



Detection and Classification of Breast Cancer Using Artificial Intelligence

Anjalakshi M

Dept. of Embedded Systems Technology Technologies

Easwari Engineering College

Anjumanivannan3499@gmail.com

ABSTRACT:

The most frequent type of cancer among Indian women is breast cancer. In a scenario involving one of two women, there is a 50% probability of mortality. Breast cancer causes the second-highest rate of death in women, particularly in European nations. It happens when breast cells begin to develop malignant, (cancerous tumors). The chance of surviving this disease can be increased with an accurate and early diagnosis. This study involves two methods, the first method is by using deep learning algorithm where detection and classification of breast cancer is done for mammogram images. The second method involves the use of machine learning algorithm where prediction of cancer with numerical data is done. The mammogram images of breast are been collected and then preprocessing and filtration is done in order to classify the type of cancer. The two widely used machine learning algorithms are Random forest classifier and KNN (K-Nearest Neighbor) where numerical data are been studied and compared with both the algorithms. The data set are been collected from the Wisconsin Diagnose Breast Cancer data set for training and testing of numerical data process. Both the algorithms are been compared with important parameters like accuracy and precision and output with the higher value is been used for the prediction process.

Keywords: breast cancer, machine learning algorithm, deep learning, random forest, k-nearest neighbor.

I. INTRODUCTION

Human tissues are composed of cells, and tissues eventually give rise to organs. Every cell must carry out specific tasks, and when they are finished, they expire. Yet, there are situations when cells may not always die after performing because of internal and external problems, and new tissues can develop without the need for them. Tumors are formed by abnormal cell division or the growth of additional cells. Early in the course of the illness, the symptoms are not well-presented, which delays diagnosis. The National Breast Cancer Fund (NBCF) advises that women over the age of forty should obtain a mammography once a year. A mammography is a breast X-ray. The characteristics such as age, family history, and genetic risk are being a reason for this cancer. Over 97% of women live for more than five years with an early detection of these malignant cells. Also, during the past few decades, the number of deaths caused by this illness has considerably increased. The likelihood of developing this specific type of cancer is typically higher in metropolitan areas, although the pace of development appears to be increasing globally. By adopting suitable knowledge finding tools, these repositories will be making a greater contribution to present and future. Early detection and screening remain the only ways to improve the outcomes of breast cancer cases at this time.

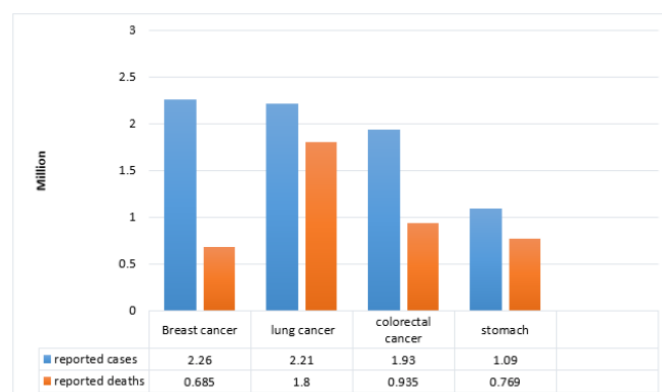


Figure 1: Reported cases of different types of cancer

Since serious breast cancer in women is caused by variances and unpredictability in breast cell tissues, research on the disease has expanded over the past ten years. Therefore, it is crucial to monitor the quantity of breast cancer-related fatalities prior to treatment. Cancerous and non-cancerous images are shown in Figure 2(a) and (b) as examples. Physicians can detect and treat breast cancer at an early stage thanks to therapeutic imaging, a non-invasive method of looking inside the body.

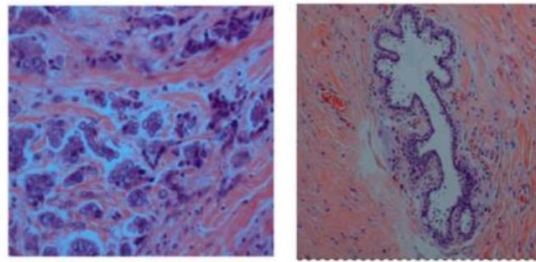


Figure 2: a. Cancerous breast images; and b. non-Cancerous breast images

Macrocalcifications and masses, which are frequent anomalies, are the root causes of breast cancer. The connective tissues and epithelia of the breast area develop microcalcifications and breast lumps. The tumour size and shape vary when they first appear in the breast. Depending on their intensity, these are categorised as malignant or benign. Benign breast lumps are non-invasive and non-cancerous, but they can grow and press against other organs, creating more problems. Malignant breast tumours are cancerous and aggressive. To prevent death, one should receive treatment at the early stages of diagnosis. While malignant tumours have irregular shapes, benign masses have restricted, smooth boundaries and are oval or circular in shape. The term "malignant breast masses" refers to lumps that are fuzzy, scratchy, or unclear. The malignant tumour also appears whiter than any nearby tissue.

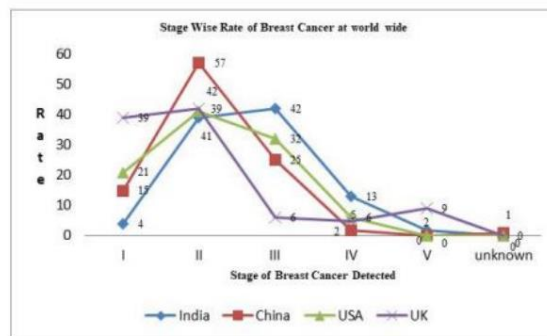


Figure 3: Stage wise breast cancer incidence detection rate

II. BACKGROUND

Both benign (not dangerous) and malignant (cancer, deadly) breast tumours exist. Most of the time, benign tumours are harmless. It does not spread to other body regions or organs. It unusually invades the nearby tissues and cells. It is typically not reversible and can be removed with the right treatment or surgery. Cancerous tumours pose a threat to life because they can invade nearby tissues and cells. They may spread to other body regions as well, which may result in death.

