7/2/2023

MAJOR PROJECT

BREAST CANCER DETECTION

AKNOWLEDGEMENT

We would like to thank all of the people who helped us with this project, without their support and guidance it wouldn’t have been possible. We appreciate **MR Akash Maurya** for his guidance and supervision which has provided a lot of resources needed in completing our project.

Our parents as well as friends were constantly encouraging us throughout the process when we felt discouraged or became frustrated because they knew how much work went into this venture so that is why we want to extend them thanks too!

We are grateful to our team partners in developing the project, for their willingness and assistance. They helped us with this project, which we appreciate dearly.

At last , I would like to express my gratitude to the Skill Vertex education company for this amazing course and resources provided to build our skillset and successful career .

ABSTRACT

Breast cancer is one of the major causes of death in women and it is difficult to prevent breast cancer as the main reasons underlying breast cancer remain unknown. Characteristics of breast cancer, such as masses and microcalcifications visible in mammograms, can be employed for early diagnosis and hence are highly beneficial for women who may be at risk of developing malignant tumors. The principal check used for screening and early diagnosis is X-ray mammography and the proper interpretation of the clinical report is vital for breast cancer prediction, but the decision may be prone to error. Mammograms are difficult to interpret, especially in the screening context. The sensitivity of screening mammography is affected by image quality and the radiologist's level of expertise

It is very necessary to detect cancer at early stages. There are various Machine Learning techniques available for the purpose of diagnosis of breast cancer data. This paper presents a Machine Learning model to perform automated diagnosis for breast cancer. This method employed CNN as a classifier model and Recursive Feature Elimination (RFE) for feature selection. Also, five algorithms SVM, Random Forest, KNN, Logistic Regression, Naïve Bayes classifier have been compared in the paper. The system was experimented on BreaKHis 400X Dataset. The performance of the system is measured on the basis of accuracy and precision.

INTRODUCTION

Breast cancer is major cause of death in women around the world. According to WHO (World Health Organisation), breast cancer accounted for maximum deaths (2.26 million cases), worldwide in 2020 out of the 10 million cases of cancer. Breast cancer starts when cells in the breast begin to grow out of control. These accumulations of cells are called tumours and they 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 making them prone to cancer. side effects of Breast Cancer are – Fatigue, Headaches, Pain and numbness (peripheral neuropathy), Bone loss and osteoporosis. There are two types of tumours. One is benign which is non-cancerous and the other one is malignant which is cancerous. Benign breast tumours are abnormal growths in the breast, but they do not spread outside. So, this means that they are not life threatening, but some types of benign tumours can increase a woman's risk of getting breast cancer. Different imaging tests are used for detecting breast cancer. Some of them are mammograms, breast ultrasound and breast MRI. Detection of breast cancer in its early stages using image processing techniques includes four parts. In the first part the digital images (mammograms) are pre-processed to remove any kind noise. Then in the second part the images undergo the segmentation process to enhance the tumor part. After this, in the third part, the important features in the segmented images are extracted. Finally, in the fourth part, with the help of the extracted features, the images are classified into normal, benign or malignant. Here, ‘normal’ represents the breast with no tumor, ‘benign’ represents the breast with non-cancerous tumor and ‘malignant’ represents breast with cancerous tumor.

BREAST CANCER

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 tumour that can often be seen on an x-ray or felt as a lump. Breast cancer occurs almost entirely in women, but men can get breast cancer, too. It’s important to understand that most breast lumps are benign and not cancer (malignant). Non-cancerous breast tumours are abnormal growths, but they do notspread outside of the breast. They are not life threatening, but some types of benignbreast lumps can increase a woman's risk of getting breast cancer. Any breast lump orchange needs to be checked by a health care professional to determine if it is benign ormalignant (cancer) and if it might affect your future cancer risk.

WHERE BREAST CANCER STARTS

Breast cancers can start from different parts of the breast.

