER 3 LITERATURE SURVEY

3.1Machine Learning Techniques to Diagnose Breast cancer by Alireza Osarehand Bita Shadgar [1].

We have investigated the issues of breast cancer diagnosis and prognostic risk evaluation of recrudescence and metastasis by using 3 well-known classifiers i.e., SVM, KNN, PNN. These classifiers were combined with SNR feature ranking method; SFS feature selection and PCA feature extraction based on FNAB dataset I and gene microarrays dataset II, respectively. Feature ranking and filtering supplied the informative and important features to classify breast tumor. It provides the physicians a valuable clue to pay more attention to these relevant features in their clinical breast tumor diagnosis. Feature ranking and filtering also improved the evaluation performance to the prognostic risk of recrudescence and metastasis.

3.2Discovering Mammography-based Machine Learning Classifiers for BreastCancer Diagnosis by Raúl Ramos-Pollán

We presented a first evaluation of a method to design mammography-based machine learning classifiers (MLC) for breast cancer diagnosis, allowing to characterize breast lesions according to BI-RADS classes (grouped by benign and malignant). The results of our investigation confirm that: (1) BCDR contains critical information to create robust datasets of features vectors allowing massive exploration of classifiers search spaces, including the biopsies constituting the golden standard against which classifier performance is evaluated; (2) the testpct40 valida tion method, or possibly some variation, could be considered to be used when one2all might not be computationally feasible,enlarging the exploration possibilities of MLC without reducing their statistical consistency; (3) using features vectors representing the same identified ROI (lesion) in both CC and MLO mammography images increased meaningfully classifiers performance.

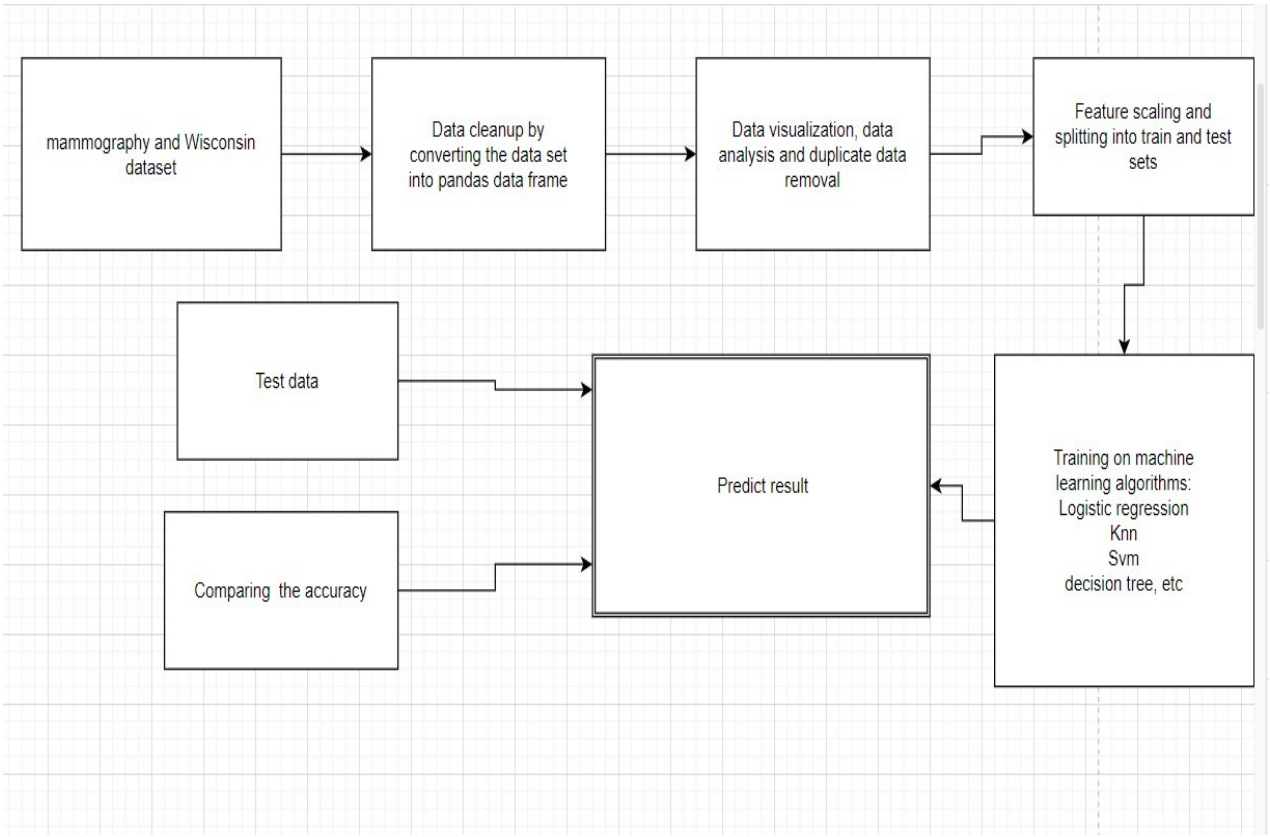
3.3Breast Cancer Classification Using Machine Learning by Meriem AMRANE1 and Ikram GAGAOUA3

On the Wisconsin Breast Cancer datasets, we used our two main algorithms, which are: NB & KNN, since our target and challenge from breast cancer classification is to build classifiers that are precise and reliable. After an accurate comparison between our algorithms, we noticed that KNN achieved a higher efficiency of 97.51%, however, even NB has a good accuracy at 96.19 %, if the dataset is larger, the KNN's time for running will increase.

CHAPTER 4
PROJECT DESCRIPTION

CHAPTER 4 PROJECT DESCRIPTION

4.1 PROPOSED DESIGN



4.2 ASSUMPTIONS AND DEPENDENCIES

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 this project. 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 modelling. In this project, we aim to review ML techniques and Mammography images in BC diagnosis and prognosis.

Firstly, extract the specific parameter data from Mammography images then we provide an overviewof Deep learning techniques including artificial neural networks (CNN), Logistic Regression , decision trees (DTs), and Random Forest. Then, we investigate their applications. Our primary data is drawn from the Wisconsin breast cancer database (WBCD) which is the benchmarkdatabase for comparing the results through different algorithms. Finally, a we separate the benign and malignant tumors.machine learning model is considered based on the highest accuracy obtained among the other models that is tested.

CHAPTER 5
REQUIREMENTS

CHAPTER 5 REQUIREMENTS

5.1 FUNCTIONAL REQUIREMENTS

Implementing Machine learning model to Breast cancer will save lives of many patients.Predicting the cancer in early stages can be cured

Python Librares

Datasets

Mammographic images

ML Libraries Algorithms

Wisconsin Dataset: consist of 580 rows and 32 columns it has the data of one year and Another Dataset from the data world which has more than 2000 instances.

parameters: radius_mean,texture_mean,perimeter_mean,Smoothness_mean,Roicurve,compatness_ mean of the cells

Pre-Processing:

The first phase we do is to collect the data that we are interested in collecting for pre-processing and to apply classification and Regression methods. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and lacking certain to contain many errors. Data pre-processing is a proven method of resolving such issues.

DATA Preparation*:*

Data Preparation, where we load our data into a suitable place and prepare it for use in our machine learning training. We'll first put all our data together, and then randomize the ordering.