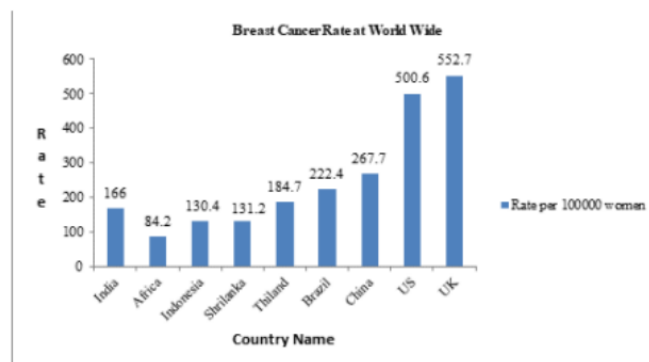


Figure 4: Crude incidence rate of women per 100000

III. MEDICAL IMAGES

Medical technology makes great use of digital image processing automation, but the biggest risk is that cancer mortality is rising. A dataset of medical images is needed to train the algorithm for cancer identification in order to improve early tumour diagnosis. By separating the image-based data into several qualities including texture, colour, and intensity, the suspected tissue images are segmented. Information like the location and extent of any ailment in the human body can be learned from medical images. Finding the precise position of pectoral tumour muscles and damaged tissues is helpful.

IV. TYPES OF MEDICAL IMAGES

To train the algorithms to identify the tumour, researchers employ a variety of medical images, including thermography, magnetic resonance imaging (MRI), mammography, X-ray, ultrasound images, and histopathological images. Thermography is a sophisticated and economical technology for detecting breast cancer that doesn't expose body cells to ionising radiation. Angiogenesis, edoema, nitric oxide vasodilatory phenomena, and oestrogen are all indications of cancer. Improved diagnosis and categorization of breast cancer are made possible by thermography [16]. Magnetic resonance imaging (MRI) is administered to patients who have a high risk of developing a tumour when other imaging methods are unable to reveal any abnormalities. Due to its high price, it is rarely used. Mammography is a method for tissue screening that is frequently used to identify tumours. Mammography used to be the gold standard for breast cancer screening, but it can be difficult to interpret because it can pick up on patients' malignancies and other small, inconspicuous characteristics.

V. MACHINE LEARNING ALGORITHM

A subset of artificial intelligence is machine learning (ML). that, in contrast to the conventional method of coding all conceivable combinations of data, imparts the capacity for learning into a system on the basis of a data set used for training outcomes in advance. There are numerous methods and procedures for developing learnable systems. Among them are decision trees, neural networks, and clustering.

VI. MACHINE LEARNING ALGORITHM TECHNIQUES:

Random Forest:

An approach for bagging that uses decision trees. It develops several trees that are used to categorise new entities according to their properties. Then, each tree designates a certain categorization and "votes" for that classification. The class with the most votes is then selected, and in the case of regression, the mean outputs from various trees are taken into account.

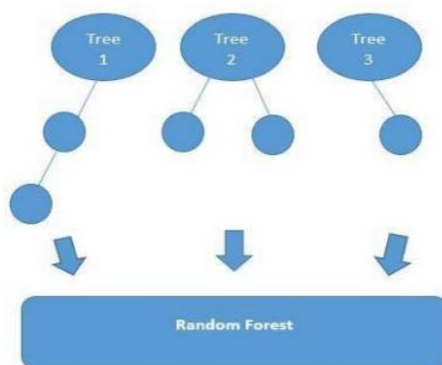


Figure 5: Random Forest works

K – Nearest Neighbour (KNN):

K can be viewed as a representation of the training data points that are near the test data points that will be used to determine the class. The k-nearest-neighbor algorithm is the process that finds a data set's group membership based on the neighbouring data sets. The classification and regression processes also use this technique, which makes use of supervised learning. KNN gathers all the nearby data points before processing a new data point. A high degree of variance in the qualities has a significant impact on the distance.

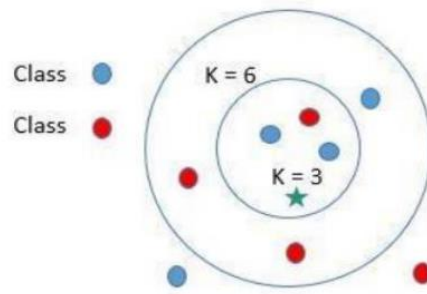


Figure 6: KNN illustrations

VII. DEEP LEARNING ALGORITHM

Deep learning is a type of machine learning and artificial intelligence that aims to imitate individuals and their actions based on specific aspects of how the human brain works to make wise decisions. It is a crucial component of data science to channel its modelling based on data-driven methods under statistical and predictive modelling. There must be some powerful forces that we often refer to as algorithms in order to drive such a human-like ability to adapt, learn, and perform accordingly.

DEEP LEARNING ALGORITHM TECHNIQUES

CONVOLUTIONAL NEURAL NETWORK:

Convolutional neural network, or CNN, is a popular artificial neural network used for object and image recognition and categorization. Therefore, in order to recognise objects in an image, Deep Learning employs a CNN. In order to solve problems with image processing, computer vision, and self-driving car obstacle detection, CNNs are widely utilised in a number of jobs and activities, such as speech recognition in natural language processing, video analysis, and video compression. CNNs are widely employed in deep learning because of their significant contributions to these fields, which are rapidly growing and changing.