- Most breast cancers begin in the ducts that carry milk to the nipple (ductal cancers)

- Some start in the glands that make breast milk (lobular cancers)

- There are also other types of breast cancer that are less common like phyllodes tumour and angiosarcoma

- A small number of cancers start in other tissues in the breast. These cancers are called sarcomas and lymphomas and are not really thought of as breast cancers.

Although many types of breast cancer can cause a lump in the breast, not all do.Many breast cancers are also found on screening mammograms, which can detect cancers at an earlier stage, often before they can be felt, and before symptoms develop.

TYPES OF BREAST CANCER

There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumours and angiosarcoma are less common.

Once a biopsy is done, breast cancer cells are tested for proteins called estrogen receptors, progesterone receptors and HER2. The tumour cells are also closely looked at in the lab to find out what grade it is. The specific proteins found and the tumour grade can help decide treatment options.

BREAST CANCER SIGNS AND SYMPTOMS

Knowing how your breasts normally look and feel is an important part of breast health. Although having regular screening tests for breast cancer is important, mammograms do not find every breast cancer. This means it's also important for you to be aware of changes in your breasts and to know the signs and symptoms of breast cancer. The most common symptom of breast cancer is a new lump or mass. A painless, hard mass that has irregular edges is more likely to be cancer, but breast cancers can be tender, soft, or round. They can even be painful. For this reason, it's important to have any new breast mass, lump, or breast change checked by an experienced health care professional.

Other possible symptoms of breast cancer include:

• Swelling of all or part of a breast (even if no lump is felt)

• Skin dimpling (sometimes looking like an orange peel)

• Breast or nipple pain

• Nipple retraction (turning inward)

• Nipple or breast skin that is red, dry, flaking or thickened

• Nipple discharge (other than breast milk)

• Swollen lymph nodes (Sometimes a breast cancer can spread to lymph nodes under the arm or around the collar bone and cause a lump or swelling there, even before the original tumor in the breast is large enough to be felt).

Although any of these symptoms can be caused by things other than breast cancer, if you have them, they should be reported to a health care professional so the cause can be found. Remember that knowing what to look for does not take the place of having regular mammograms1 and other screening tests2. Screening tests can help find breast cancer early, before any symptoms appear. Finding breast cancer early gives you a better chance of successful treatment.

COMMON TESTS FOR BREAST CANCER

MAMMOGRAMS

Mammograms are low-dose x-rays of the breast. Regular mammograms can help find breast cancer at an early stage, when treatment is most successful. A mammogram can often find breast changes that could be cancer years before physical symptoms develop. Results from many decades of research clearly show that women who have regular mammograms are more likely to have breast cancer found early, are less likely to need aggressive treatment like surgery to remove the breast (mastectomy) and chemotherapy, and are more likely to be cured.

BREAST MRI

Breast MRI (magnetic resonance imaging) uses radio waves and strong magnets to make detailed pictures of the inside of the breast.

BREAST ULTRASOUND

Breast ultrasound uses sound waves to make a computer picture of the inside of the breast. It can show certain breast changes, like fluid-filled cysts, that are harder to identify on mammograms.

BREAST BIOPSY

When other tests show that there is breast cancer, then biopsy is done. Needing a breast biopsy doesn't necessarily mean that there is cancer. Most biopsy results are not cancer, but a biopsy is the only way to find out for sure. During a biopsy, a doctor will remove small pieces from the suspicious area so they can be looked at in the lab to see if they contain cancer cells.

PROJECT OUTLINE

METHODOLOGY

1.**Dataset:** Attribute Information:

- ID number
- Diagnosis (M = malignant, B = benign)

  Ten real-valued features are computed for each cell nucleus:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

2.Data collection

After collecting data, we need to know what are the shape of this dataset, Here we have attribute(property) called data.shape.