Feature Extraction:

In machine learning and statistics, feature selection, also known as variable selection, attribute selection, is the process of selection a subset of relevant features for use in model construction.

Data File and Feature Selection Breast Cancer Wisconsin Data Set from Kaggle repository and out of 31 parameters we have selected about 8-9 parameters and we will extract parameters from the Mammography Images. Our target parameter is breast cancer diagnosis – malignant or benign. We have used Wrapper Method for Feature Selection. The important features found by the study are:

SHAPE DOMIAN	TEXT DOMIAN
Area	Mean
Centroid	Skewness
Perimeter	corelation
Median	Homogentity
Circularity	Contrast
Roundness	Energy
Area Function	kurtosis

5.2 NON-FUNCTIONAL REQUIREMENTS

- Performance: The performance of the developing system will be as good as possible.
- Reliability: The recommendation system will be highly reliable
- Robustness: The system will be robust and secure
- Portability: The system will be highly portable.
- Maintainability: The system can be easily used and maintained.

CHAPTER 6
METHODOLOGY

CHAPTER 6 METHODOLOGY

6.1 PROPOSED METHODOLOGY

Logistic Regression

Logistic regression was introduced by statistician DR Cox in 1958 and so predates the field of machine learning. It is a supervised machine learning technique, employed in classification jobs (forpredictions based on training data). Logistic Regression uses an equation like Linear Regression, butthe outcome of logistic regression is a categorical variable whereas it is a value for other regression models. Binary outcomes can be predicted from the independent variables.

k-Nearest Neighbor (k-NN)

K-Nearest Neighbor is a supervised machine learning algorithm as the data given to it is labelled. It is a nonparametric method as the classification of test data point relies upon the nearest training data points rather than considering the dimensions (parameters) of the dataset.

Random Forest

Random forest, like its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction.

Regression Analysis: Regression analysis describes the relationships between a set of independent variables and the dependent variable. It produces an equation where the coefficients represent the relationship between each independent variable and the dependent variable.

6.2 HARDWARE REQUIREMENTS

1) Intel i5 CPU, with an i7 recommended: A Core i7 will typically be better for multitasking, media-editing and media-creation tasks, high-end gaming, and similar demanding workloads.

2) 8GB RAM, with 16GB recommended: With 16 GB of RAM, you have enough memory to run as many programs as you want without slowing your computer down. This amount of memory is enough for hardcore gamers, video editors, gaming streamers, and anyone using AutoCAD or other demanding software.

3) 1920 x 1080 resolution display.

6.3 SOFTWARE REQUIREMENTS

1) Windows 10 OS/ Unix/ Mac OS.
2) Python3 or Python2 version.
3) Anaconda Prompt + Jupyter Notebook.
4) TensorFlow, scikit learn.

CHAPTER 7

EXPERIMENTATION

CHAPTER 7 EXPERIMENTATION

We are Importing dataset and reading the csv file

```
Breast Cancer Prediction Using Python

[ ]  # importing libraries
     import numpy
     import matplotlib.pyplot as plt
     import pandas as pd
     import seaborn as sns

[ ]  # reading data from the file
     df=pd.read_csv("data.csv")

[ ]  df.head()
```

Shaping and Cleaning or droping the data

```
[8]  # return the size of dataset
     df.shape

     (569, 33)

     # remove the column
     df=df.dropna(axis=1)

[10] # shape of dataset after removing the null column
     df.shape

     (569, 32)

     # describe the dataset
     df.describe()
```

Counting the images

```
[12] # Get the count of malignant<M> and Benign<B> cells
     df['diagnosis'].value_counts()

     B     357
     M     212
     Name: diagnosis, dtype: int64

     sns.countplot(df['diagnosis'],label="count")
```

Performing the Skewdness Graph for the parametrs

```
[13] df_temp = df.drop(columns=['diagnosis'], axis=1)

     fig, ax = plt.subplots(ncols=6, nrows=5, figsize=(20, 20))
     index = 0
     ax = ax.flatten()

     for col in df_temp.columns:
       if index ==30:
            break;
       sns.distplot(df[col], ax=ax[index])
       index+=1


     plt.tight_layout(pad=0.5, w_pad=0.7, h_pad=5.0)
```

performing the heat map for parameters

```
# visualize the correlation
plt.figure(figsize=(10,10))
sns.heatmap(df.iloc[:,1:10].corr(),annot=True,fmt=".0%")
```

Implementing the alogrithums Random forest,logistic regression,decisiontree classifier

```
[21] # split the dataset into dependent(X) and Independent(Y) datasets
     X=df.iloc[:,2:31].values
     Y=df.iloc[:,1].values

[22] # spliting the data into trainning and test dateset
     from sklearn.model_selection import train_test_split
     X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.20,random_state=0)

[23] # feature scaling
     from sklearn.preprocessing import StandardScaler
     X_train=StandardScaler().fit_transform(X_train)
     X_test=StandardScaler().fit_transform(X_test)

[24] # models/ Algorithms

     def models(X_train,Y_train):
             #logistic regression
             from sklearn.linear_model import LogisticRegression
             log=LogisticRegression(random_state=0)
             log.fit(X_train,Y_train)


             #Decision Tree
             from sklearn.tree import DecisionTreeClassifier
             tree=DecisionTreeClassifier(random_state=0,criterion="entropy")
             tree.fit(X_train,Y_train)

             #Random Forest
             from sklearn.ensemble import RandomForestClassifier
             forest=RandomForestClassifier(random_state=0,criterion="entropy",n_estimators=10)
             forest.fit(X_train,Y_train)

             print('[0]logistic regression accuracy:',log.score(X_train,Y_train))
             print('[1]Decision tree accuracy:',tree.score(X_train,Y_train))
             print('[2]Random forest accuracy:',forest.score(X_train,Y_train))

             return log,tree,forest
```

Combing two algorithums using ensemble methods

```
[29] from joblib import dump
     dump(model[2],"Cancer_prediction.joblib")

     ['Cancer_prediction.joblib']

     from sklearn.tree import DecisionTreeClassifier
     from sklearn.ensemble import BaggingClassifier

     bag_model=BaggingClassifier(
         base_estimator=DecisionTreeClassifier(),
         n_estimators=100,
         max_samples=0.8,
         oob_score=True,
         random_state=0
     )

     bag_model.fit(X_train, Y_train)
     bag_model.oob_score_

     0.9516483516483516
```

Experimentation from Mammography images

Importing libraries

```
[ ] import matplotlib.pyplot as plt
    import seaborn as sns

    import numpy as np

    import keras
    from keras.models import Sequential
    from keras.layers import Dense, Conv2D, Flatten , Dropout , BatchNormalization, MaxPooling2D, GlobalAveragePooling2D
    from keras.preprocessing.image import ImageDataGenerator
    from keras.callbacks import Callback, ModelCheckpoint, CSVLogger

    import tensorflow as tf

    import pickle

    from sklearn.metrics import classification_report,confusion_matrix
```

Building the CNN model

```
model = Sequential()
model.add(Conv2D(32, (3, 3), input_shape=(224,224,3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(32, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Flatten())
model.add(Dense(64, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(1, activation='sigmoid'))

model.compile(loss='binary_crossentropy',
              optimizer='rmsprop',
              metrics=['accuracy'])
```