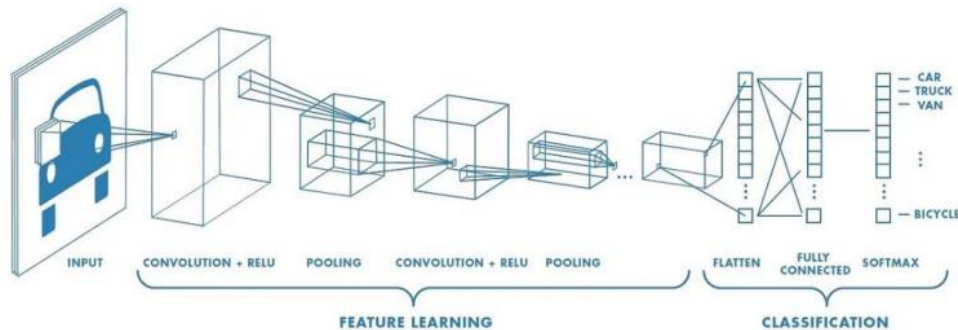


Figure 7: Convolutional neural network

VIII. PROPOSED SYSTEM

OVERALL FRAMEWORK

NUMERICAL PROCESS:

We divide the dataset into two sets where one for training and the other for testing—and classify the breast cancer using the data.

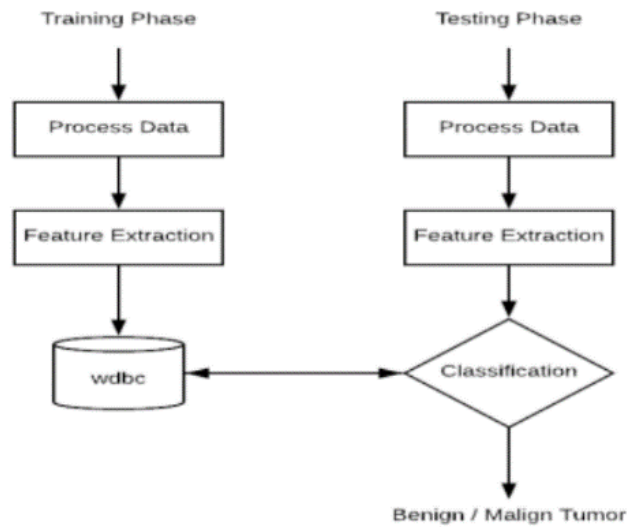


Figure 8: Proposed breast cancer detection model

TOOLS AND DATASET

Tools and libraries

The Python programming language was utilised in this work, together with libraries like Numpy, Pandas, Scikit-Learn, and Matplotlib, to develop machine learning methods. Google collaboratory is the tool used for these processes.

Dataset

Utilising data from the UCI repository, we are utilising the Wisconsin Breast Cancer Data Set There are a total of 569 samples in it out of which 357 are benign, and the remaining 212 are cancerous(malignant).

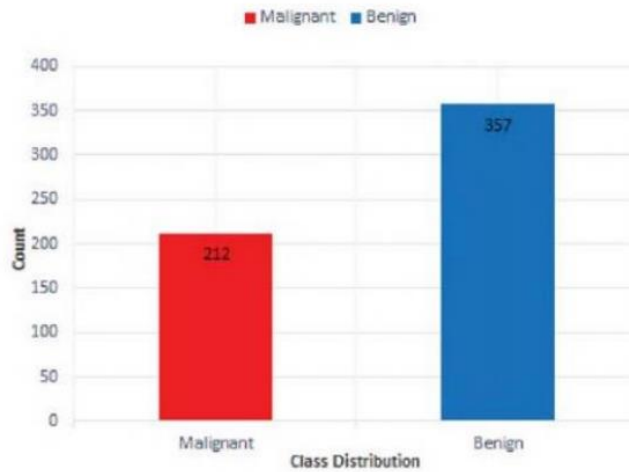


Figure 9: Dataset for breast cancer

FEATURES USED

Table 1: features

radius	The mean distance between the centre and points around the perimeter
texture	Deviation of standard values in grayscale
perimeter	The nuclear perimeter is the sum of the distances between the snake points.

area	the snake's internal pixel number plus one half of a pixel added to the perimeter
smoothness	Local variations in radius length can be found by calculating the distance between a radial line's length and the mean length of the lines surrounding it..
compactness	$\text{Perimeter}^2 / \text{area}$
concavity	the degree of the concave portions of the contour
Concave points	The ratio of the contour's concave areas.
symmetry	The length difference between parallel lines to the major axis and the cell boundary in both directions.
Fractal dimensions	An approximate coast. A higher number denotes a less regular shape and, hence, a greater possibility of malignancies.

IX. METHODOLOGY

This project aims to predict the breast cancer using machine learning algorithm. The various steps involved in the development of the model

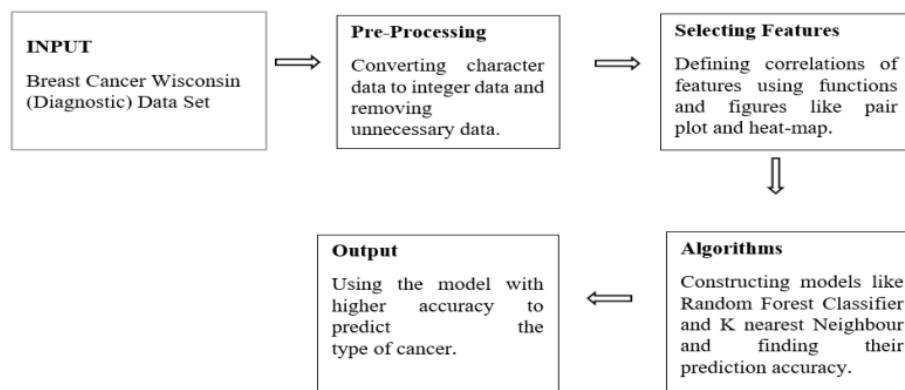


Figure 10: Block diagram of proposed model

STEP 1: DATA COLLECTIONS

The first step in predicting breast cancer is data collection. The final overall accuracy of the model is influenced by the accuracy of the training data. We'll use the dataset from the UCI Breast cancer Repository and perform some data cleaning (e.g., make the data usable by removing the empty columns). Google Colaboratory was used for implementing the machine learning algorithms in the Python programming language in this study along with the libraries such as Numpy, Pandas and Matplotlib.

