MACHINE LEARNING BASED SYSTEM FOR CANCER PREDICTION

Mini Project-2A Report

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Engineering (Computer Engineering)

by:

Soham Chakraborty	**TU3F1920046**
Ganesh Sunil Raje	**TU3F1920004**
Akhil Anjneulu Boddul	**TU3F1920011**
Vivek Dilip Patil	**TU3F1920024**

Under the Guidance of

Dr. Shaveta Malik

Department of Computer Engineering

TERNA ENGINEERING COLLEGE

Nerul (W), Navi Mumbai 400706

(2021-2022)

Internal Approval Sheet

TERNA ENGINEERING COLLEGE, NERUL

Department of Computer Engineering

Academic Year 2020-21

CERTIFICATE

This is to certify that the mini project entitled **"MACHINE LEARNING BASED SYSTEM FOR CANCER PREDICTION"** is a bonafide work of

Soham Chakraborty	**TU3F1920046**
Ganesh Sunil Raje	**TU3F1920004**
Akhil Anjneulu Boddul	**TU3F1920011**
Vivek Dilip Patil	**TU3F1920024**

submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the Bachelor of Engineering (Computer Engineering).

Guide **Head of Department** **Principal**

Approval Sheet

Project Report Approval

This Mini Project Report – entitled "**MACHINE LEARNING BASED SYSTEM FOR CANCER PREDICTION**" by following students is approved for the partial fulfillment of degree of **B.E. in "Computer Engineering"**.

Submitted by:

Soham Chakraborty	**TU3F1920046**
Ganesh Sunil Raje	**TU3F1920004**
Akhil Anjneulu Boddul	**TU3F1920011**
Vivek Dilip Patil	**TU3F1920024**

Examiners Name & Signature:

1._____-

2._____

Date: _____

Place: _____

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Soham Chakraborty	TU3F1920046	----------
Ganesh Sunil Raje	TU3F1920004	----------
Akhil Anjneulu Boddul	TU3F1920011	----------
Vivek Dilip Patil	TU3F1920024	----------

Date: _____

Place: _____

Acknowledgement

We would like to express our sincere gratitude towards our guide **Dr. Shaveta Malik** and Project Coordinators for their help, guidance and encouragement, they provided during the project development. This work would have not been possible without their valuable time, patience and motivation. We thank them for making my stint thoroughly pleasant and enriching. It was great learning and an honor being their student.

We are deeply thankful to **Dr. Archana Mire (H.O.D Computer Department)** and entire team in the Computer Department. They supported us with scientific guidance, advice and encouragement, they were always helpful and enthusiastic and this inspired us in our work.

We take the privilege to express our sincere thanks **to Dr. L. K. Ragha** our Principal for providing the encouragement and much support throughout our work.

Soham Chakraborty	TU3F1920046	----------
Ganesh Sunil Raje	TU3F1920004	----------
Akhil Anjneulu Boddul	TU3F1920011	----------
Vivek Dilip Patil	TU3F1920024	----------

Date: _____

Place: _____

Table of contents

Abstract

Cancer is one of the deadliest and most widespread diseases known to humanity. The successful treatment of this disease is directly correlated to the stage at which it is detected. It's both fortunate and unfortunate that cancer presents a wide array of symptoms which indicates it's existence. This wide array of symptoms have proved to be often ignored by physicians and specialists as they are too fine to be detected by human doctors. These symptoms present themselves as a tumour and manifest in the form of specific nature of the tumour. Some of these are dimension of tumour, texture of tumour, location of tumour etc.

However these specific details about the tumour are present in the form of raw numeric data which often follows predictable trends. This data though mind boggling and impossible to process for a human physician can be analysed by computers in very short time. Further upon analysis we stand to discover important trends which can now be used to predict a diagnosis on further data from patients.

In this particular project implementation we intent upon analysing already available raw, anonymous patient data by using basic data processing and machine learning techniques. We then can demonstrate the accuracy of different machine learning models and also gain insightful data which indicates which parameters related to a cancer tumour most significantly show that a tumour is malignant and needs to be treated.

I. List of Figures

II. List of Tables

1. Introduction

1.1 A Brief Introduction

One of the most critical and important tasks within the field of bio-informatics is the detection and classification of a disease by using raw biopsy data obtained from a patient. The amount of data which is presented to medical officers and physicians is extremely vast and very little has been done till date to make this process automated and efficient. The requirement of human intervention for classifying medical data leaves circumstances of errors and unwanted biases which may prove to be detrimental towards patient health.

Referring to Indian government data, breast cancer is the most common type of cancer affecting Indian women as well as women world wide. It accounted for nearly 37.7% of all cancer cases among Indian women. A record 2,76,480 cases were reported in 2020 meaning that a new case was reported every 4 minutes.

Breast cancers firmly establishes itself as the leading cause of death in Indian women. Breast cancer is caused due to the presence of abnormal cells with damaged DNA sequences. This cells multiple uncontrollably and form the “lump of cells” which is one of the earliest indicators for presence of cancer. Many techniques have been now introduced for screening and detection of such cancers, commonly used ones are x-ray and mammography scans for determining cancer cell presence.

In the very recent years, various machine learning models were used for performing medical diagnosis. Though a large number of parameters were taken into account, the machine learning models were able to provide consistently accurate results. In traditional medical settings a physician would visually check and go through a large number of scans, this will make the diagnostic process less accurate and also time consuming taking upwards 1.5 hours per patient.

In this project we discuss, analyse and implement a machine learning based system for providing automated diagnosis of breast cancer by applying data mining techniques on patient biopsy data.

1.2 Aim and Objective

In the elementary system proposed here the following objectives would be the the guidelines directing the progress of this project.

1. To create a simple and accurate breast cancer prediction model.
2. The system must not require extensive development process involving professional software engineering teams.
3. The system must be able to classify tumours into malignant or benign categories in less amount of time.
4. To use a machine learning classification method to fit a function that can predict the disease.