Dividing the images into train ,test,validation sets

```
train = datagen.flow_from_directory('/content/gdrive/My Drive/breast_cancer/train', target_size=(224, 224), class_mode='binary', batch_size=64)
# load and iterate validation dataset
val = datagen.flow_from_directory('/content/gdrive/My Drive/breast_cancer/val/', target_size=(224, 224), class_mode='binary', batch_size=64)
# load and iterate test dataset
test = datagen.flow_from_directory('/content/gdrive/My Drive/breast_cancer/test/', target_size=(224, 224), class_mode='binary', batch_size=64)
```

```
Found 3816 images belonging to 2 classes.
Found 1908 images belonging to 2 classes.
Found 1908 images belonging to 2 classes.
```

```
[ ] imgs, labels = next(train)
```

```
[ ] imgs.shape
```

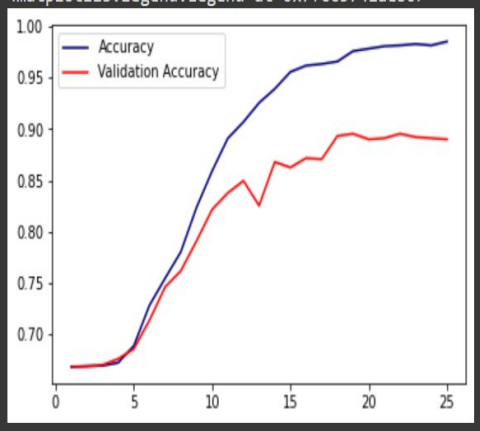
```
(64, 224, 224, 3)
```

```
[ ] train.class_indices
```

```
{'Benign': 0, 'Malignant': 1}
```

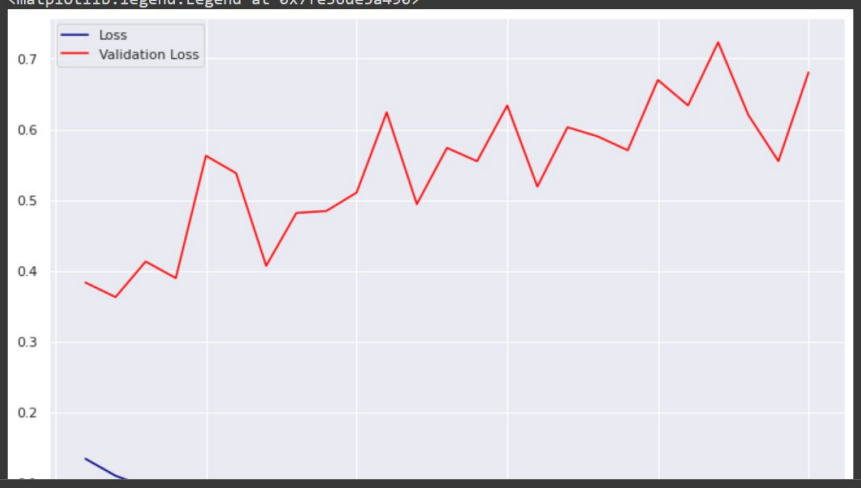
Plotting graph after training the dataset

```
[ ] import pandas as pd
    log_data = pd.read_csv('/content/gdrive/My Drive/breast_cancer/training.log', sep=',', engine='python')

    plt.plot(np.arange(1, len(history.history['accuracy'])+1,1), history.history['accuracy'], color='navy', label = 'Accuracy')
    plt.plot(np.arange(1, len(history.history['accuracy'])+1,1), history.history['val_accuracy'], color='red', label='Validation Accuracy')
    plt.legend()

    <matplotlib.legend.Legend at 0x7fcc5741aeb0>
```



Plotting loss and validation score

```
plt.plot(np.arange(1, len(history.history['loss'])+1,1), history.history['loss'], color='navy', label = 'Loss')
plt.plot(np.arange(1, len(history.history['loss'])+1,1), history.history['val_loss'], color='red', label='Validation Loss')
plt.legend()

<matplotlib.legend.Legend at 0x7fe36de5a490>
```



ROC

```
[ ]  from sklearn.metrics import roc_curve, auc

     fpr , tpr , thresholds = roc_curve (y_test , y_pred_prob)

     area_under_curve = auc(fpr, tpr)
```

```
[ ]  plt.plot([0, 1], [0, 1], 'r--')
     plt.plot(fpr, tpr, label='AUC = {:.3f}'.format(area_under_curve))
     plt.xlabel('False positive rate')
     plt.ylabel('True positive rate')
     plt.title('ROC curve')
     plt.legend(loc='best')
     plt.show()
```

checking scores testing data

```
     score = model.evaluate(X_test, y_test, verbose=0)
     print(f'Test loss: {score[0]} / Test accuracy: {score[1]}')

     Test loss: 0.5016220808029175 / Test accuracy: 0.8900862336158752
```

```
[ ]  score = restored_model.evaluate(X_test, y_test, verbose=0)
     print(f'Test loss: {score[0]} / Test accuracy: {score[1]}')

     Test loss: 0.5016220808029175 / Test accuracy: 0.8900862336158752
```

```
[ ]  y_pred_prob = model.predict(X_test)

     58/58 [==============================] - 2s 27ms/step
```

```
[ ]  # Using the saved model
     y_pred_prob = restored_model.predict(X_test)

     58/58 [==============================] - 1s 20ms/step
```

Developing vgg16 model

```
[ ]  VGG_model = Sequential()
     VGG_model.add(backbone)
     VGG_model.add(Flatten())
     VGG_model.add(Dense(512, activation='relu'))
     VGG_model.add(BatchNormalization())
     VGG_model.add(Dropout(0.5))
     VGG_model.add(Dense(1, activation='sigmoid'))
```

```
[ ]  VGG_model.compile(
             loss='binary_crossentropy',
             optimizer=tf.keras.optimizers.Adam(learning_rate=0.00005),
             metrics=['accuracy']
         )
```

Plotting confusion matrix on VGG16 model

```
from sklearn.metrics import confusion_matrix

sns.set(rc={'figure.figsize':(7.7,6.27)})

sns.heatmap(confusion_matrix(y_test,y_pred),cmap=plt.cm.Blues,annot=True,annot_kws={"size": 32}, fmt='g')
plt.xticks([0.50,1.50], ['Malignant','Benign'], fontsize=20)
plt.yticks([0.50,1.50],['Malignant','Benign'], fontsize=20)

plt.ylabel('True label')
plt.xlabel('Predicted label')

plt.title('Confusion Metrix for Breast Cancer')
```

Resize and displaying Mammography images

```
plt.figure(figsize=(15,15))
for i, (img, prediction, prob, true_label) in enumerate(
    zip(sample_test_images, max_prediction, prediction_probs, sample_test_labels)):
  plt.subplot(5,5,i+1)
  plt.xticks([])
  plt.yticks([])
  plt.grid('off')

  plt.imshow(img)
  plt.xlabel('{} ({:0.3f})'.format(cancer_labels[prediction], prob))
  plt.ylabel('{}'.format(true_label))
```