STEP 2: SEPARATING THE DATA'S

The number of benign and malignant cells will be identified, and then a pair plot can be made with regard to each characteristic to show which feature influences the prediction the most. Using a heat map, the relationships can be seen more clearly.

STEP 3: MODEL TRAINING

Our target variable and predictor variable are chosen after studying the correlations. Afterward, divide the data into training set (75% of the total data) and testing set (25% of the total data).

STEP 4: ALGORITHMS FOR MODELLING

Then, a function is created that contains the K_Nearest Neighbour model and Random Forest classifier model. Then, accuracy of these two models will be evaluated, and the one that has the greatest accuracy in predicting cancer is used.

STEP 5: RESULTS AND OUTPUT

The final step is to predict the cancer kind and compare it to values in our testing dataset.

X. OUTPUT AND ANALYSIS

DATA PROCESSING:

The dataset's M (malignant) and B (benign) columns are converted to 1 and 0 respectively. The data splitting percentage is:

Data for training: 75% (426 records).

Data from the test: 25% (143 records).

HEAT MAP

The Heat Map provides information about the impact of other variables (columns) on the diagnosis column and improves our ability to see correlations.

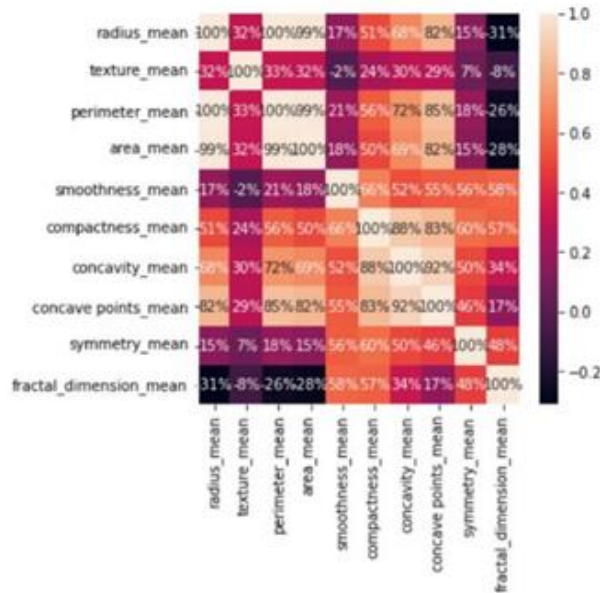
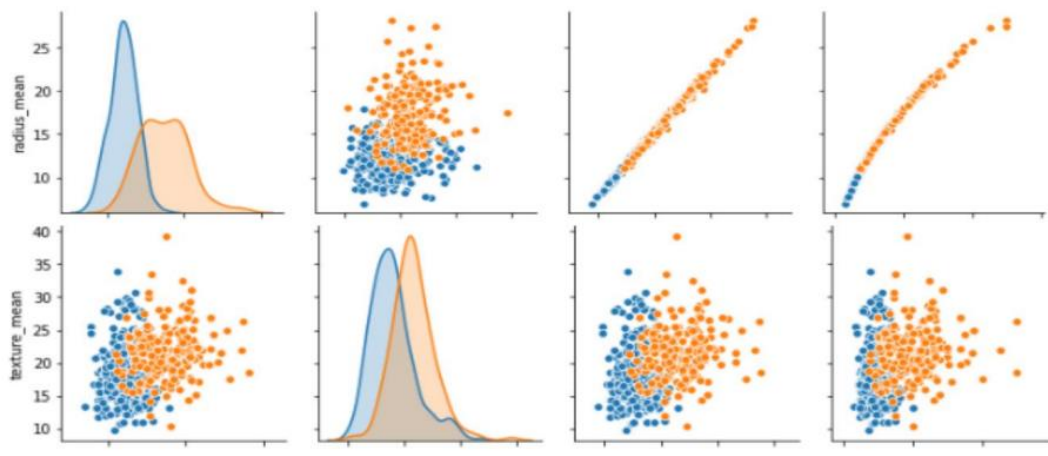


Figure 11: Heat map for feature correlation

COMPARED DIAGNOSIS PAIR PLOT WITH OTHER FEATURES

The pair plot illustrates visually how many cancerous and benign cells there are in relation to the features. This enables us to determine the features to examine when predicting outcomes. There are totally 569 records, out of which 357 are benign (non-cancerous), and 212 are malignant (cancerous). The following image compares the cancerous (Malignant) and non cancerous (Benign) cells in our data set.