1.3 Scope of project

The proposed system would be functional to the following extents-

- It uses Linear Regression, Decision Tree and Decision Forest algorithms to predict the probability of an instance having a certain outcome.
- Scikit-learn machine learning library is used for applying various algorithms for building the model.
- These algorithms will analyse the data from the dataset to predict whether the patient has breast cancer or not and it will also tell us it is malignant or benign.
- The model helps to minimize the amount of money and time spent by patients to undergo tests.
- Well-validated predictions using this model could assist in a better way personalized care and treatment, and improve the control over the cancer development.

2. Literature Survey

The use of data mining techniques for classifying cancer tumours has been looked into and tried for a long time, Due to the extensive research performed by data scientists, physicians and mathematicians, a wide variety of data, conclusions and techniques are now available as technical literature. From this vast wealth of information the following are some of the papers found most relevant to this project and provided valuable insights.

Y Chen 2021[1] et al have given an in depth study on how ML based systems can be made integrated within the current diagnostic procedures used by physicians, further they have outlined on measures to deal with typical anomalies usually present within medical data.

In their exhaustive paper **Sultana 2018[2]** et al have studied the performance of various classification algorithms when put to the task of classifying cancerous growths they concluded that Logistic regression had the best performance in this regard.

Cruz, Joseph A, and David S Wishart 2014[3] have excellently summarised most of the research performed till now in the field of Machine learning assisted cancer care in the three critical fields of predicting existence of cancer, risk factors and chances of recurrence. so the the accuracy of using data mining based methods were compared with regard to traditional expert based methods and the conclusion was that data mining greatly increases predictive cancer diagnosis accuracy but it must be used alongside traditional bio-marker based techniques to be fit for clinical implementation.

3. Proposed System

In order to achieve the required functionality of classifying tumour data with satisfactory accuracy levels, the system is designed according to the following framework:-

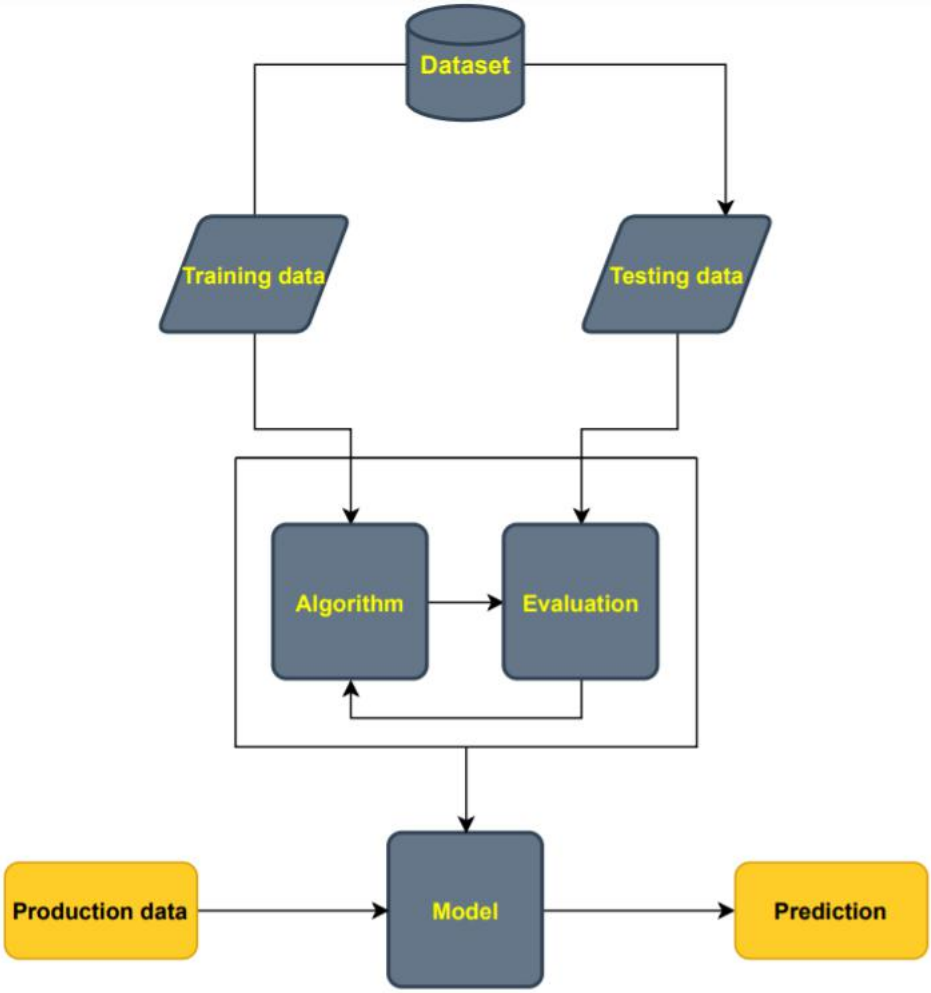

Figure 3.1- Proposed system

First raw mammogram scans are analyzed and the images in them are scaled to collect tumour dimensions on multiple parameters and those are stored in comma seperated values(CSV) format documents.

The dataset is then separated into training and testing datasets having an 80-20 ratio respectively.

The training dataset is then fed into the classification algorithms to train them.

Once the algorithms have been trained, They are now used to predict the category of the remaining 20% of data earlier classified as test set.This allows us to gauge the accuracy of our model and modify parameters accordingly for more accurate predictions.

Now that the model is trained, tested and their accuracy is noted, we can use this model as a backend service for web based applications which make using machine learning based predictive cancer treatment more accessible for physicians.

