PROJECT REPORT

“Breast Cancer Detection using machine learning and Cloud Computing ”

Submitted By:
Shikha 2021UCS15**31**
Isha 2021UCS15**52**
Ananya 2021UCS15**73**

Cloud Computing
COCSC15
Department of Computer Science

November 2023

Netaji Subhas University of Technology

A STATE UNIVERSITY UNDER DELHI ACT 06 OF 2018, GOVT. OF NCT OF DELHI

Azad Hind Fauj Marg, Sector-3, Dwarka, New Delhi-110078

Table of Contents

INTRODUCTION

Breast cancer is a global cause for concern owing to its high incidence around the world. The alarming increase in breast cancer cases emphasizes the management of disease at multiple levels.

Breast cancer is a significant health challenge, especially for women:

1. **Survival in Metastatic Breast Cancer**: Metastatic breast cancer has a five-year survival rate below 30%, even with chemotherapy [1], emphasizing the need for improved detection and treatment.

2. **Global Impact:** In 2018, the International Agency for Research on Cancer (IARC) reported 2.3 million new breast cancer cases worldwide, accounting for 11.7% of all cancer diagnoses, with a 6.9% mortality rate [2]. This highlights breast cancer's global reach and severity.

3. **Incidence Disparities:** Breast cancer incidence rates vary, with higher-income countries reporting 571 cases per 100,000 individuals. Lower-income countries face disparities due to socioeconomic factors and healthcare access, reflecting global disparities.

4. **Complex Disease**: Breast cancer comprises over 100 distinct biological subtypes, each with unique characteristics, making diagnosis and treatment complex.

These facts underscore breast cancer's urgency as a public health issue, requiring immediate action for improved detection, treatment, and awareness to reduce its impact on women's health worldwide.

Problem Statement of Breast Cancer Detection using Ensemble Learning

Despite the global awareness and efforts to address the rising incidence of breast cancer, effective management of the disease remains a complex challenge. This project aims to enhance breast cancer diagnosis through the utilization of three distinct Ensemble models (voting classifier, Bagging, Boosting) applied to a dedicated Breast Cancer Diagnosis dataset. The evaluation of these models is conducted with a 75:25 training-test data split and 10-fold cross-validation, seeking to optimize accuracy and reliability in the identification and management of breast cancer cases. The objective is to contribute to the advancement of diagnostic tools for more efficient and precise intervention in the face of the escalating breast cancer crisis.

METHODOLOGY

1. Dataset Description:

- **Source**: The dataset used for this project is sourced from a breast cancer research database.

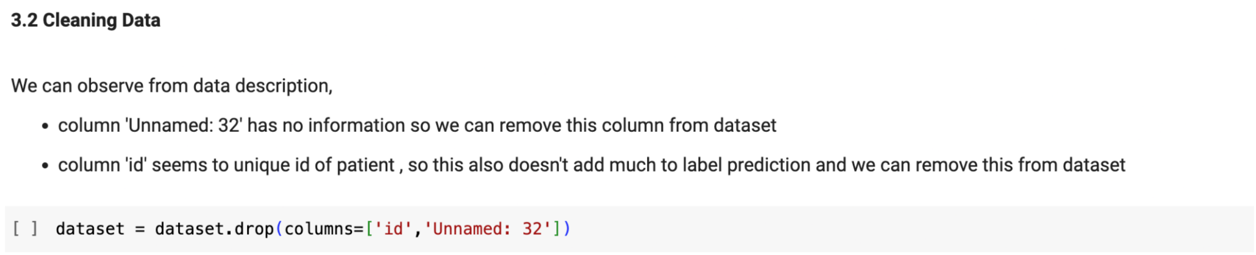
- **Division**: The dataset is divided into a **75:25** ratio for training and testing, respectively.

- **Size**: The dataset consists of 569 entries with 31 columns.

- Label Column: The target variable is **'Diagnosis'** indicating whether a tumor is malignant or benign.

- Feature Columns: The dataset includes 30 feature columns, representing various measurements related to the characteristics of cell nuclei.

- **Data Cleaning**:
  - Check for missing values: The data chosen by us does not contain any missing values but it did contain a NULL value column attribute. This column has been dropped from the final dataset.