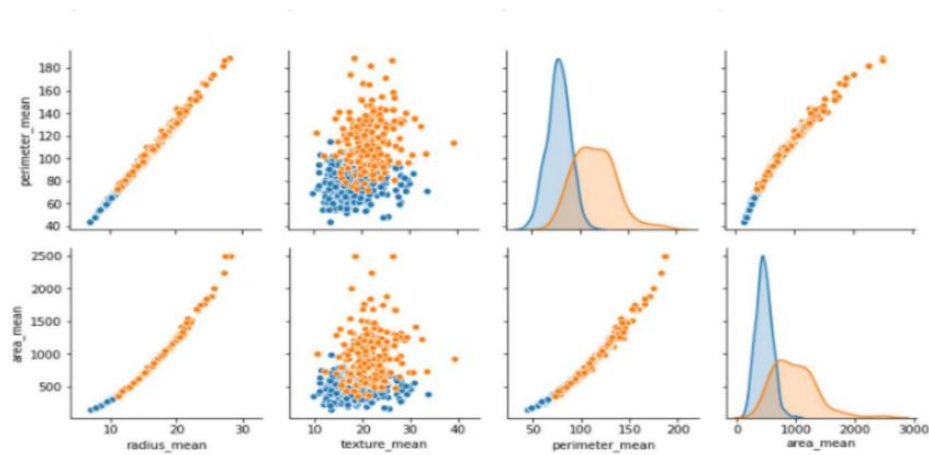


Figure 12: Pair plot diagram

ACCURACIES

The K-Nearest Neighbour and Random Forest Classifier are the two algorithms that are being compared because both were expected to produce highly accurate predictions. However, we needed one of the algorithms to have the highest accuracy possible so that our predictions would be more accurate. The image below illustrates how well both models predicted the data and helps in our choice of which model to apply.

```

from sklearn.ensemble import RandomForestClassifier
modell = RandomForestClassifier()
## fit the model to training set
modell.fit(X_train, Y_train)

RandomForestClassifier()

from sklearn.metrics import accuracy_score
acc_train = accuracy_score(Y_train, modell.predict(X_train))
print("Training Accuracy:", round(acc_train,2))

Training Accuracy: 1.0

acc_test = modell.score(X_test, Y_test)
print("Test Accuracy:", round(acc_test,2))

Test Accuracy: 0.96

```

Figure 13: Accuracy of Random Forest classifier


```

from sklearn.neighbors import KNeighborsClassifier
model2 = KNeighborsClassifier()
## fit the model to training set
model2.fit(X_train, Y_train)

KNeighborsClassifier()

acc_train = accuracy_score(Y_train, model2.predict(X_train))
print("Training Accuracy:", round(acc_train,2))

Training Accuracy: 0.98

acc_test = model1.score(X_test,Y_test)
print("Test Accuracy:", round(acc_test,2))

Test Accuracy: 0.96

```

Figure 14: K- Nearest Neighbors classifier

BIOIMAGING PROCESS

The data set which is been collected is been done various process for the classification of the breast cancer.

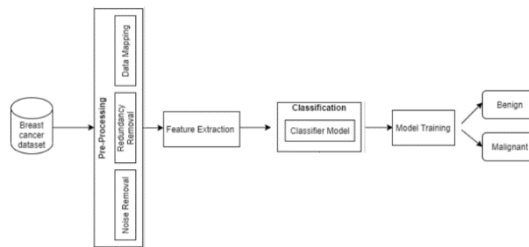


Figure 15: Proposed model for the bioimaging method

Tools And Dataset

TOOLS: MatLab was been used for the image processing method for the tumour identification in the breast cancer images obtained. Convolutional neural network for medical images is the tool which was used to identify tumour in mammographic images.

DATASET: The dataset is been collected from the well-known scan center -SARAVANA SCANS and the doctor referred is Dr.Saravanan. It has 600 mammographic images and these images are been used for the tumour identification and classification

MAMMOGRAPHIC IMAGE

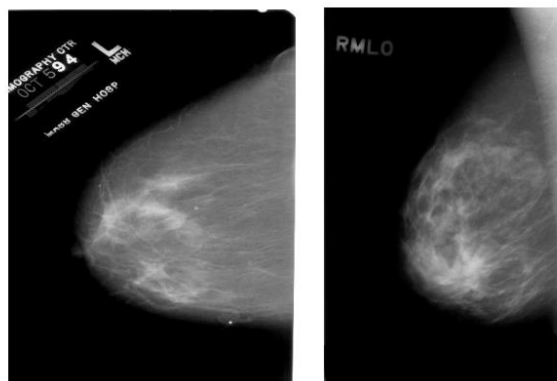
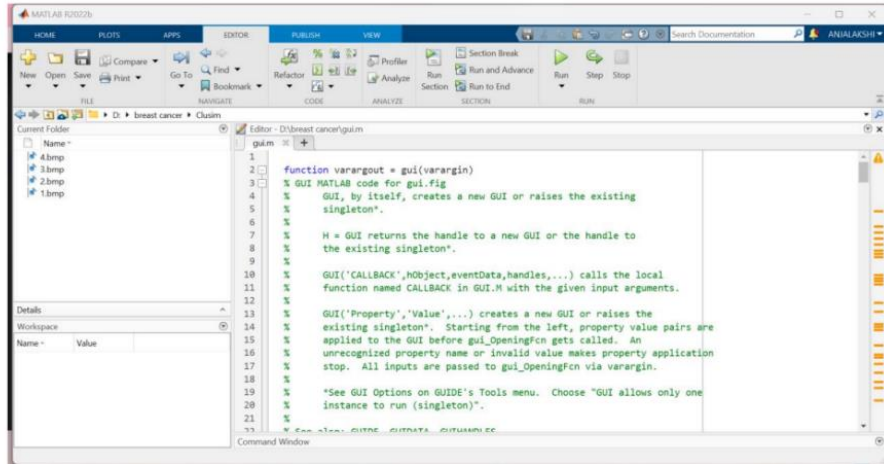


Figure 16: Mammographic medical images

Mammographic images are medical images that are used for tumor identification in humans. In this we are taking mammographic images of breast to detect the tumor and then with machine learning classification is been done.