3.Analysis of data work

- data information
- graphical representation
- dependent and independent data
- missing values

4.**Data Modelling**

- different algorithms
- confusion matrices
- F1 score
- Accuracy and prediction

ALGORITHM

Logistic Regression

In linear regression, the linear regression hyperplane that is obtained cannot be used to predict the dependent variable by using the independent variable. Hence, when there is categorical data, logistic regression is used. Logistic Regression predicts whether something is true or false instead of predicting something continuous. It is used for classification. The sigmoid function is used to convert the independent variable into an expression of probability which ranges from 0 and 1 concerning the dependent variable. The ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular Machine Learning algorithm. A drawback of Logistic Regression is the assumption of linearity between the dependent and independent variables.

Random Forest

Random forest is a supervised learning algorithm. It is a collection of Decision Trees. Decision Tree is hierarchical in nature in which nodes represent certain conditions on a particular set of features, and branches split the decision towards the leaf nodes. Leaf determine the class labels. Decision Tree can be constructed either by using Recursive Partitioning or by Conditional Inference Tree. Recursive Partitioning is the step-by-step process by which a Decision Tree constructed by either splitting or not splitting each node. We can say that the tree is learned by splitting the source set into subsets based on an attribute value test. The recursion is completed when the subset at a node has all the same value of the target variable. Conditional Inference Tree is a statistical based approach that uses non parametric tests as splitting criteria that is corrected for multiple testing to avoid over fitting. Random Forest is suitable for high dimensional data modeling as it can handle missing values, continuous, categorical and binary data but for very data sets, the size of the trees can take up a lot of memory. It can tend to over-fit, so there is a need to tune the hyper-parameters.

DECISION TREE

- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes**

represent the features of a dataset, branches represent the decision rules and **each leaf node represents the outcome.**

- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Analysis of the classifiers

In this project I have used three main classifiers to analyse which algorithms has most potential to give precise prediction.

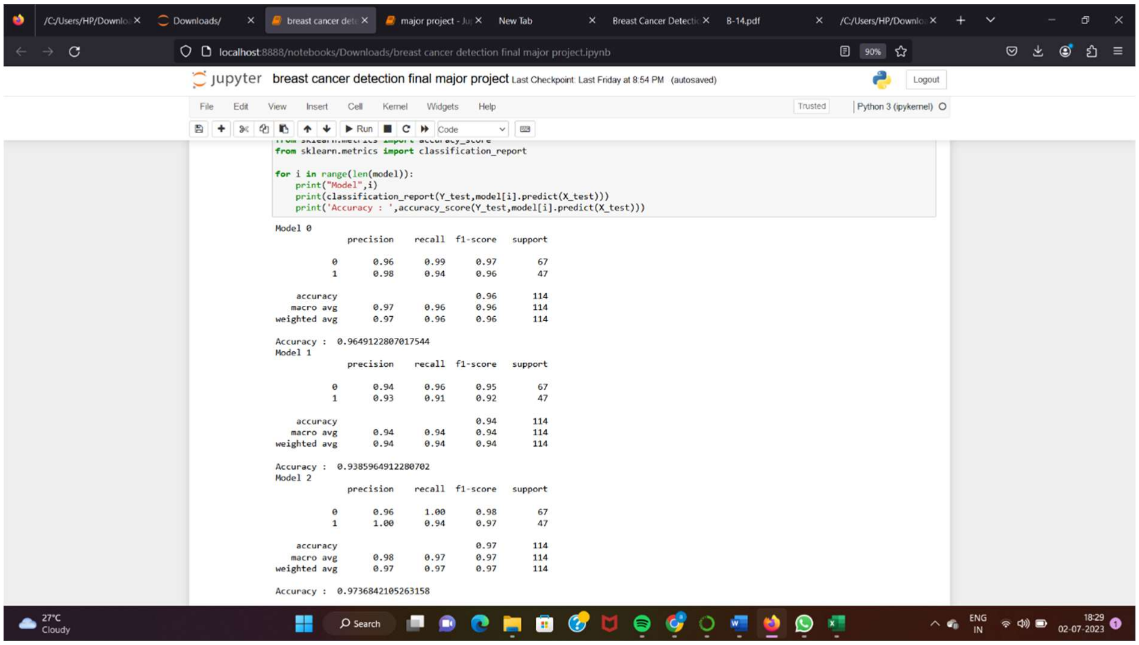
Main classifier used are:

- Logistic regression
- Decision tree
- Random forest

From Confusion Matrix ,we can retrieve the accurate score for the precision,F1 score ,recall support.

The accuracy of the used classifiers as follow:

- logistic regression accuracy: 0.9912087912087912
- Decision tree accuracy: 1.0
- Random forest accuracy: 0.9978021978021978



```
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report

for i in range(len(model)):
    print("Model",i)
    print(classification_report(Y_test,model[i].predict(X_test)))
    print('Accuracy : ',accuracy_score(Y_test,model[i].predict(X_test)))
```

```
Model 0
              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

Accuracy :  0.9649122807017544
Model 1
              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

Accuracy :  0.9385964912280702
Model 2
              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

Accuracy :  0.9736842105263158
```

In this above image,

Model 0 represents logistic regression

Model 1 represents decision tree

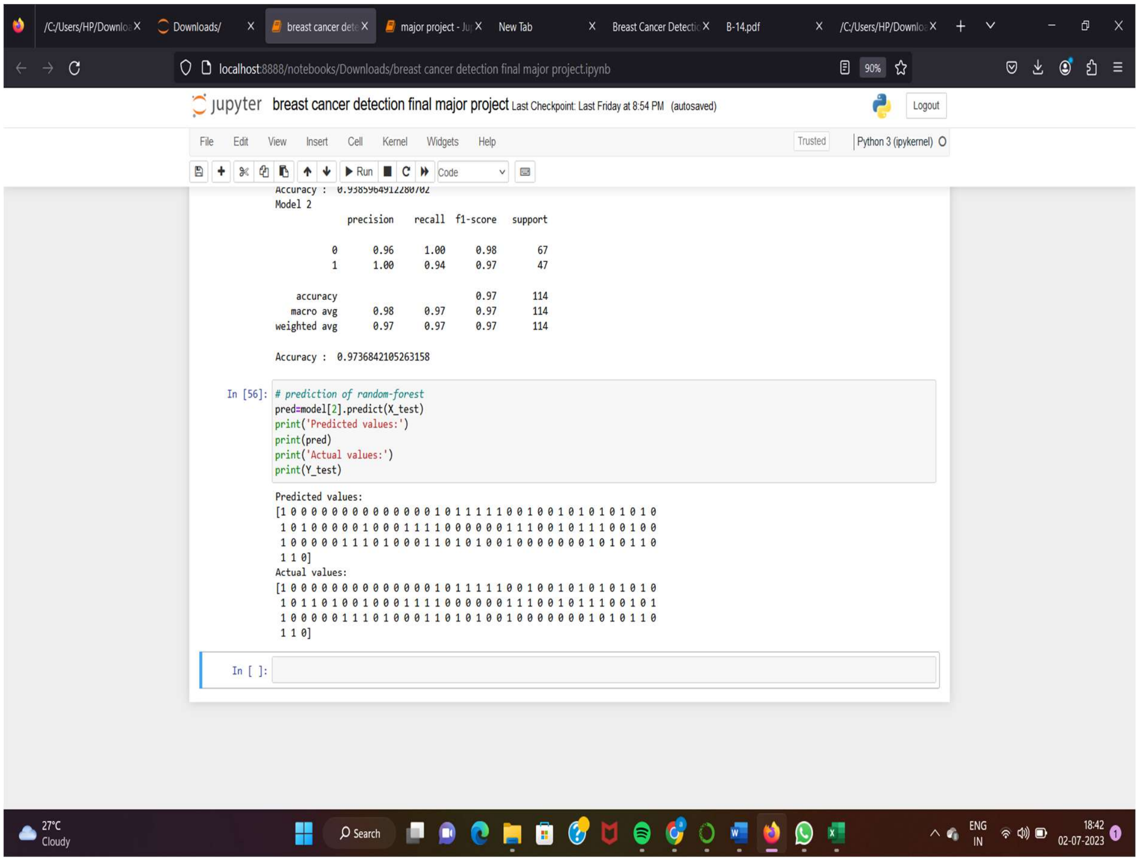
Model 2 represents Random forest

From this we clearly can analyse that the Random forest classifier is more accurate than the other two which have been tested.

Thus we pick Random forest classifier as model to be used for precise result for the breast cancer prediction.

RESULT

As the model used does not give 100% percentage true result. we can only make the most accurate prediction using this model .thus, we can predict the values with reference to actual values.



```
Accuracy :  0.9385964912280702
Model 2
              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

Accuracy :  0.9736842105263158
```

```
In [56]: # prediction of random-forest
         pred=model[2].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 0 0 1 0 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 0 0 0 1 0 0 0 1 1 1 1 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 1 0 1 0 0 1 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 0 0 1 0 1 1 1 1 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 1 1 1 1 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 1 0 1 0 1 1 0
 1 1 0]
```

```
In [ ]:
```

From this above image, we can see that most of the predicted values match with actual values also some of the prediction has gone wrong ,which maybe an issue as it is the matter of human life .but there is no model which can be deployed for 100% yield thus Random forest classifier is far better than the other.

FUTURE SCOPE IN BUSINESS

FROM RESEARCH