4. System methodology

The following diagram indicates the flow of operations in the proposed system:-

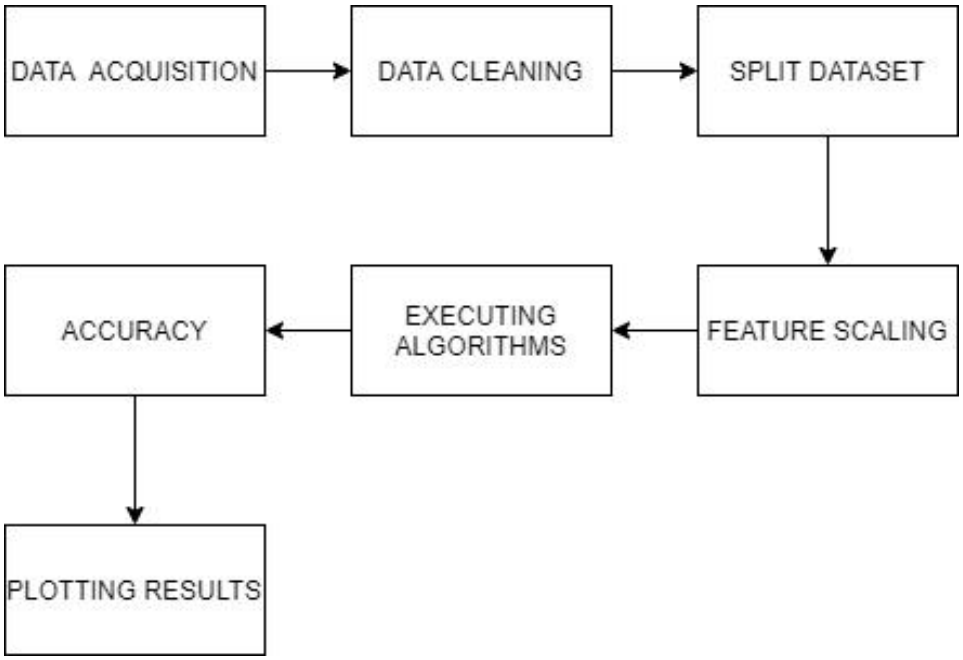

Figure 4.1-System methodology

4.1 Project processes

- Data is first collected and stored in CSV format.
- Data is cleaned by dropping Columns having NULL values and replacing outliers from the data set by replacing them with mean values.
- Feature scaling is done to normalize all values in the dataset.
- All the three algorithms - logistic regression, decision tree & decision forest are executing one by one on the cleaned and normalized dataset.
- The classifications predicted by the three algorithms on the test set data are compared with the actual classification of data for determining the most accurate algorithm of the three.
- The results are plotted into scatter charts and heat maps so that the results can be easily interpreted.

5. System requirements

5.1 Software requirements

Software name	Version/specification
1.Visual Studio Code	Version 1.51
2.Visual Studio code python extension	LTE version
3.Python Interpreter	Version 3.9.0 (5th October 2020 release)
4.Operating System	OS X Yosemite and above Windows 7 (with .NET Framework 4.5.2), 8.0, 8.1 and 10 (32-bit and 64-bit) Linux (Debian): Ubuntu Desktop 14.04, Debian 7

Table 5.1-Software requirements

5.2 Hardware requirements

Hardware name	Version/Specification
1.Processor	2.5 Ghz or faster
2.RAM	Minimum 1GB
3.Memory	Minimum 400MB

Table 5.2-Hardware requirements

6. Design and Implementation

6.1 System Architecture

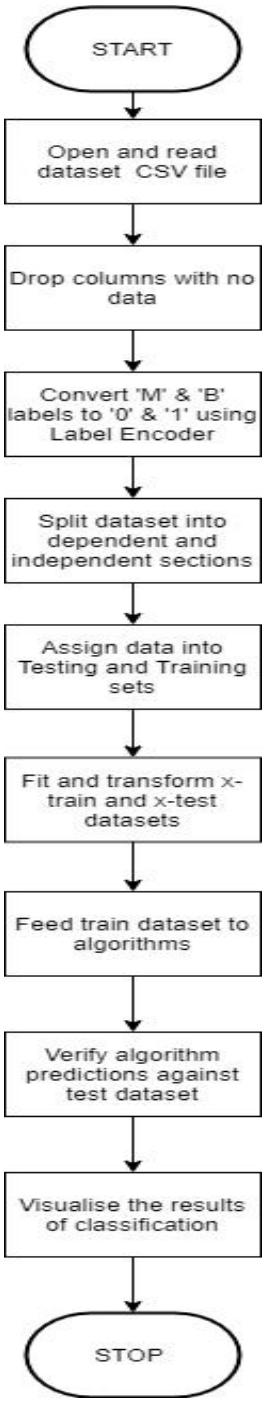

Figure 6.1-System Architecture

6.2 Architecture modules

Architecture steps	Description
1.Open file	First the dataset stored in CSV format is opened and read.
2.Drop blank columns.	Here the columns with missing values are dropped.
3.Label encoding	The dataset indicated malignant and benign tumours as ‘B’ & ‘M’ which are difficult to work with. They are instead replaced bu ‘0’ & ‘1’.
4.Splitting into dependent and independent data sets.	From the values present in the dataset they are split into X and Y dimensions.
5.Train test split	25% of data is designated as test data & remaining 75% of data is designated as training data set.
6.Fitting and transforming data.	The X dimension data is transformed and fitted so that it can be fed to classifying algorithms.
7.Running of algorithms	All the test & training data is analyzed by the algorithms.
8.Verify algorithm predictions.	The outputs obtained by algorithms is checked against the test dataset output to determine accuracy of classification.
9.Visualise the results	The summary of results is presented in the form of visual graphs.