3.2 Cleaning Data

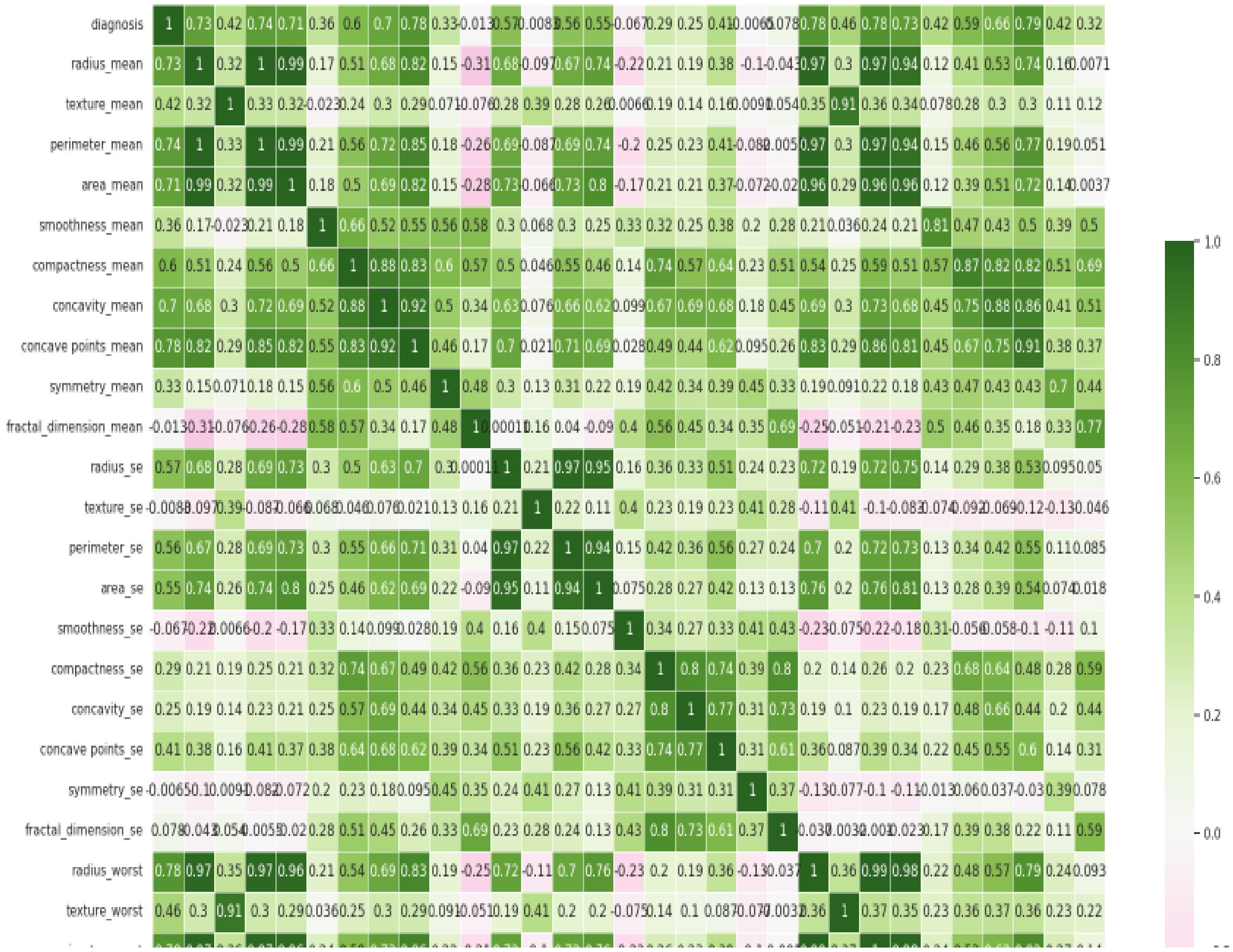
We can observe from data description,

- column 'Unnamed: 32' has no information so we can remove this column from dataset
- column 'id' seems to unique id of patient , so this also doesn't add much to label prediction and we can remove this from dataset

```
[ ]  dataset = dataset.drop(columns=['id','Unnamed: 32'])
```

  - Outlier treatment: Absence of any outliers.

- **Correlation Analysis:**
  - Computed the correlation matrix to identify highly correlated features. Less Correlated features are found out by getting a ranking of the features via RFE (recursive feature elimination).



