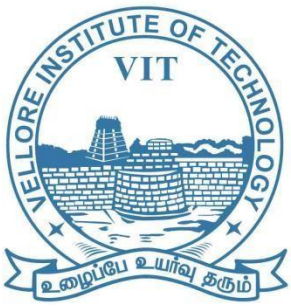


VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Electrical Engineering

Special Term 5_ 2021-2022 Slot : TCC1+TCC2

Technical Answers for Real World Problems

J-Component Project Report on

Breast Cancer Prediction

Submitted by :

Divyansh Singh Raghav : 19BEE0067

August 2022

Under the Guidance of

JACOB RAGLEND I

Professor Grade 2, SELECT, VIT-Vellore.

CONTENTS

S.No	Topic	Page Number
1	Introduction	3
2	Literature Survey	3
3	Objective	4
4	Methodology	4
5	Work Plan	7
6	Results	8
7	Expected Outcomes	9
8	References	10

Introduction

One of the most prevalent malignant tumours of women in the world is breast cancer, which typically affects older women. However, in recent years, younger women are now developing breast cancer. As is well known, postmenopausal women have been the subject of less research on breast cancer, and more research is still needed to fully understand its characteristics .The World Health Organization disclosed that more than 620,000 women died from breast cancer in the world in 2018 alone, which represents approximately 15% of all female cancer deaths. Thus, breast cancer diagnosis presents one of the main challenges that need to get timely treatments. In this context, multiple image modalities, namely mammography, echography and magnetic resonance Imaging (MRI) are used for breast tumour diagnosis. One of the main treatments of this pathology is chemotherapy. However, several secondary effects can occur due this treatment, and cancer can not respond to it.

Literature Survey

[1] Amin ,Abdus et.al. dealt with usage of ML algorithms, they suggested a novel breast cancer detection approach in the clinical data. In the proposed method, supervised and unsupervised techniques were utilized to select related features from a data set, and SVM and K fold validation is utilized.

[2] Hein, Umadevi demonstrates the methodology reliability by combining modern segmentation procedures with machine learning. The proposed method is useful for distinguishing between types of tumours. The simulated findings are reviewed to establish the method's suitability for early breast cancer detection

[3] Azour , Azzedine suggested classifying existing deep learning research on mammography types according on the approaches used by researchers in their empirical studies. Assessmentt efficacy of this approach for the early detection of breast cancer has proved to be high.

[4] Irum , Ayaz worked on identification and categorization of breast cancer on histopathology pictures using an unique patch-based deep learning approach. Unsupervised training and supervised tuning are used to extract features

[5] Jing, Denan dealt with the choosing and extraction of the image data.Then they deployed the model by investigating the use of CNN-based learning method to characterise breast masses for various diagnostic, predictive, or prognostic tasks in a variety of imaging modalities.

[6] Bolei, Jon researchers presented a deep selective attention strategy for selecting valuable areas in original pictures for categorization. In this technique, a decision network is created to determine where to crop and if the cropped patch is required for categorization. The patches are trained further using co-evolution training technique.

[7] Lei, Zhang evaluated two machine learning algorithms for automatically classifying breast cancer data into benign and malignant sub-classes. The first method relies on the extraction of a set. two coding models encode handmade characteristics and trained using support vector machines, whereas the second method is based on the architecture of convolutional neural networks. Networks of neurons

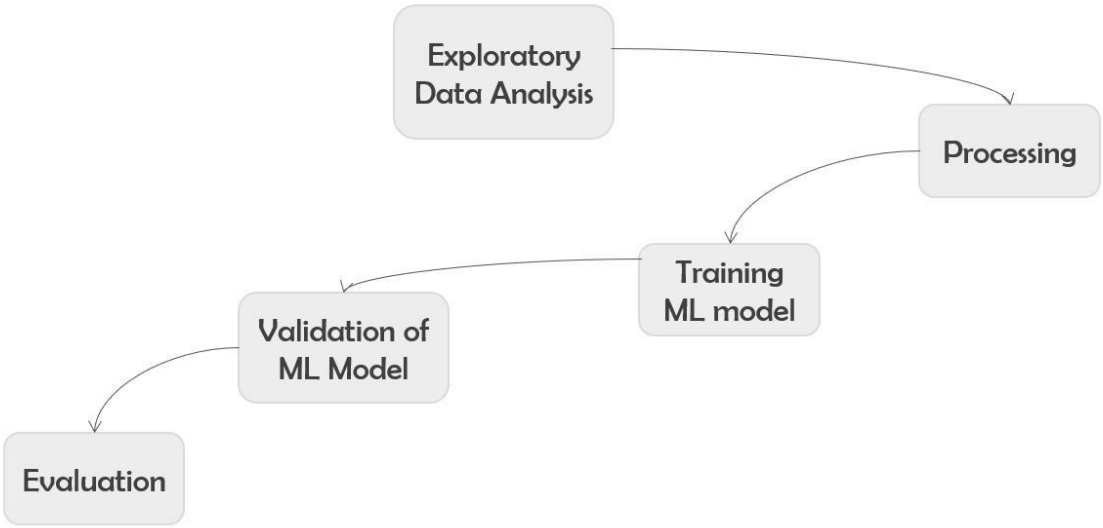
Objectives :