Table 6.1-Architecture modules

As illustrated in Figure 6.1, the first obtains data from a CSV file by extracting the data values from the file and creating a dataframe. This data is not perfect, hence to further refine data the columns with no values are dropped from the dataframe. The class of a cancer case is labeled as ‘M’ indicating a malignant tumour and ‘B’ indicating a benign tumour, this values are instead replaced by ‘1’ & ‘0’ respectively for making the task of processing them simpler. The dataset is now split into dependent and independent datasets. Again another split is made where 75% data is reserved for training the final model and 25% is reserved for testing the model. The algorithms are now fed the training dataset for building their prediction models. Once the process of building is completed the testing dataset reserved earlier is used to test

the dataset and evaluate it's performance metrics. The final results are then visualised in the forms of charts and graphs.

6.3 Language Used

- For implementing this project the language of choice was python.
- Due to it's very expressive syntax which allowed for reducing the lines of code and making production and testing process much more efficient.
- Also the presence of large number of application based libraries makes this a lucrative choice for such projects.

6.4 Libraries used

The following open source libraries were used for implementing this project:-

Name of Library	Purpose of Library
1.Numpy	Used for creating array objects.
2.math	For performing mathematical operations.
3.Matplotlib	For plotting graphs.
4.Pandas	For creating dataframes.
5.seaborn	For visualising dataframes and graphs.
6.sklearn	For applying machine learning algorithms.

Table 6.2-Libraries used

6.5 Detailed description of modules

- **Numpy**
  - numpy is the fundamental package for scientific computing in Python. Using it, we can create multidimensional array objects and perform various operations and manipulations on them.

- **Math**
  - math module is a standard module in Python by which we can use a variety of mathematical functions like trigonometric, representation, logarithmic, etc.

- **Matplotlib & pyplot**
  - Matplotlib is a low level graph plotting library in python that serves as a visualization utility.
  - Most of the matplotlib utilities lies under the pyplot submodule. matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure.

- **Seaborn**
  - seaborn is a data visualization library built on top of matplotlib predominantly used for making statistical graphics and is closely integrated with pandas dataframes. The graphs can also be customized easily.

- **Scikitlearn**
  - Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering, etc.

6.6 Limitations

While implementing this system the following limitations were encountered:-

1. Small training dataset

- The small dataset being used leads to highly accurate over fitted models which may not be as accurate while dealing with real world data

2. Requirement of highly refined dataset

- The dataset used within the created system is highly refined and cleaned. It might prove to be difficult datasets with such high purity.

3. Insufficient resources & time

- Machine learning and it's implementation in various utilities is something that requires rigorous research and testing.

- Such operations often demand significant amount of financial resources as well as other research tools.
- This lack of resources often impedes research work and leads to production of very specialized systems which are not economical neither are they widely and easily available.

7. Results on implementation

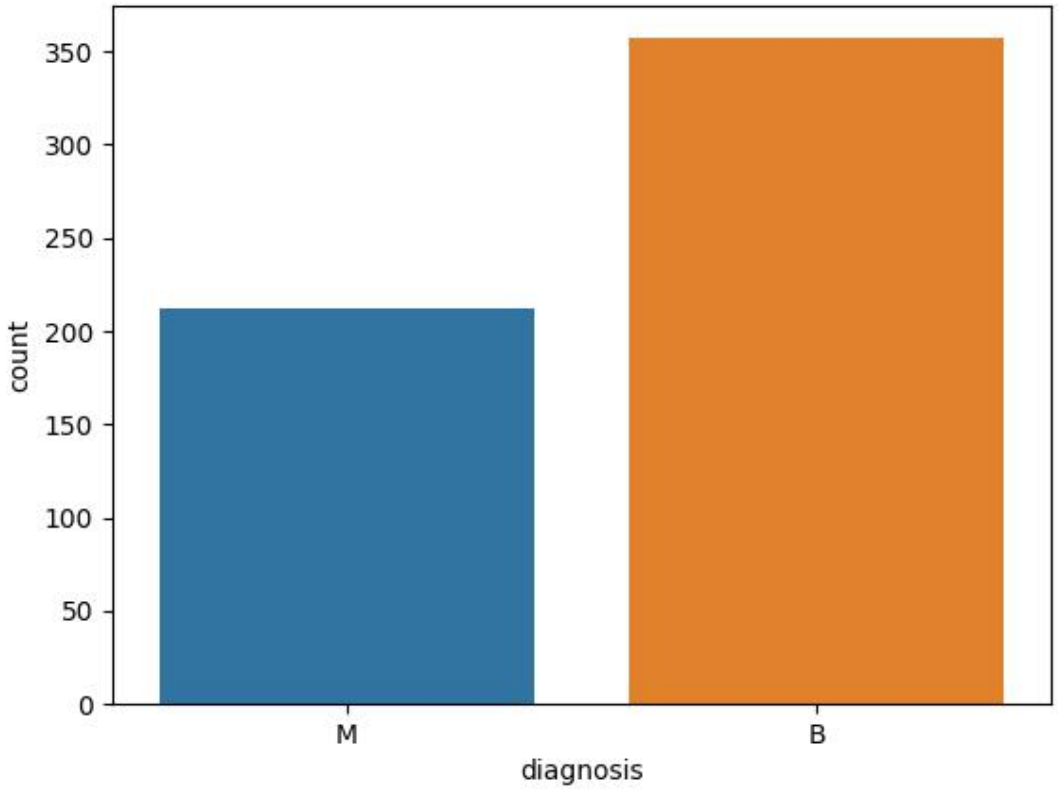

Figure 9.1-This shows the number of benign and malignant tumour cases present in the dataset.

Benign tumors are those that stay in their primary location without invading other sites of the body. They do not spread to local structures or to distant parts of the body. Benign tumors tend to grow slowly and have distinct borders.

Malignant tumors have cells that grow uncontrollably and spread locally and/or to distant sites. Malignant tumors are cancerous (ie, they invade other sites). They spread to distant sites via the bloodstream or the lymphatic system. This spread is called metastasis. Metastasis can occur anywhere in the body and most commonly is found in the liver, lungs, brain, and bone.
Malignant tumors can spread rapidly and require treatment to avoid spread. If they are caught early, treatment is likely to be surgery with possible chemotherapy or radiotherapy. If the cancer has spread, the treatment is likely to be systemic, such as chemotherapy or immunotherapy.