Correlation matrix of our features

3. Machine Learning Algorithms and Ensemble Methods:

- **Ensemble Learning:**
  - Voting Classifier: Utilizing a Voting Classifier to combine predictions from multiple base models. This approach aggregates the decisions of multiple algorithms to enhance overall predictive performance.

- **Base Models:**
  - **Decision Tree (DT):** Decision trees are used for their ability to capture complex relationships within the data. They are robust and can handle non-linear patterns.

  - **Logistic Regression:** Logistic regression is chosen for its simplicity and interpretability, providing insights into the contribution of each feature to the model.

  - **K Nearest Neighbors (KNN):** KNN is a non-parametric algorithm that can capture local patterns and is effective when there is a clustering effect in the data.

4. Justification of Choices:

- **Ensemble Learning:**
  - **Advantages:** Ensemble methods often lead to improved generalization, robustness, and accuracy by combining diverse models.

- **Base Models**:
  - Decision Tree (DT): Effective at capturing complex relationships, useful for identifying intricate patterns in breast cancer data.

  - **Logistic Regression**: Provides a baseline linear model and helps in understanding the impact of individual features on the prediction.

  - **K Nearest Neighbors (KNN)**: Suitable for identifying local patterns and relationships, which might be beneficial in the context of breast cancer diagnosis.

  - **Random Forest:** This algorithm stands out for its reduced susceptibility to overfitting in comparison to decision trees, showcasing high levels of robustness and optimization as an ideal candidate for a base model.

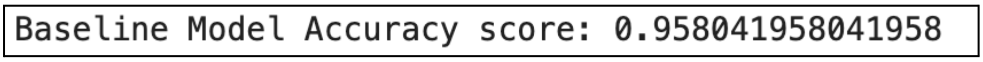
FEATURE SELECTION:

Our selected features range from **‘radius_mean’** to **‘fractal_dimension_worst’**.

- On these features we have applied RFE(Recursive Feature Elimination).
- The **RFE** is a recursive feature elimination (RFE) method where an estimator is trained on the initial set of features. Feature importance is assessed, and the least important features are iteratively pruned until the desired number of features is reached. This process **refines the feature set to enhance model efficiency and interpretability.**

Impact Of RFE on Our Model:

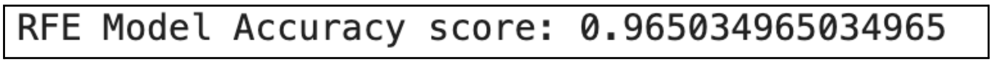
Before applying RFE:
The accuracy of model comes out to be: 95.8%



```
Baseline Model Accuracy score: 0.958041958041958
```

(only random forest as base model)

After applying RFE:
The accuracy of model comes out to be: 96.5%



```
RFE Model Accuracy score: 0.965034965034965
```

(random forest as base model as well as RFE for feature elimination)

Observation:

We infer that the model with feature selection is giving improved accuracy and f1 score values.

Hence, we apply this RFE feature selection to our Ensemble Learning Technique so that we can increase the prediction power of our algorithms by selecting the most critical variables and eliminating the redundant and irrelevant ones.

MODEL TRAINING:

Our Dataset has been trained and tested on the **Ensemble Learning Model**.

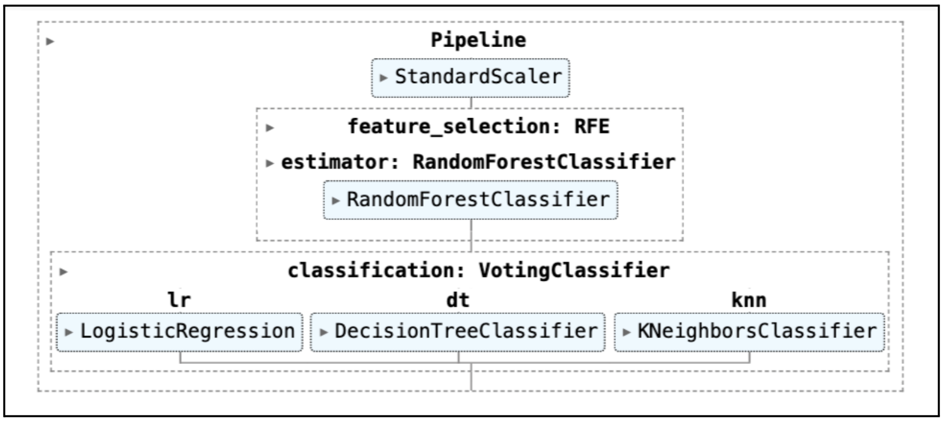
An ensemble model is a machine learning technique that combines the predictions of multiple individual models to create a stronger, more accurate, and robust predictor than any of its constituent models alone. The fundamental idea behind ensemble learning is that by aggregating diverse sources of information, the weaknesses of individual models are mitigated, and their collective strength enhances overall predictive performance.