In collaboration with colleagues at DeepMind, Cancer Research UK Imperial Centre, Northwestern University and Royal Surrey County Hospital, we set out to see if artificial intelligence could support radiologists to spot the signs of breast cancer more accurately.

The model was trained and tuned on a representative data set comprised of de-identified mammograms from more than 76,000 women in the U.K. and more than 15,000 women in the U.S., to see if it could learn to spot signs of breast cancer in the scans. The model was then evaluated on a separate de-identified data set of more than 25,000 women in the U.K. and over 3,000 women in the U.S. In this evaluation, our system produced a 5.7 percent reduction of false positives in the U.S, and a 1.2 percent reduction in the U.K. It produced a 9.4 percent reduction in false negatives in the U.S., and a 2.7 percent reduction in the U.K.

We also wanted to see if the model could generalize to other healthcare systems. To do this, we trained the model only on the data from the women in the U.K. and then evaluated it on the data set from women in the U.S. In this separate experiment, there was a 3.5 percent reduction in false positives and an 8.1 percent reduction in false negatives, showing the model's potential to generalize to new clinical settings while still performing at a higher level than experts.

Next steps

looking forward to future applications, there are some promising signs that the model could potentially increase the accuracy and efficiency of screening programs, as well as reduce wait times and stress for patients. Google's Chief Financial Officer Ruth Porat shared her optimism around potential technological breakthroughs in this area in a post in October reflecting on her personal experience with breast cancer. But getting there will require continued research, prospective clinical studies and regulatory approval to understand and prove how software systems inspired by this research could improve patient care.

This work is the latest strand of our research looking into detection and diagnosis of breast cancer, not just within the scope of radiology, but also pathology. In 2017, we published early findings showing how our models can accurately detect metastatic breast cancer from lymph node specimens. Last year, we also developed a deep learning algorithm that could help doctors spot breast cancer more quickly and accurately in pathology slides.

We're looking forward to working with our partners in the coming years to translate our machine learning research into tools that benefit clinicians and patients.

REFERENCE

- https://www.kaggle.com/code/vikasukani/breast-cancer-prediction-using-machine-learning/notebook
- https://blog.google/technology/health/improving-breast-cancer-screening