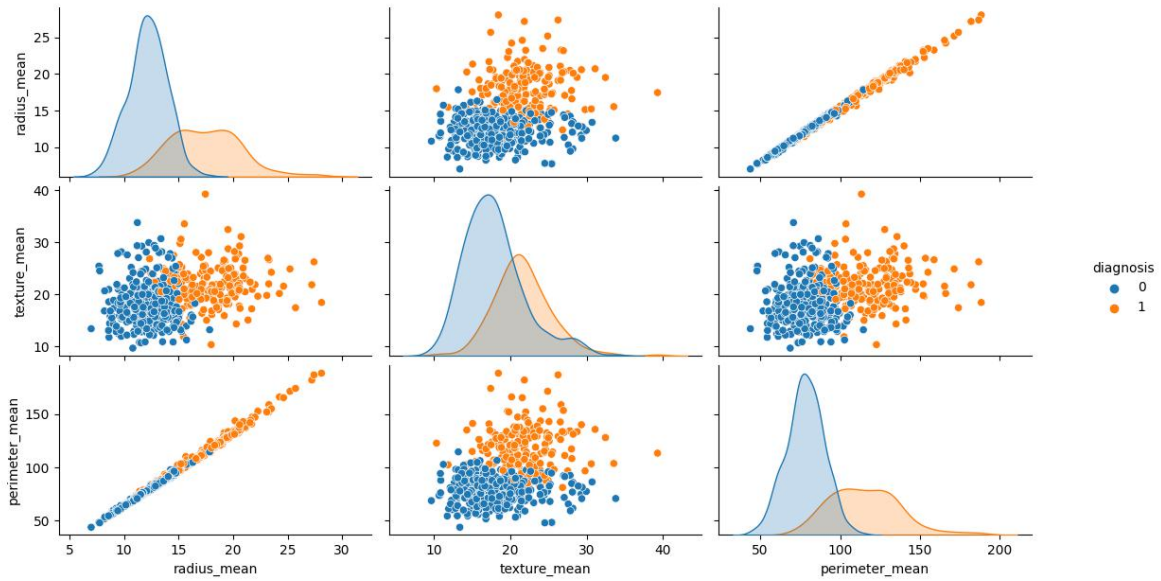

Figure 9.2 - Several scatter plots in this diagram show how different biological parameters vary over malignant and benign tumours.

A scatter plot (or scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.Scatter plots' primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

Identification of correlational relationships are common with scatter plots. In these cases, we want to know, if we were given a particular horizontal value, what a good prediction would be for the vertical value.

We can often see the variable on the horizontal axis denoted an independent variable, and the variable on the vertical axis the dependent variable. Relationships between variables can be described in many ways: positive or negative, strong or weak, linear or nonlinear.

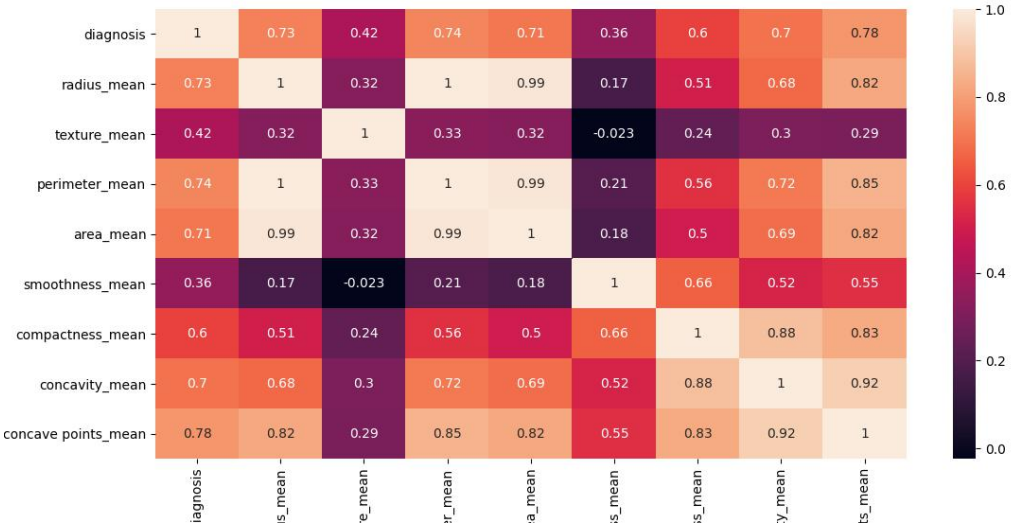

Figure 9.3-Heat map demonstrating the correlation between different parameters of malignant tumours.

A heat map (or heatmap) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

There are two fundamentally different categories of heat maps: the cluster heat map and the spatial heat map. In a cluster heat map, magnitudes are laid out into a matrix of fixed cell size whose rows and columns are discrete phenomena and categories, and the sorting of rows and columns is intentional and somewhat arbitrary, with the goal of suggesting clusters or portraying them as discovered via statistical analysis. The size of the cell is arbitrary but large enough to be clearly visible.