Types of Ensemble Models:

1. VOTING CLASSIFIER:

Voting, or majority voting, combines the predictions of multiple models by selecting the most common prediction. It can be either hard voting (simple majority) or soft voting, where the class probabilities are averaged.

In our model, we have used the following models among which further selection has been done.



a. K- Nearest Neighbour:

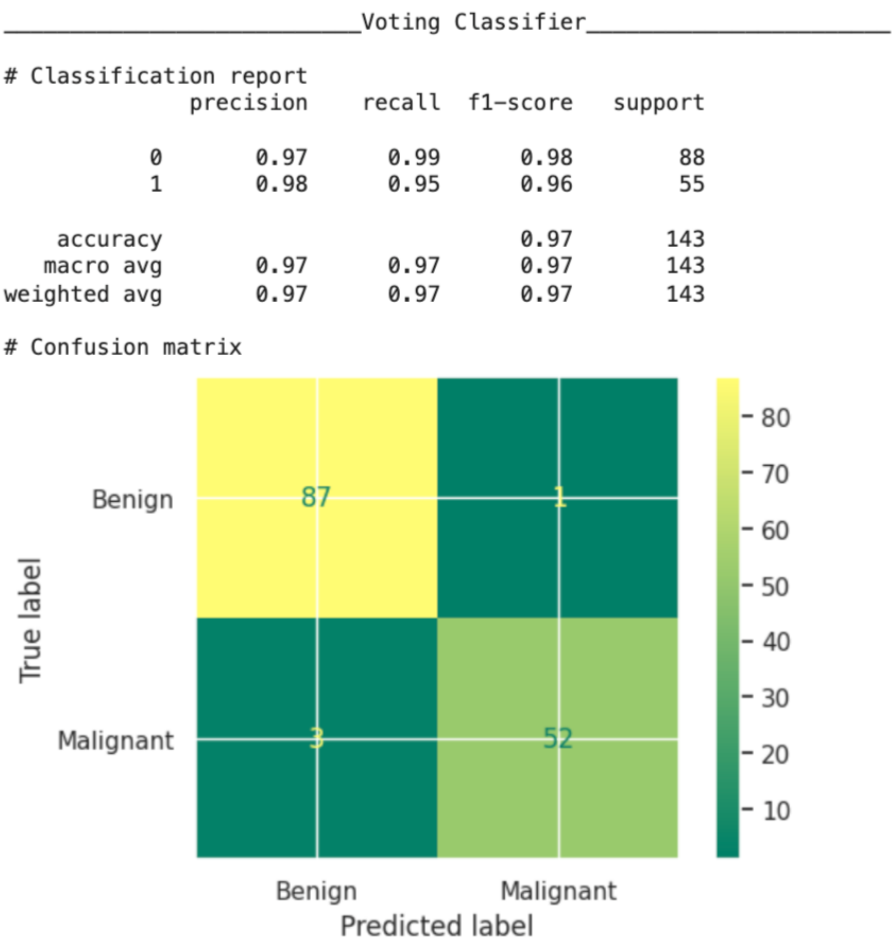
Accuracy obtained is: 93.6%

b. Decision Tree Classifier:

Accuracy obtained is: 94.6%

c. Logistic Regression

Accuracy obtained is: 95.1%



```
___________________________Voting Classifier_______________________

# Classification report
              precision    recall  f1-score   support

           0       0.97      0.99      0.98        88
           1       0.98      0.95      0.96        55

    accuracy                           0.97       143
   macro avg       0.97      0.97      0.97       143
weighted avg       0.97      0.97      0.97       143

# Confusion matrix
```