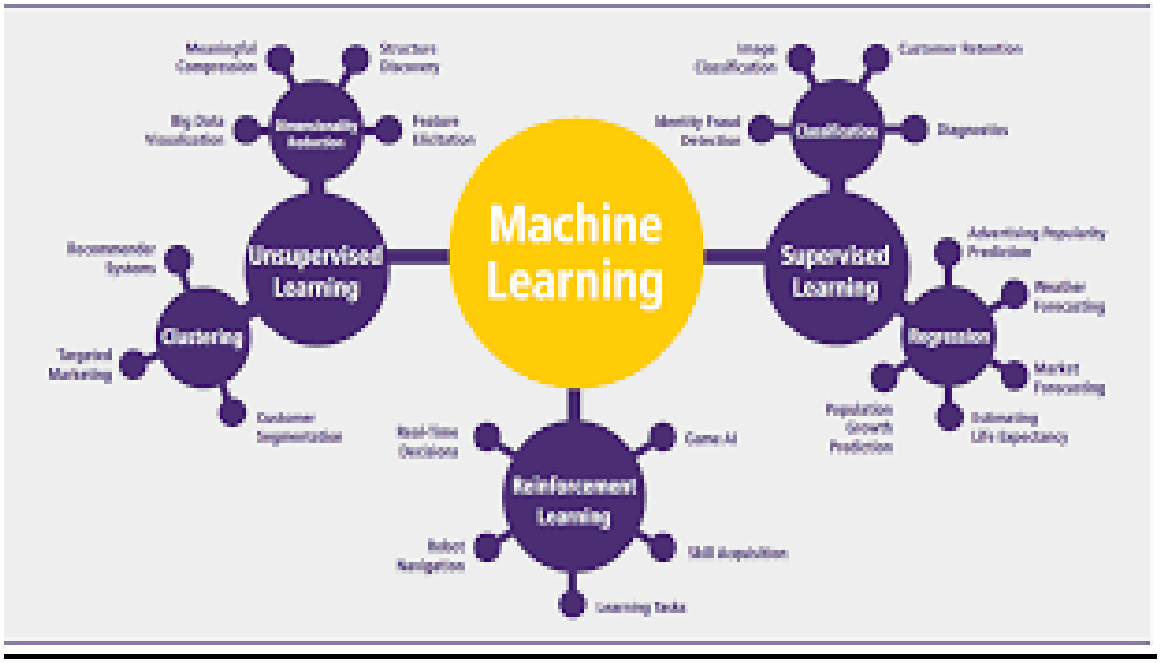
1. Our Primary objective would be to predict the condition , existence and the type of the breast cancers using the clinical data extracted .
2. Deployment and validation of Machine Learning model in the divergent manner i.e. considering more number of parameters, so that diagnose can be looked for accordingly
3. Types of diagnosis is also the crucial concern , thus Our aim would also be towards type of tumour that is emerging in the women's body.

Methodology :

Our motive is to build a model which could detect the of breast cancer in the prior stage.

- First step would be surely to gather the clinical information (dataset) and this data would be used to train and predict the disease.
- Data is divided into two different sets, training and testing sets. Testing dataset would be utilized at all for the validation.
- Validation and exploration would be performed on the data to know the other clinical factors contributing towards tumour cells.
- At last ,evaluation would be initialised in which, accuracies would be estimated and improvisation can also be done.





Models Utilized In The Project

1. ***Logistics Regression :*** Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

Advantages :

- Easy to Implement with clinical parameters
- Work well with binary datasets.
- It proves to be very efficient when the dataset has features that are linearly separable.