Fine tuining the VGG16 model

```
set_trainable = False
for layer in backbone2.layers:
    if layer.name == 'block4_conv1':
        set_trainable = True
    if set_trainable:
        layer.trainable = True
    else:
        layer.trainable = False
```

```
[ ] VGG_model_2 = Sequential()
    VGG_model_2.add(backbone2)
    VGG_model_2.add(GlobalAveragePooling2D())
    VGG_model_2.add(BatchNormalization())
    VGG_model_2.add(Dropout(0.5))
    VGG_model_2.add(Dense(1, activation='sigmoid'))
```

```
[ ] VGG_model_2.compile(
            loss='binary_crossentropy',
            optimizer=tf.keras.optimizers.Adam(learning_rate=0.00005),
            metrics=['accuracy']
        )
```

Building tranfer learning Resnet50 model

```
from tensorflow.keras.applications.resnet50 import ResNet50
```

+ Code + Text

```
[ ] backbone3 = ResNet50(input_shape = (224, 224, 3), include_top=False, weights='imagenet')
```

```
Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/resnet/resnet50_weights_tf_dim_ord
94765736/94765736 [==============================] - 3s 0us/step
```

```
[ ] backbone3.training = False
```

```
[ ] ResNet50_model = Sequential()
    ResNet50_model.add(backbone3)
    ResNet50_model.add(GlobalAveragePooling2D())
    ResNet50_model.add(Dropout(0.5))
    ResNet50_model.add(Dense(1, activation='sigmoid'))
```

```
[ ] ResNet50_model.summary()
```

```
Model: "sequential_3"
```

Plotting the Confusion Matrix on Resnet50 Model

```
sns.set(rc={'figure.figsize':(7.7,6.27)})

sns.heatmap(confusion_matrix(y_test,y_pred_4),cmap=plt.cm.Blues,annot=True,annot_kws={"size": 32}, fmt='g')
plt.xticks([0.50,1.50], ['Malignant','Benign'], fontsize=20)
plt.yticks([0.50,1.50],['Malignant','Benign'], fontsize=20)

plt.ylabel('True label')
plt.xlabel('Predicted label')

plt.title('Confusion Metrix for Breast Cancer')
```

CHAPTER 8
TESTING AND RESULTS

CHAPTER 8 TESTING AND RESULTS

Reading the dataset

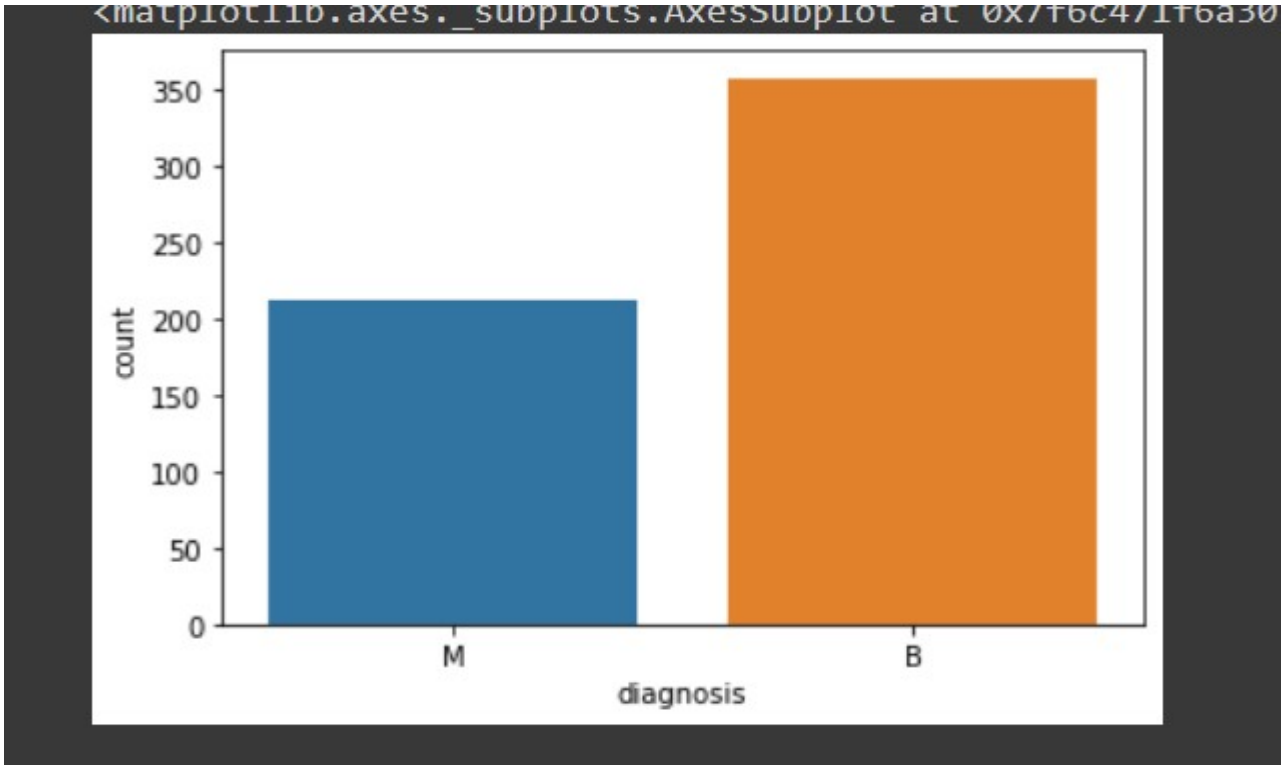
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture_worst
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	17.33
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	23.41
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	25.53
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	26.50
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	16.67