By contrast, the position of a magnitude in a spatial heat map is forced by the location of the magnitude in that space, and there is no notion of cells; the phenomenon is considered to vary continuously. Here in figure 9.3 we can observe that radius and mean area of a tumour are heavily correlated whereas the same cannot be said for mean smoothness and texture.

```
Model 1 - Logistic Regression
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        67
           1       0.98      0.94      0.96        47

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114

Confusion Matrix:
 [[66  1]
 [ 3 44]]
Accuracy :  0.9649122807017544
Specificity :  0.9850746268656716
```

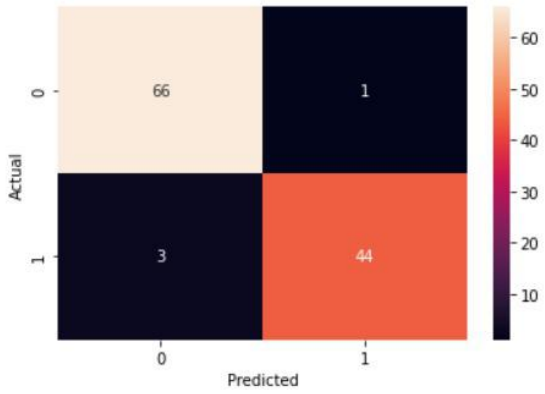

Figure 9.4 - Accuracy achieved by Logistic regression algorithm

```
Model 2 - Decision Tree
              precision    recall  f1-score   support

           0       0.94      0.96      0.95        67
           1       0.93      0.91      0.92        47

    accuracy                           0.94       114
   macro avg       0.94      0.94      0.94       114
weighted avg       0.94      0.94      0.94       114

Confusion Matrix:
 [[64  3]
 [ 4 43]]
Accuracy :  0.9385964912280702
Specificity :  0.9552238805970149
```

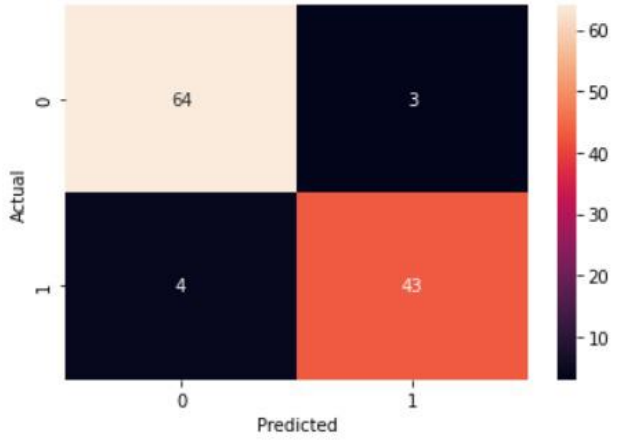

Figure 9.5 - Accuracy achieved by Decision tree algorithm

```
Model 3 - Random Forest
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        67
           1       1.00      0.94      0.97        47

    accuracy                           0.97       114
   macro avg       0.98      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114

Confusion Matrix:
 [[67  0]
 [ 3 44]]
Accuracy :  0.9736842105263158
Specificity :  1.0
```

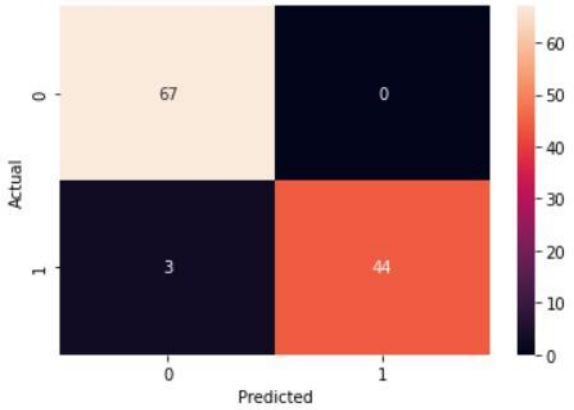

Figure 9.6 - Accuracy achieved by Decision forest algorithm

Figures 9.4,9.5 & 9.6 depict confusion matrices for all the three algorithms respectively. A confusion matrix, in predictive analytics, is a two-by-two table that tells us the rate of false positives, false negatives, true positives and true negatives for a test or predictor. We can make a confusion matrix if we know both the predicted values and the true values for a sample set.

In machine learning and statistical classification, a confusion matrix is a table in which predictions are represented in columns and actual status is represented by rows. Sometimes this is reversed, with actual instances in rows and predictions in columns. The table is an extension of the confusion matrix in predictive analytics, and makes it easy to see whether mislabeling has occurred and whether the predictions are more or less correct.

A confusion matrix is also known as an error matrix, and it is a type of contingency table.

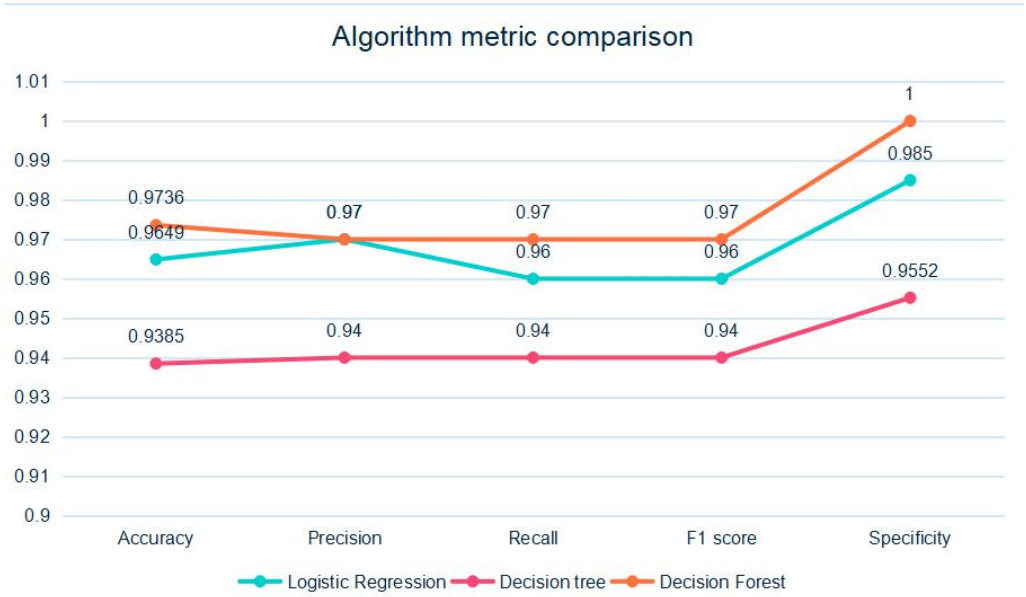

Figure 9.7 - Comparison of all the three algorithms in terms of their accuracy, precision, Recall, F1 score and specificity