Disadvantages :

- Non linear problems can't be solved with logistic regression
- It is difficult to capture complex relationships using logistic regression

2. ***Random Forest Classifier:*** The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Advantages -

- Random Forest can be used to solve both classification as well as regression problems.
- Random Forest works well with both categorical and continuous variables.
- Random Forest can automatically handle missing values.

Disadvantages :

- Random Forest is a complex model to implement .
- Forest require much more time to train as compared to decision trees as it generates a lot of trees.

3. ***Decision Tree Classifier:*** The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

Advantages :

- A decision tree does not require scaling of data as well
- Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.

Disadvantages :

- A small change in the data can cause a large change in the structure of the decision tree causing instability.
- For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
- Decision tree often involves higher time to train the model

4. ***Linear SVC****:* SVM or Linear SVC is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

   Advantages :
   - Effective when the number of features are more than training examples.
   - Best algorithm when the classes are separable.
   - Suited for extreme case binary classification.

   Disadvantages :
   - Not suitable for large datasets, because takes a long time to process.
   - Does not perform well with overlapping classes.
   - Selecting the appropriate kernel function can be tricky.

Work Plan :

Review 1 :

Investigating the strategy and collecting papers and other essential resources.

Review 2 :

Contemplating for the efficient algorithm for the training and implementing the algorithm for the Cancer detection.

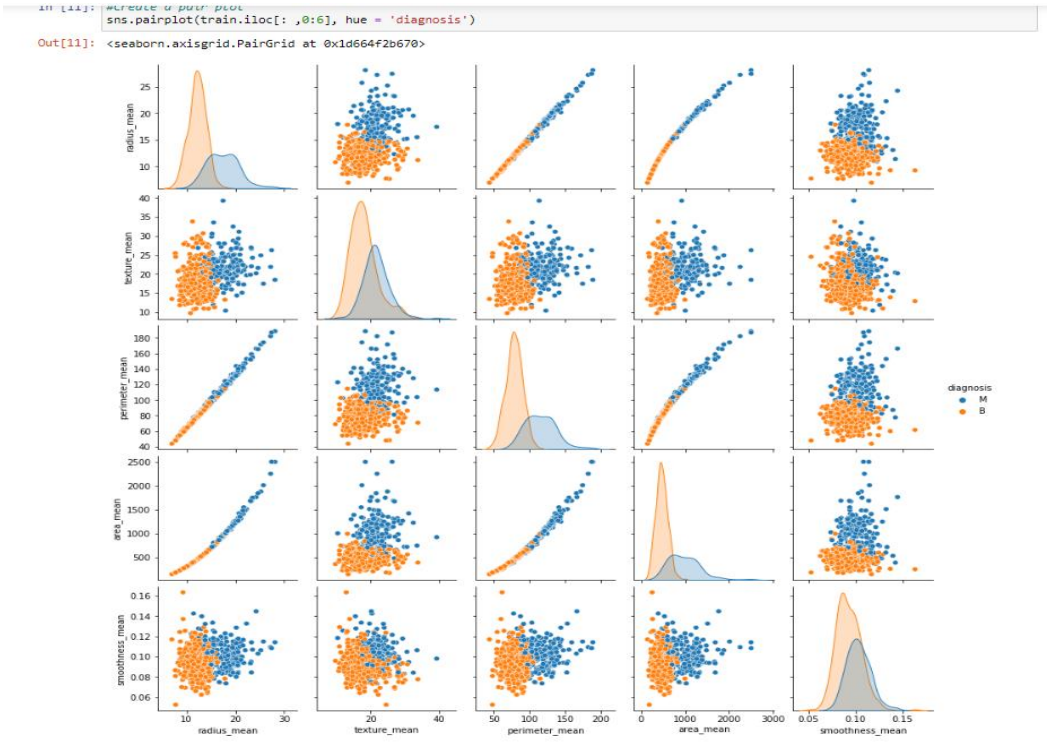
Review 3 :

Final demonstration with proper Validation and improving accuracy for the particular trained model.

Results/Accuracies :

Dataset-https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

- Preprocessing results like Distributed dataset on the basis of diagnosis typecan be viewed here. Moreover Many plots are plotted in order to get centralized distribution of the data based on diagnosis.



- After the modelling we get to know about parameters like the accuracy , precision, F1-Score etc for every model implemented. We can clearly see that the best parametric accuracy is achieved by random forest classifier.