FILTRATION AND PREPROCESSING

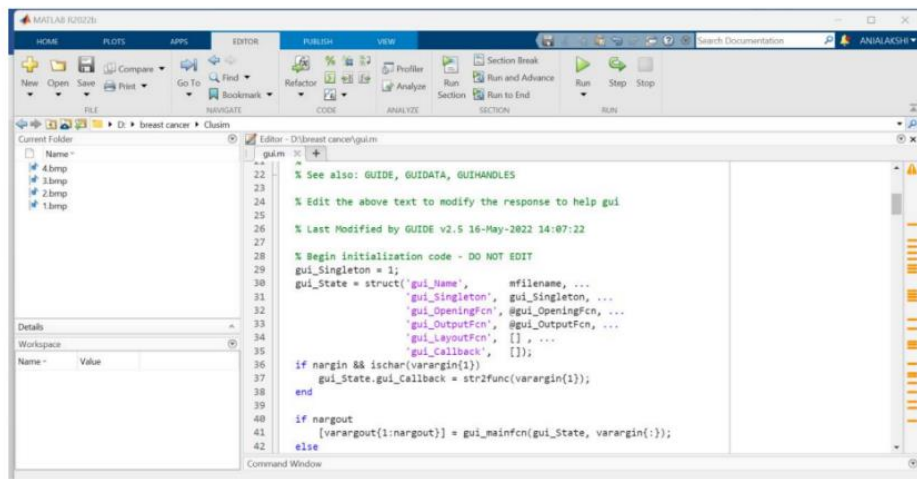
The image which is been collected should be preprocessed and filtered before getting the required output. So specific code is been applied to the data set and preprocessing of data is done.



```

1  function varargout = gui(varargin)
2
3  % GUI MATLAB code for gui.fig
4  % GUI, by itself, creates a new GUI or raises the existing
5  % singleton*.
6
7  % H = GUI returns the handle to a new GUI or the handle to
8  % the existing singleton*.
9
10 % GUI('CALLBACK',hObject,eventData,handles,...) calls the local
11 % function named CALLBACK in GUI.M with the given input arguments.
12
13 % GUI('Property','Value',...) creates a new GUI or raises the
14 % existing singleton*. Starting from the left, property value pairs are
15 % applied to the GUI before gui_OpeningFcn gets called. An
16 % unrecognized property name or invalid value makes property application
17 % stop. All inputs are passed to gui_OpeningFcn via varargin.
18 %
19 %*See GUI Options on GUIDE's Tools menu. Choose "GUI allows only one
20 % instance to run (singleton)".
21
22 % -----
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73

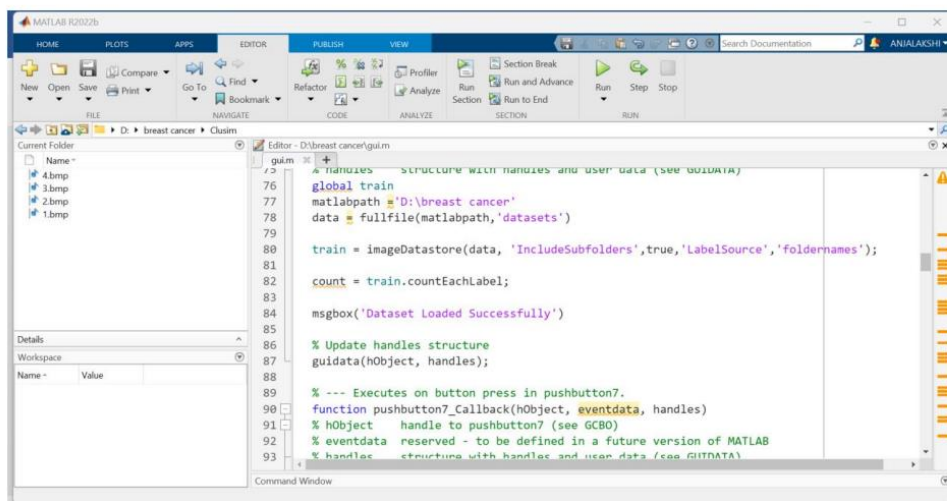
```



```

22 % See also: GUIDE, GUIDATA, GUIHANDLES
23
24 % Edit the above text to modify the response to help gui
25
26 % Last Modified by GUIDE v2.5 16-May-2022 14:07:22
27
28 % Begin initialization code - DO NOT EDIT
29 gui_Singleton = 1;
30 gui_State = struct('gui_Name',       mfilename, ...
31                  'gui_Singleton',   gui_Singleton, ...
32                  'gui_OpeningFcn', @gui_OpeningFcn, ...
33                  'gui_OutputFcn',  @gui_OutputFcn, ...
34                  'gui_LayoutFcn',  [], ...
35                  'gui_Callback',    []);
36
37 if nargin && ischar(varargin{1})
38     gui_State.gui_Callback = str2func(varargin{1});
39 end
40
41 if nargin
42     [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
43 else
44     varargout{1:nargout} = gui_mainfcn(gui_State, varargin{:});
45 end
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73

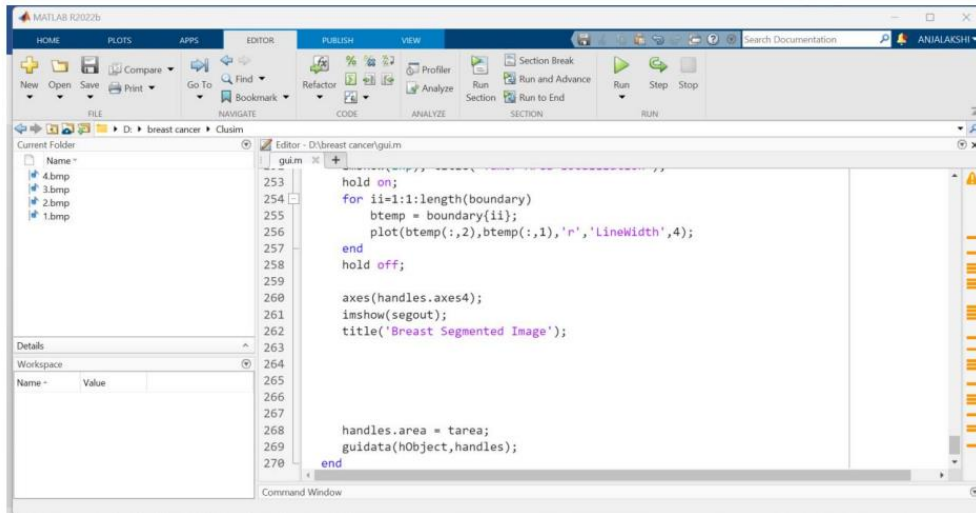
```



```

76 % -----
77 % Analyzes STRUCTURE WITH HANDLES and user data (see GUIDATA)
78 global train
79 matlabpath = 'D:\breast cancer'
80 data = fullfile(matlabpath,'datasets')
81
82 train = imageDatastore(data, 'IncludeSubfolders',true,'LabelSource','foldernames');
83
84 count = train.countEachLabel;
85
86 msgbox('Dataset Loaded Successfully')
87
88 % Update handles structure
89 guidata(hObject, handles);
90
91 % --- Executes on button press in pushbutton7.
92 function pushbutton7_Callback(hObject, eventdata, handles)
93 % hObject handle to pushbutton7 (see GCBO)
94 % eventdata reserved - to be defined in a future version of MATLAB
95 % handles structure with handles and user data (see GUIDATA)