5 rows × 33 columns

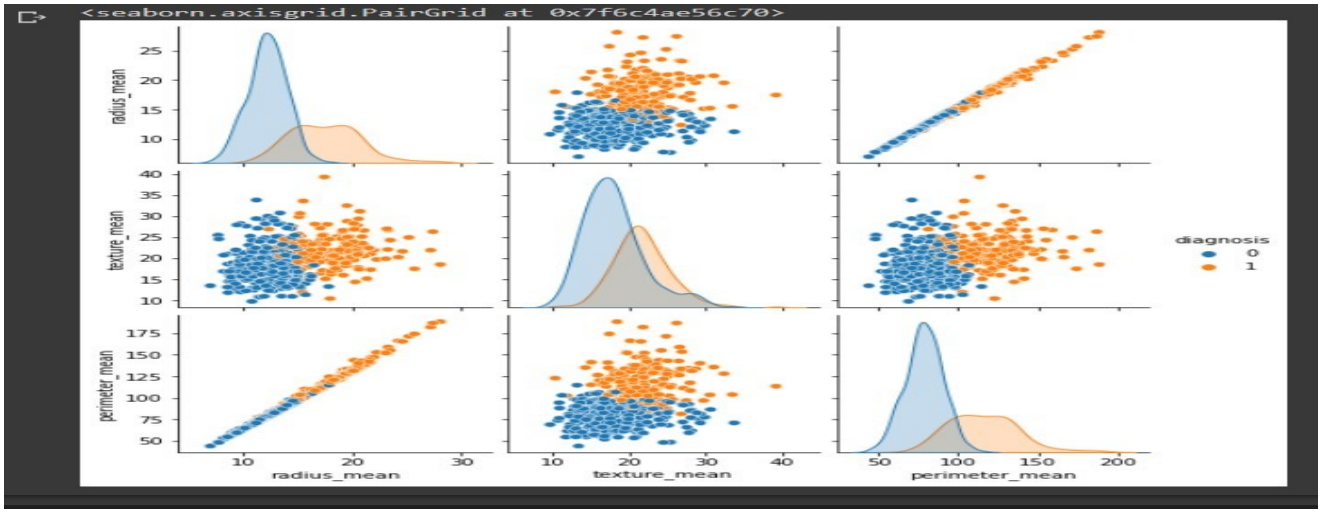
Parameters after droping

```
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
 32  Unnamed: 32              0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

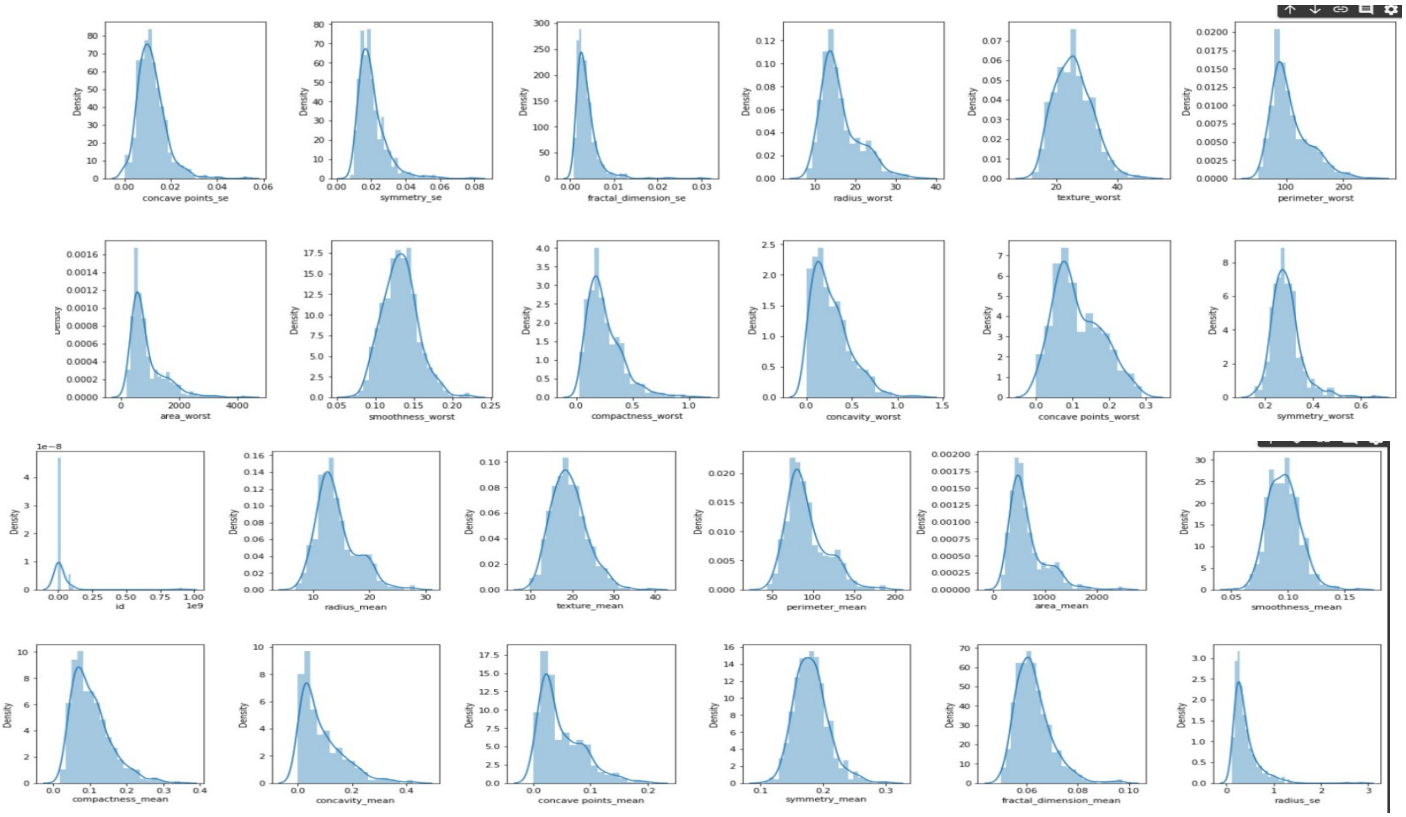
Counting the data from the dataset



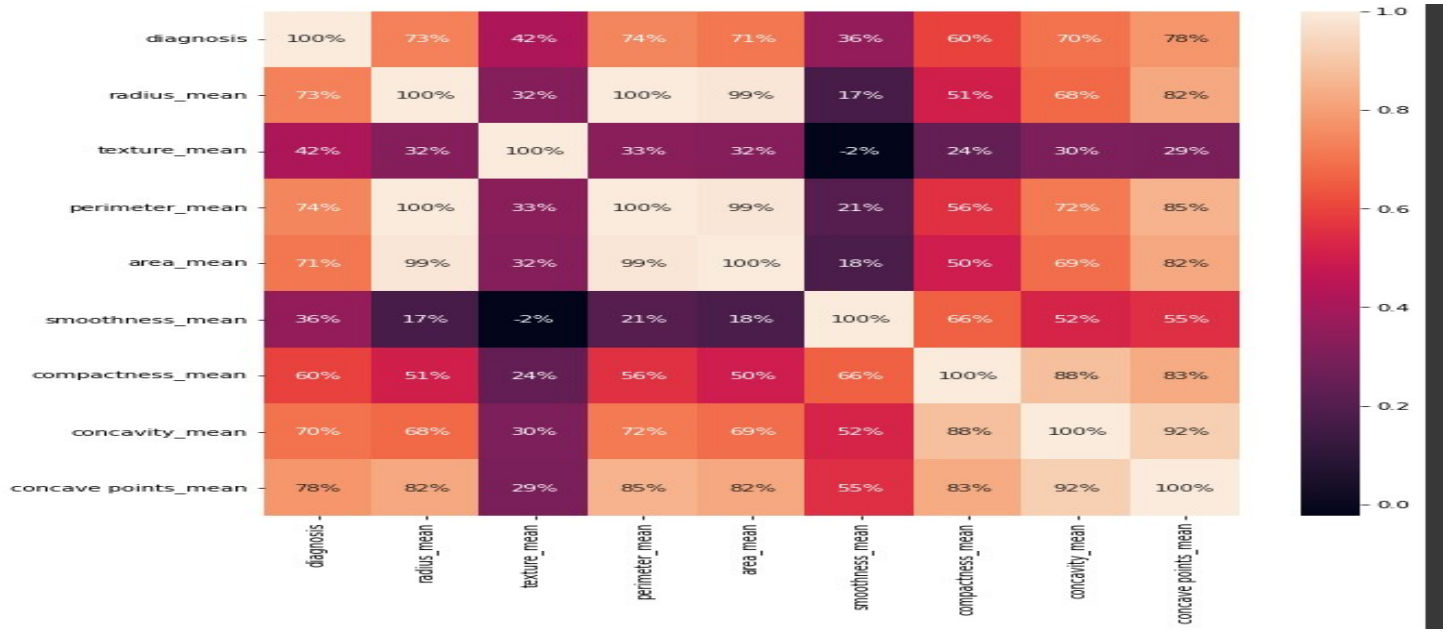
Performing the Visulization on Parameters



Visulizing the skewdness of parameters



Heatmap for parameters



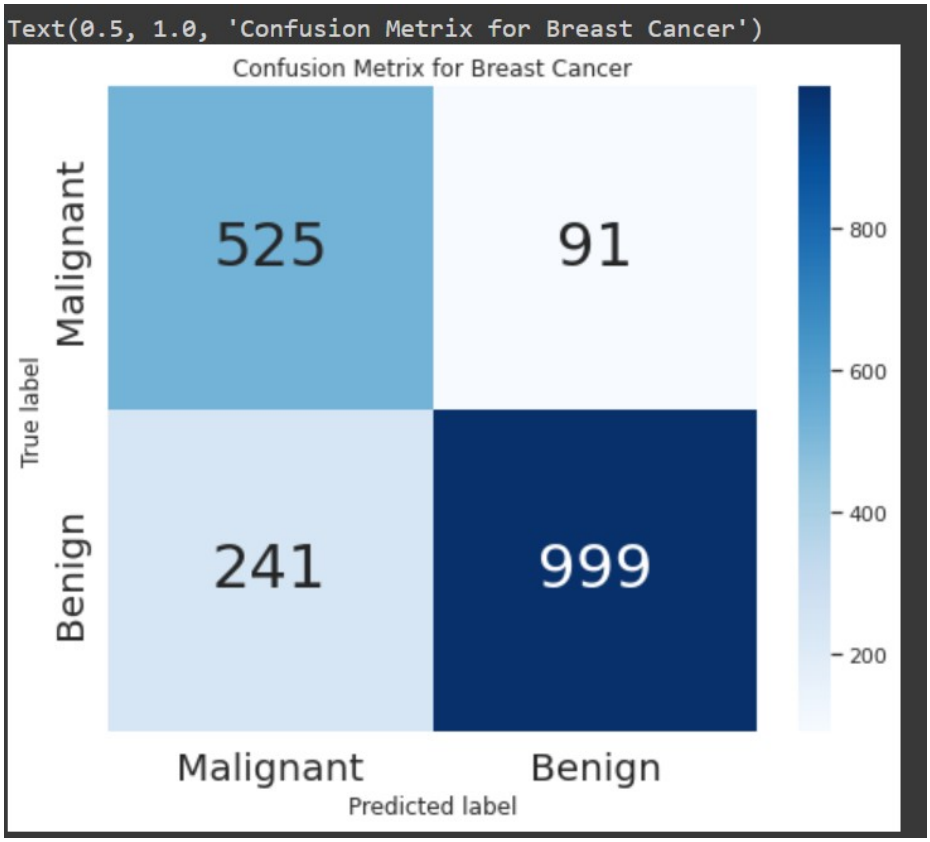
Results from Random forest,logistic regression,decisiontree classifier