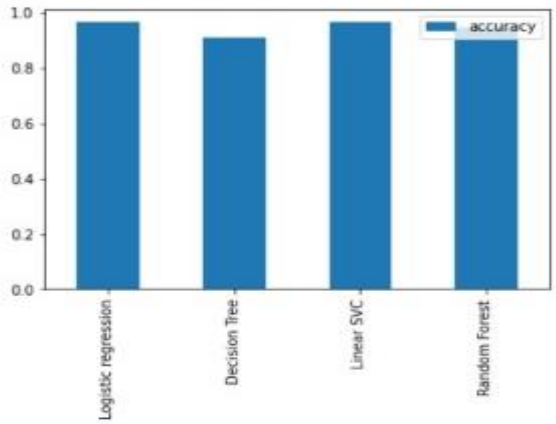
```
from sklearn.metrics import classification_report, accuracy_score
for i in range (len(model)):
    print('Model : ',model[i])
    print(classification_report(Y_test,model[i].predict(X_test)))
    print(accuracy_score(Y_test,model[i].predict(X_test)))
    print()
```

```
Model :  LogisticRegression()
              precision    recall  f1-score   support

           B       0.98      0.96      0.97       108
           M       0.94      0.97      0.95        63

    accuracy                           0.96       171
   macro avg       0.96      0.97      0.96       171
weighted avg       0.97      0.96      0.97       171

0.9649122807017544

Model :  DecisionTreeClassifier()
              precision    recall  f1-score   support

           B       0.95      0.92      0.93       108
           M       0.87      0.92      0.89        63

    accuracy                           0.92       171
   macro avg       0.91      0.92      0.91       171
weighted avg       0.92      0.92      0.92       171

0.9181286549707602

Model :  RandomForestClassifier()
              precision    recall  f1-score   support

           B       0.99      0.96      0.98       108
           M       0.94      0.98      0.96        63

    accuracy                           0.97       171
   macro avg       0.96      0.97      0.97       171
weighted avg       0.97      0.97      0.97       171

0.9707602339181286
```

- With 30% of testing data , we are getting maximum accuracy using random forest algorithm along with logistic regression model. Then we have accuracy plot for all the models.

```
Out[36]: {'Logistic regression': 0.9649122807017544,
          'Decision Tree': 0.9064327485380117,
          'Linear SVC': 0.9298245614035088,
          'Random Forest': 0.9649122807017544}
```

```
In [25]: model_compare = pd.DataFrame(scores, index=["accuracy"])
         model_compare.T.plot.bar();
```



Outcomes & Social Impacts:

- The project would be helpful for the aged women, as they could detect the cancer and type of tumour they are having at an early stage and get diagnosed.

- Clinical alerts could be sent to the respective patients for whom the condition is more vulnerable towards breast cancer. Mostly hospitals could give a notification according to the evaluated results.
- Other lifestyle parameters could be monitored which are most leading to breast cancer, thus user could try to counter them at the earliest.
- Correct type of diagnosis must be known as it differs according to the type of the tumour cells. So the model would be differentiating among the type of cancer a particular patient have.
- This model could be modified in future using more no of dataset and more sophisticated algorithms.

References :

[1] U. Haq et al., "Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques," in IEEE Access, vol. 9, pp. 22090-22105, 2021, doi: 10.1109/ACCESS.2021.3055806.

[2] P. E. Jebarani, N. Umadevi, H. Dang and M. Pomplun, "A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection," in IEEE Access, vol. 9, pp. 146153-146162, 2021, doi: 10.1109/ACCESS.2021.3123425.

[3] F. Azour and A. Boukerche, "Design Guidelines for Mammogram-Based Computer-Aided Systems Using Deep Learning Techniques," in IEEE Access, vol. 10, pp. 21701-21726, 2022, doi: 10.1109/ACCESS.2022.3151830..

[4] I. Hirra et al., "Breast Cancer Classification From Histopathological Images Using Patch-Based Deep Learning Modeling," in IEEE Access, vol. 9, pp. 24273-24287, 2021, doi: 10.1109/ACCESS.2021.3056516.

[5] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He and J. Fan, "Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and

Early Diagnosis," in IEEE Access, vol. 8, pp. 96946-96954, 2020, doi: 10.1109/ACCESS.2020.2993536.

[6] B. Xu et al., "Attention by Selection: A Deep Selective Attention Approach to Breast Cancer Classification," in IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 1930-1941, June 2020, doi: 10.1109/TMI.2019.2962013.

[7] D. Bardou, K. Zhang and S. M. Ahmad, "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks," in IEEE Access, vol. 6, pp. 24680-24693, 2018, doi: 10.1109/ACCESS.2018.2831280.

Thank You!!!