On comparison of classification performance of the three algorithms the following results we obtained:-

Algorithm	Accuracy	Precision	Recall	F1 score	Specificity
Logistic Regression	0.9649	0.97	0.96	0.96	0.985
Decision tree	0.9385	0.94	0.94	0.94	0.9552
Decision Forest	0.9736	0.97	0.97	0.97	1

Table 7.1 - summary of performance metrics of algorithms

The following metrics were used to compare the algorithms:-

1. **Accuracy**:- It is the ratio of number of correct predictions to the total number of input samples. It works well only if there are equal number of samples belonging to each class.

2. **Precision**:- It is the number of correct positive results divided by the number of positive results predicted by the classifier.

3. **Recall**:- It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

4. **F1 score**:- F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

5. **Specificity**:- Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative). This implies that there will be another proportion of actual

negative, which got predicted as positive and could be termed as false positives. This proportion could also be called a false positive rate.

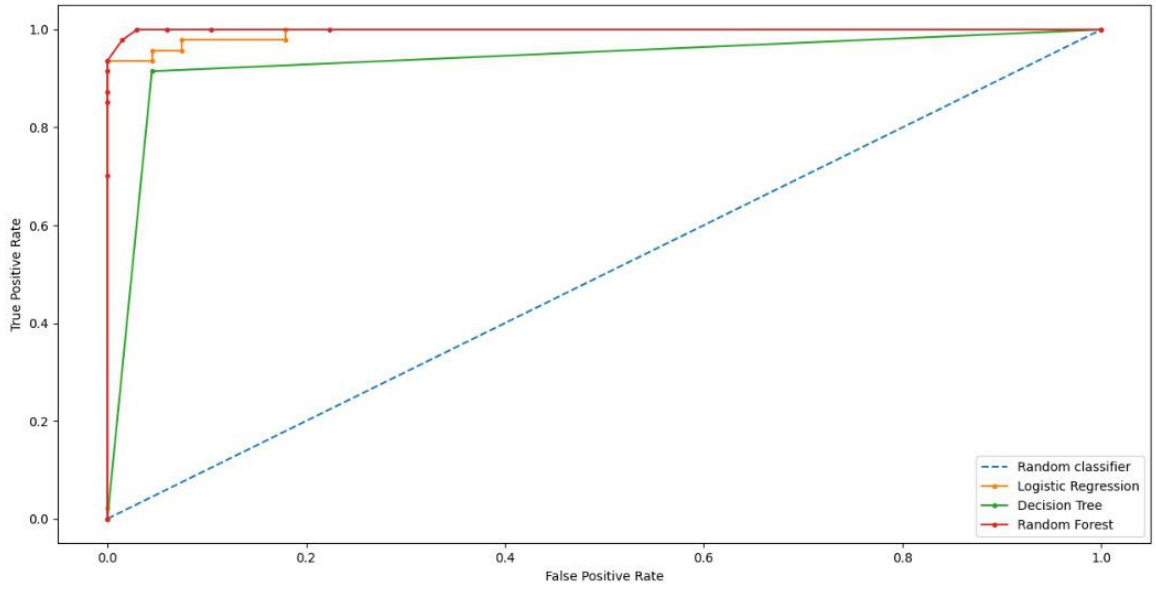

Figure 9.8 - Comparison of all the three algorithms and their classification performance on a receiver operating characteristic curve.

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly. If, however, the AUC had been 0, then the classifier would be predicting all Negatives as Positives, and all Positives as Negatives.When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly. If, however, the AUC had been 0, then the classifier would be predicting all Negatives as Positives, and all Positives as Negatives. When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives.When AUC=0.5, then the classifier is not able to distinguish between Positive and Negative class points. Meaning either the classifier is predicting random class or constant class for all the data points.

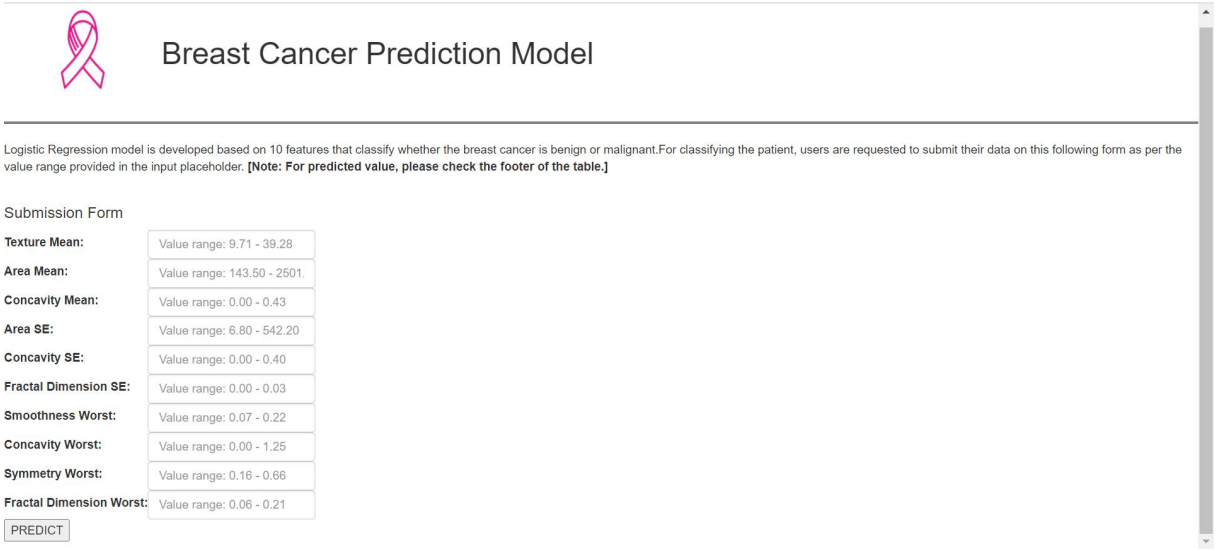

Figure 9.9 - Snapshot of the webpage interface for entering tumour biopsy parameter values which are then processed by Decision forest algorithm present in backend.