```



XI. RESULT

NUMERICAL DATA OUTPUT

The comparison between the actual test value and the value predicted by our model is given to show whether the model works as expected.

```

# prediction of random-forest
pred=model1.predict(X_test)
print('Predicted values:')
print(pred)
print('Actual values:')
print(Y_test)
    
```

Predicted values:
 [1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 0 1 0 0 1 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 0
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0]

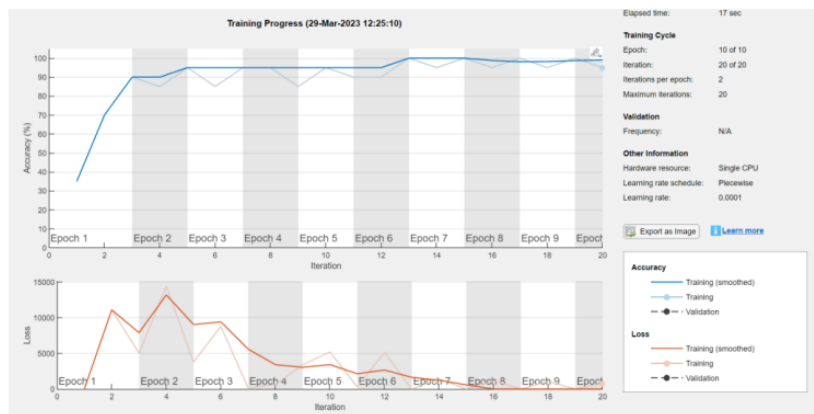
Actual values:
 [1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0]

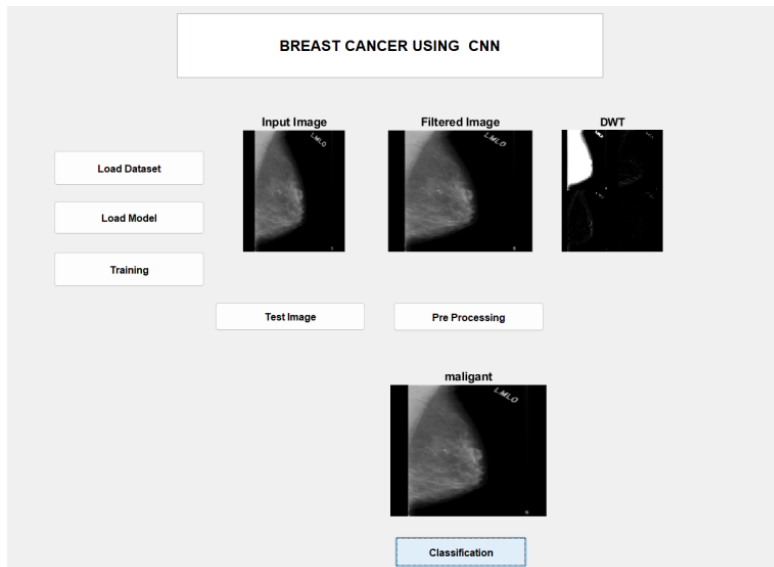
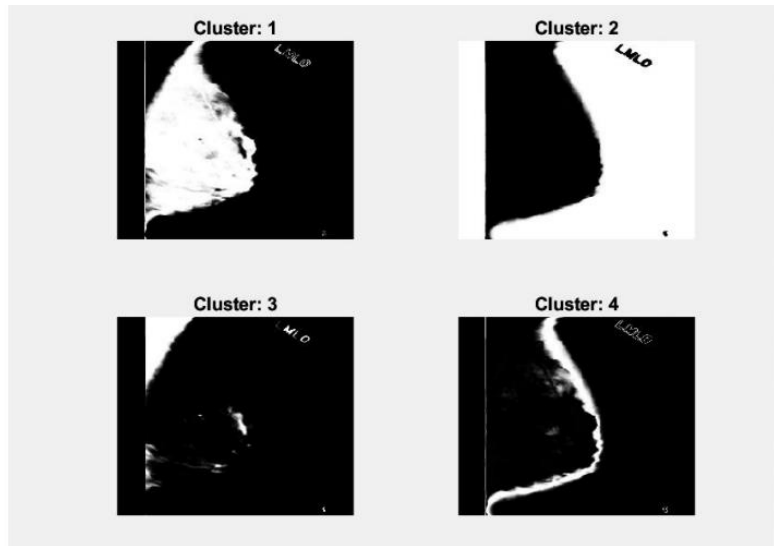
Figure 17: verified output of random forest

BIOIMAGE OUTPUT:

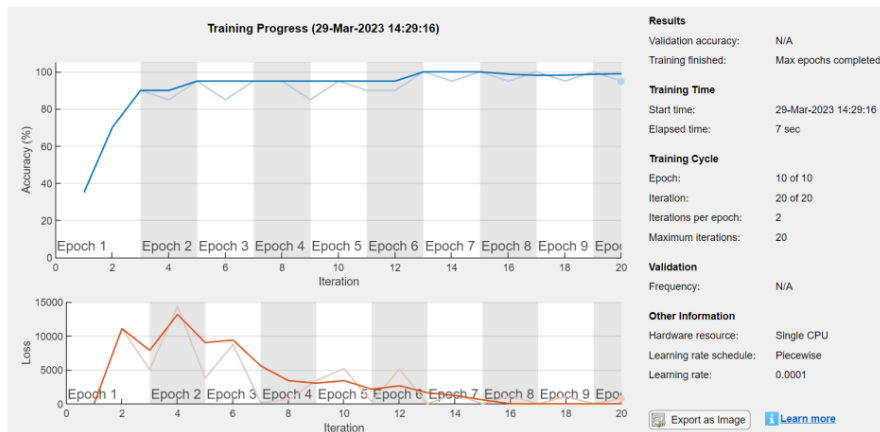
The output for the tumor identification is obtained and classification of the cancer is done and thus the predicted output is verified. Tumour size depends on the persons stages