```
25] model=models(X_train,Y_train)

    [0]logistic regression accuracy: 0.9912087912087912
    [1]Decision tree accuracy: 1.0
    [2]Random forest accuracy: 0.9978021978021978
```

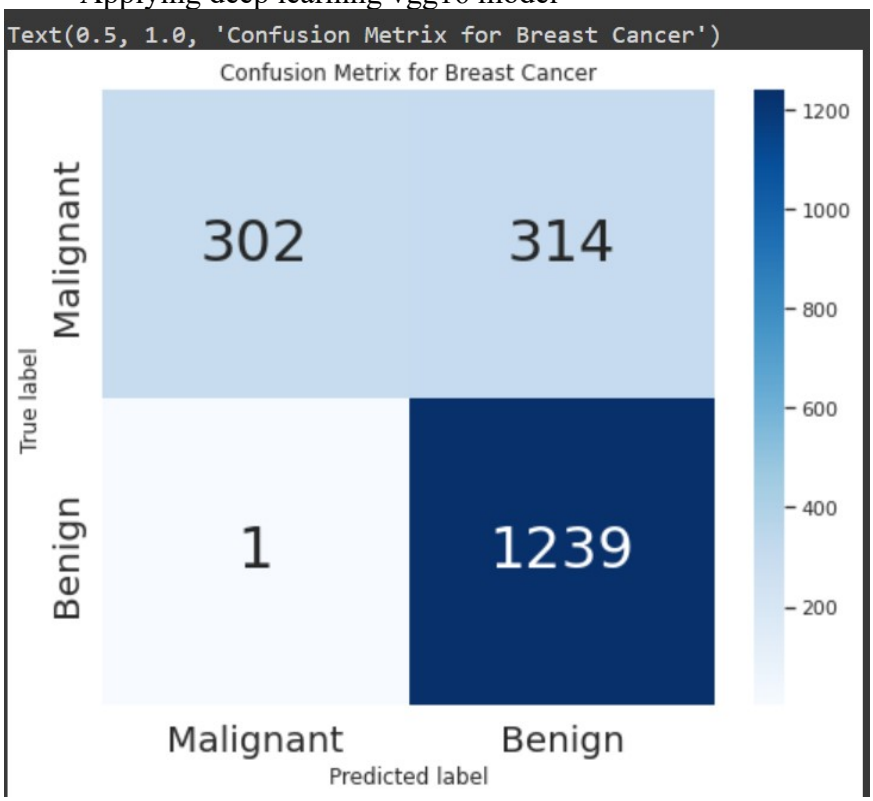
Result from ensembling methods