This website works by asking the user to input ten tumour parameter values which greatly influence the diagnosis process. These values are then passed to the pre-trained algorithm present in backend. The result is then returned in the form of a clear positive or negative diagnosis.

8. Conclusion

Cancers still remains responsible for the untimely death of millions of people each year. Out of this Breast cancer claims the number one reason behind the deaths of an unfortunate number of women each year globally. However the silver lining is the fact that breast cancer also presents itself as one of the more easily detectable cancers, thus giving physicians an opportunity to reduce the damage done by this dreadful disease but as the number of cases keep on increasing each year, the existing health care institutions feel choked and unable to cater to the ever rising number of eligible patients. By creating a machine learning based classification model we demonstrated that data obtained from patients by analyzing their tumour biopsy reports could be efficiently and accurately classified into data indicating a malignant tumour or a benign tumour. This was made possible when using logistic regression algorithms which consistently yielded accuracy of over 96% while classifying tumour data. Further it revealed an interesting correlation between various tumour parameters and how likely they were to indicate the presence of cancer.

Future research into this field should be to further increase the accuracy of such models either by training with more extensive data sets or by refining the classification algorithms themselves. A study on the parameters whose variations are highly correlated to the presence of cancer would be useful for both current machine learning research as well physicians conducting medical research.

Also introducing more doctors to this novel method of diagnosing this dreadful disease will be infinitely beneficial to all the parties involved here.

9. System Code Implementation

```
#Mini Project 2A
#Breast cancer prediction using machine learning
#Batch 2019-23

import numpy
import math
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report



df = pd.read_csv("data.csv")



print(df.isna().sum())

df = df.dropna(axis = 1)
print(df.shape)

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

sns.countplot(df['diagnosis'],label = 'count')
#plt.show()

LabelEncoder_y = LabelEncoder()    #convert M and B into 0 & 1
```

```
df.iloc[:,1] = LabelEncoder_y.fit_transform(df.iloc[:,1].values)
print(df.head())

sns.pairplot(df.iloc[:,1:5], hue = "diagnosis")    #plotting relation of attributes between M and
B
plt.show()



#getting the correlation
print(df.iloc[:,1:32].corr())



#visualise the correlation using heatmap
sns.heatmap(df.iloc[:,1:10].corr(), annot=True)
plt.show()



#split dataset into dependent and independent datasets
x = df.iloc[:,2:31].values
y = df.iloc[:,1].values



x_train,x_test,y_train,y_test = train_test_split(x,y,train_size=0.20,random_state=0)



#feature scaling
x_train = StandardScaler().fit_transform(x_train)
x_test= StandardScaler().fit_transform(x_test)



#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_test,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

# running the function
model = models(x_train,y_train)



#list of model accuracies
accu    = []

#testing accuracy of model
print("Model 1")
print(classification_report(y_test,model[0].predict(x_test)))
print('Accuracy : ',accuracy_score(y_test,model[0].predict(x_test))*100)
a = (accuracy_score(y_test,model[0].predict(x_test)))*100

print("Model 2")
print(classification_report(y_test,model[1].predict(x_test)))
```

```
print('Accuracy : ',accuracy_score(y_test,model[1].predict(x_test))*100)
b = (accuracy_score(y_test,model[0].predict(x_test)))*100

print("Model 3")
print(classification_report(y_test,model[2].predict(x_test)))
print('Accuracy : ',accuracy_score(y_test,model[2].predict(x_test)))
c = (accuracy_score(y_test,model[0].predict(x_test)))*100



#plotting accuracy of 3 models
x = ["logistic_regression","decision_tree","random_forest"]
y = [a,b,c]
low = min(y)
high = max(y)
plt.ylim(0,100,1)
plt.bar(x,y)
plt.yticks(numpy.arange(0, 101, 5))
plt.xlabel("prediction models")
plt.ylabel("% accuracy")
plt.title("Accuracy of models when predicting cancerous tumours from test dataset")

plt.show()
```

References

❖ Books

[1] Summerfield Mark,Programming in Python 3,Developer's library:2018,Pearson education

❖ Journal Publications

[1] Probabilistic Machine Learning for Healthcare Irene Y. Chen, Shalmali Joshi, Marzyeh Ghassemi, Rajesh Ranganath Annual Review of Biomedical Data Science 2021 4:1, 393-415

[2] Sultana, J.. (2018). Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers. International Journal of Engineering and Technology(UAE). 7. 10.14419/ijet.v7i4.20.22115.

[3] Cruz, Joseph A, and David S Wishart. "Applications of machine learning in cancer prediction and prognosis." Cancer informatics vol. 2 59-77. 11 Feb. 2014

[4] Serban A, Crisan-Vida M, Mada L, Stoicu-Tivadar L. User Interface Design in Medical Distributed Web Applications. Stud Health Technol Inform. 2016;223:223-9. PMID: 27139407.

❖ Websites

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675494/

[6]https://www.researchgate.net/publication/331233978_Predicting_Breast_Cancer_using_Logistic_Regression_and_Multi-Class_Classifiers

[7]https://www.breastcancerindia.net/statistics/latest_statistics_breast_cancer_india.html

[8]https://mainafoundation.org/breast-cancer-fast-facts/

[9]https://jamanetwork.com/journals/jamaoncology/fullarticle/2768634

[10]https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234