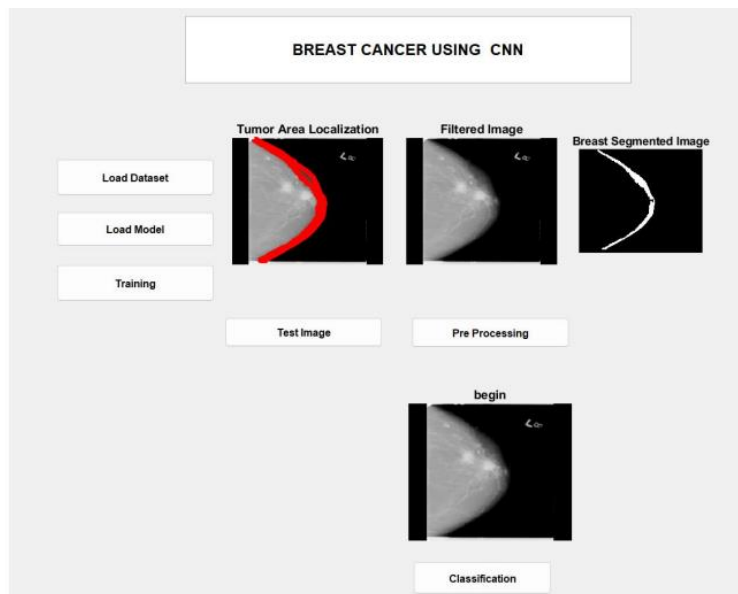
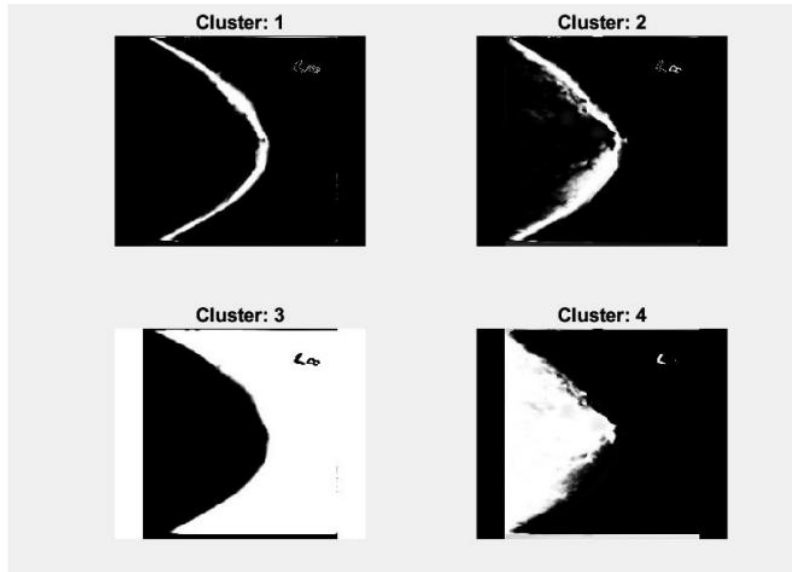
SOFTWARE OUTPUT FOR MALIGNANT



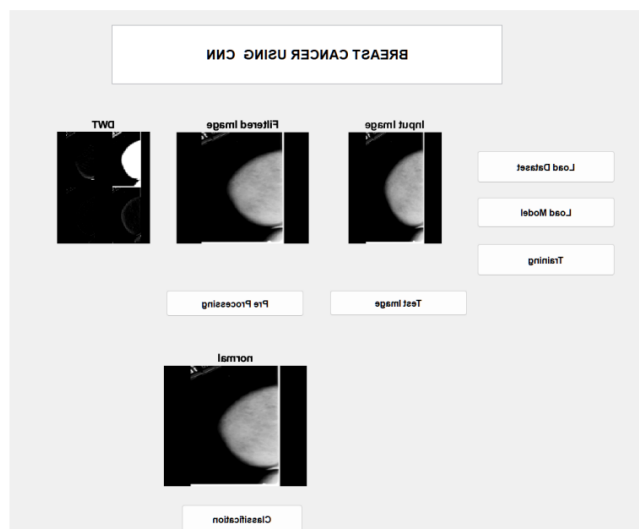


SOFTWARE OUTPUT FOR BENIGN





SOFTWARE OUTPUT FOR NORMAL



XII. CONCLUSION

Two machine learning algorithms— K-Nearest Neighbor(KNN) and Random Forest Classifier are used to identify breast cancer using numerical data. To determine which method is more suitable, the accuracy of each is compared with one another. The best algorithm for prediction is the Random Forest Classifier, which on the "Breast Cancer Wisconsin (Diagnostic) Data Set" has a pinpoint prediction accuracy. Therefore, using our Random Forest Classifier method and the features from this dataset, breast cancer may be predicted with virtually perfect accuracy. For the bioimaging based breast cancer, the tumor is been identified and the classification is done using the given dataset. Thus, for both the dataset the output is obtained and verified.

XIII. FUTURE SCOPE

Thus, by using machine learning algorithm and deep learning algorithm one can easily identify the stages of breast cancer and the tumor is been identified and classified. Future depends on AI and so for medical field they are various methods to predict and classify each disease. One such method is the breast cancer detection with Artificial intelligence.