```
0.9516483516483516
```

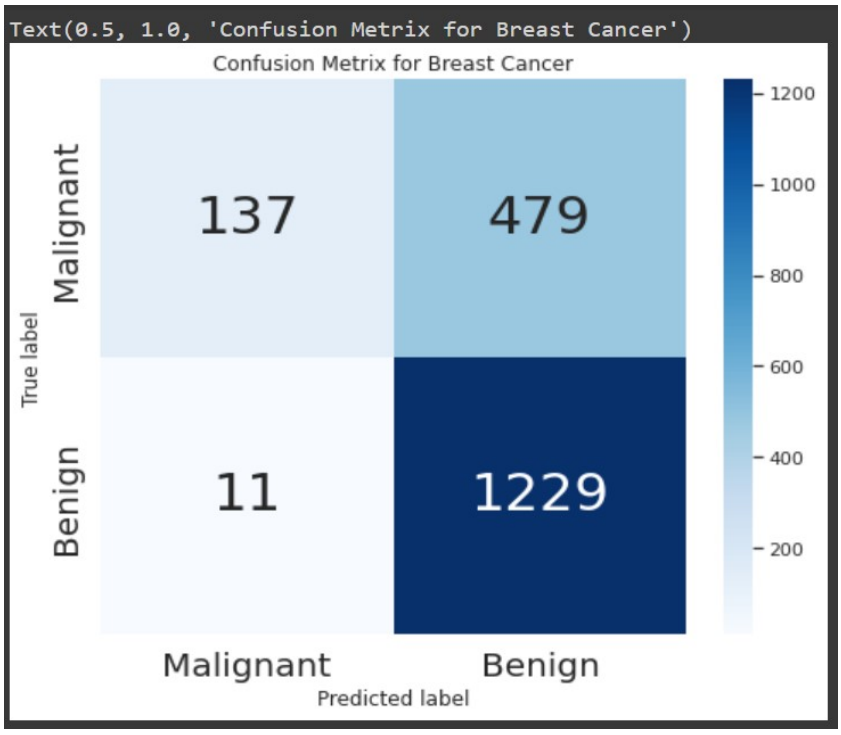
Applying and fine tuning the confusion matrix on mammography images



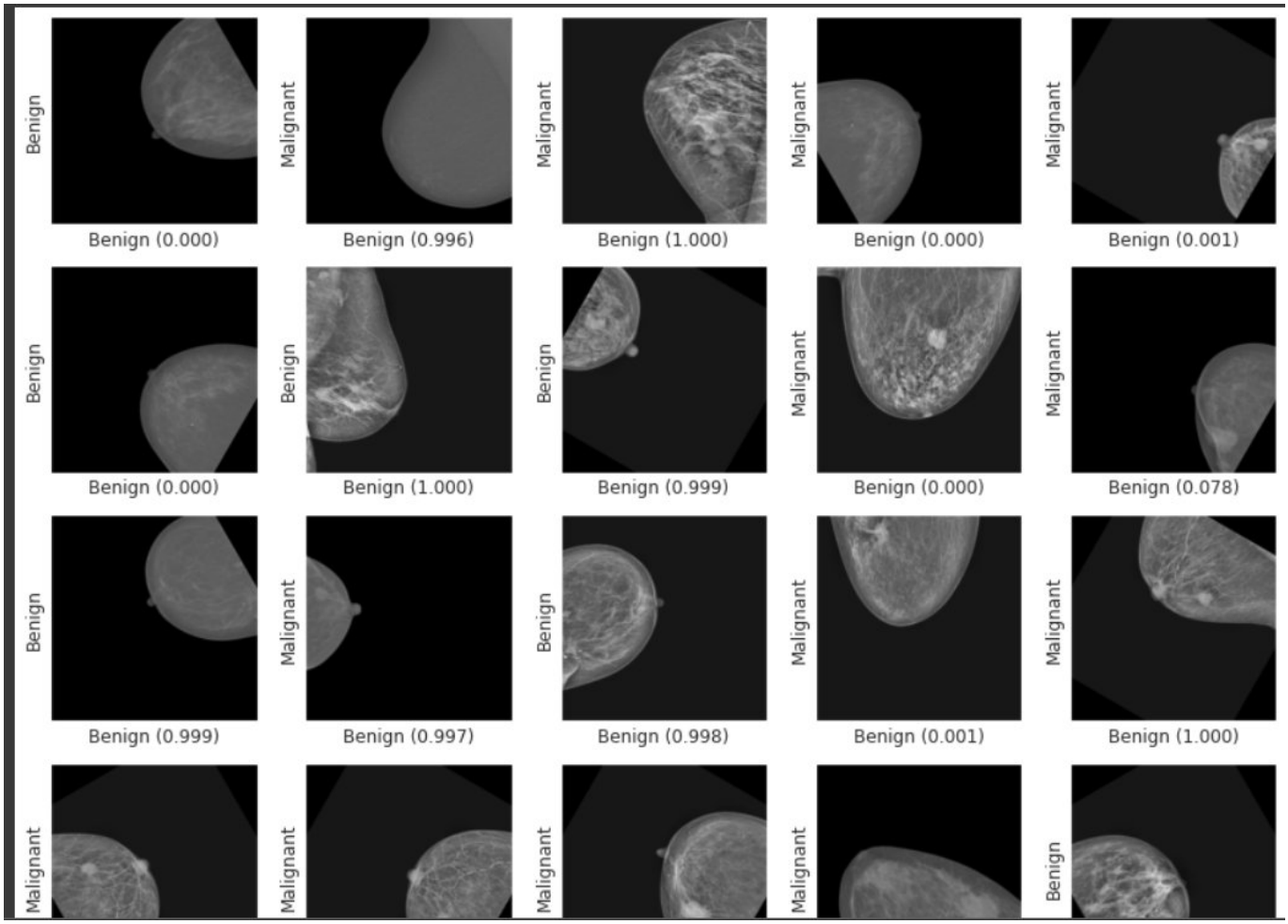
Applying deep learning vgg16 model



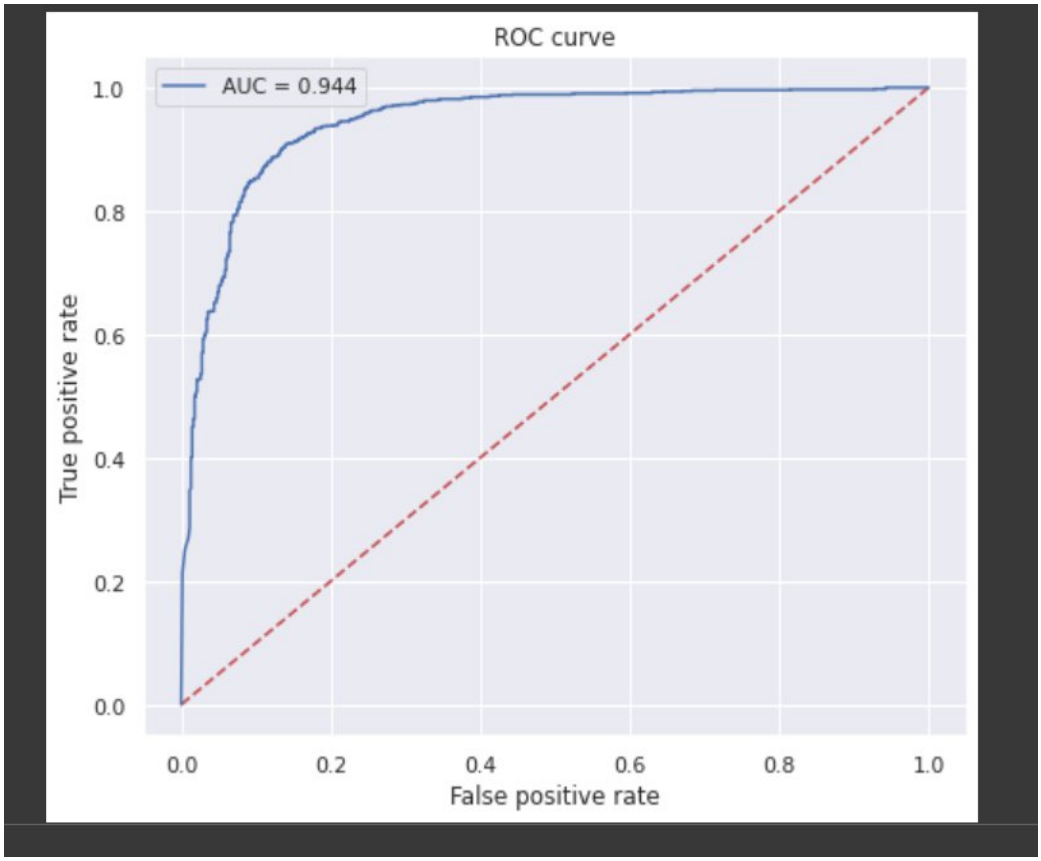
Applying tranfer learning Resnet50 model



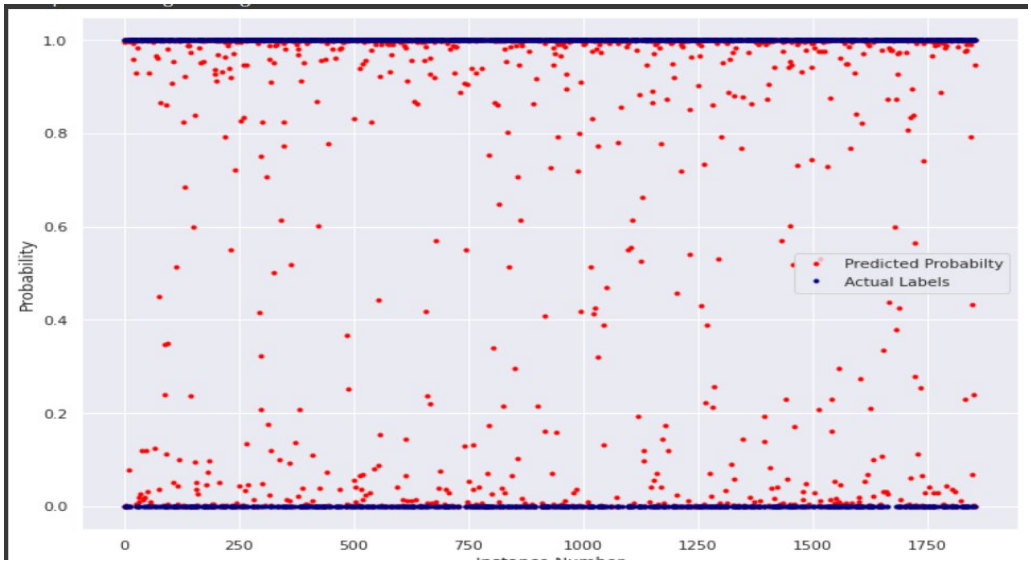
Display the Mammography images



Plotting the ROC curve



Plotting legend graph



CHAPTER 9
CONCLUSION

CHAPTER 9

9.1 CONCLUSION

Our project will deliver a more accurate model by using a quality and managable data. The prediciton obtained by taking parameters from multiple datasets make the model unique and efficient. Along with it, all the different clustering algorithms will be employed at the same time and compared which provides us the solution of which algorithm will best suit this problem case. We will develop a machine learning model that can classify Malignant and Bengin tumor for a Cancer patients using the demographic information from each individual. We'll be employing different clustering approaches to achieve the best accuracy of our model. The models we planned includes Decision Tree algorithm, Liner regression ,Random Forest. Our model will predict which type of cancer tumor for each patients must be provided. Depending on this result, patients can get the cure.

9.2 DELIVERABLES

the Mammography Images can be converted into numerical data and analysis can be drawn from it .Moreover based on this data we can predict the large dataset images and apply confusion matrix by implementing this method we can predict more accurately whether the tumor is bengin or malignant. For normal dataset ,We'll be employing different clustering approaches to achieve the best accuracy of our model. The models we planned includes Decision Tree algorithm, Liner regression ,Random Forest. Our model will predict which type of cancer tumor for each patients must be provided. Depending on this result, patients can get the cure.

CHAPTER 10

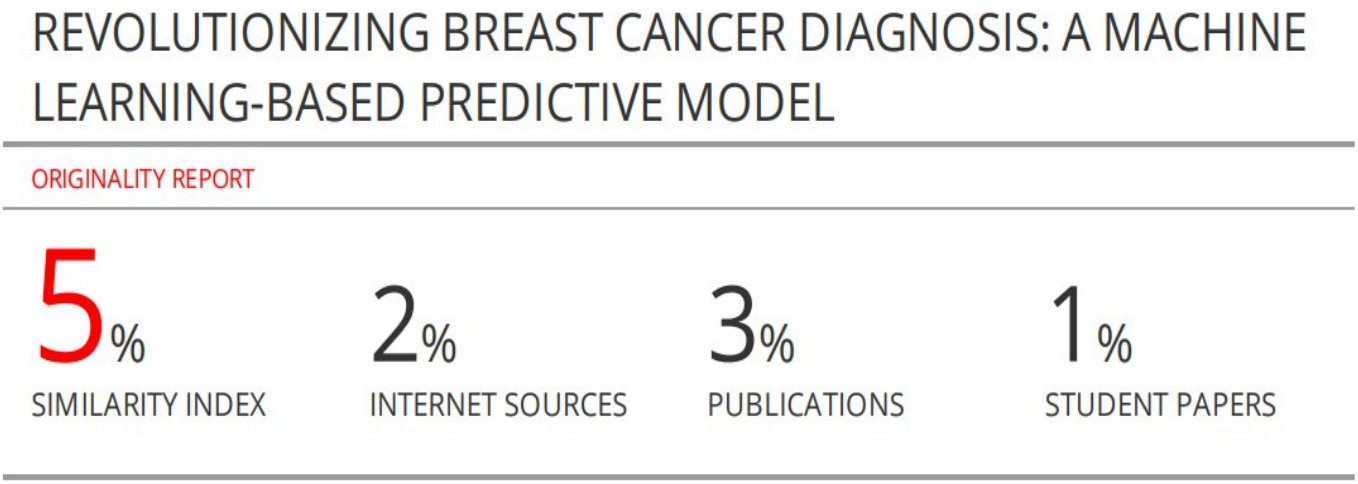
REFERENCES

CHAPTER 10 REFERENCES

- MF. Akay. “Support vector machines combined with feature selection for breast cancer diagnosis”.
- S.Chakraborty, “Bayesian kernel probit model for microarray based cancer classification”.
- T. Jinshan, R.R., X. Jun, I. El Naqa, Y. Yongyi, “Computer-Aided Detection and Diagnosis of Breast Cancer With Mammography.
- JF McCarthy, M.K., PE Hoffman, “Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management”.
- A. Pradesh, A.o.F.S.w.C.B.C.D.,” Indian J. Comput. Sci. Eng., vol. 2, no. 5, pp. 756–763, 2011
- N. Bhatia, "Survey of Nearest Neighbor Techniques", International Journal of Computer Science and Information Securit
- Verma, B., et al., Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer. Expert Syst. Appl. 37:3344–3351, 2010.
- Mavroforakis, M., et al., Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines. Eur. J. Radiol. 54:80–89, 2005.
- Sweilam, N. H., et al., Support vector machine for diagnosis cancer disease: A comparative study. Egypt. Inform. J. 11:81–92, 2010
- Efron, B., Estimating the error rate of a prediction rule: Improvement on cross-validation. J. Am. Stat. Assoc. 78:316– 331, 1983
- Setiono, R., Generating concise and accurate classification rules for breast cancer diagnosis. Artif. Intell. Med. 18:205–219, 2000.

PAPER PUBLICATION DETAILS

Rahul Kumar, Ruchith B M, Sagar M A, Sai lakshmi Sridhar and Shaila S G Wrote Paper on Breast Cancer Prediction



GITHUB REPO FOR CODE:

https://github.com/CSE-DSU/TEAM-84-BREAST-CANCER-PREDICTION-MODEL-USING-ML