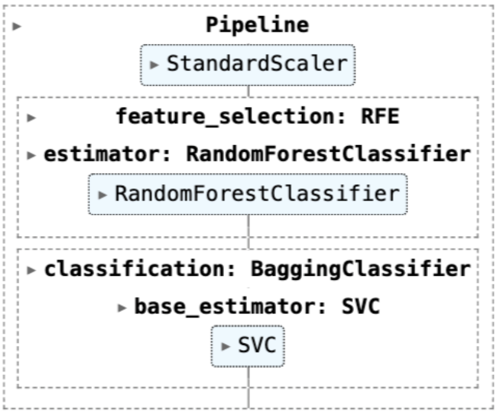
Final Accuracy of voting classifier along with it's Confusion matrix

Important:

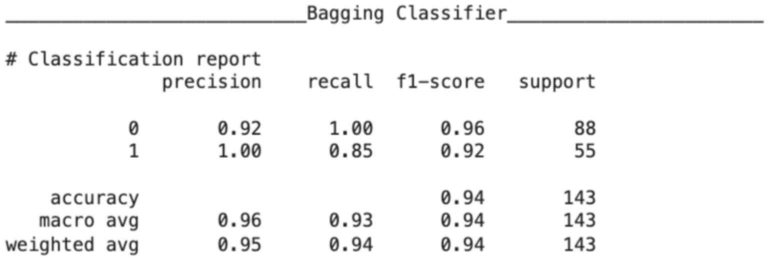
FN: 3 positive cases are misclassified (wrong negative predictions)

2. BAGGING CLASSIFIER:

Bagging involves training multiple instances of the same learning algorithm on different subsets of the training data, often created through bootstrapping (random sampling with replacement). The final prediction is typically an average or majority vote of the predictions from individual models.



In our problem we took a bagging ensemble of ***Support Vector Classifier*** as base estimator.

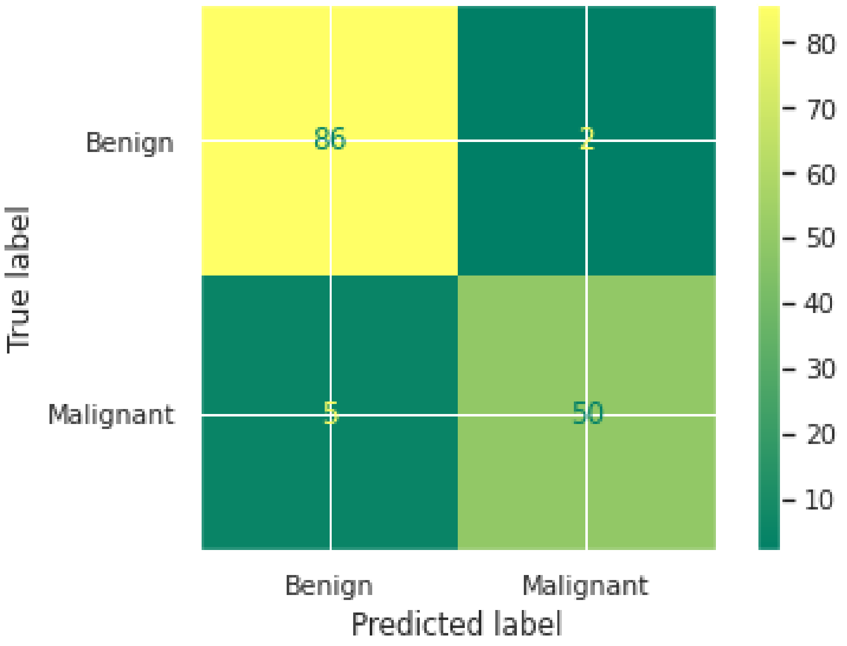


```
__________________________Bagging Classifier______________________

# Classification report
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        88
           1       1.00      0.85      0.92        55

    accuracy                           0.94       143
   macro avg       0.96      0.93      0.94       143
weighted avg       0.95      0.94      0.94       143
```



Final Accuracy of Bagging classifier along with it's Confusion matrix

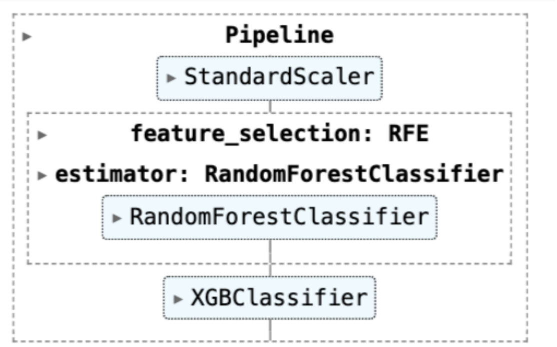
Important:

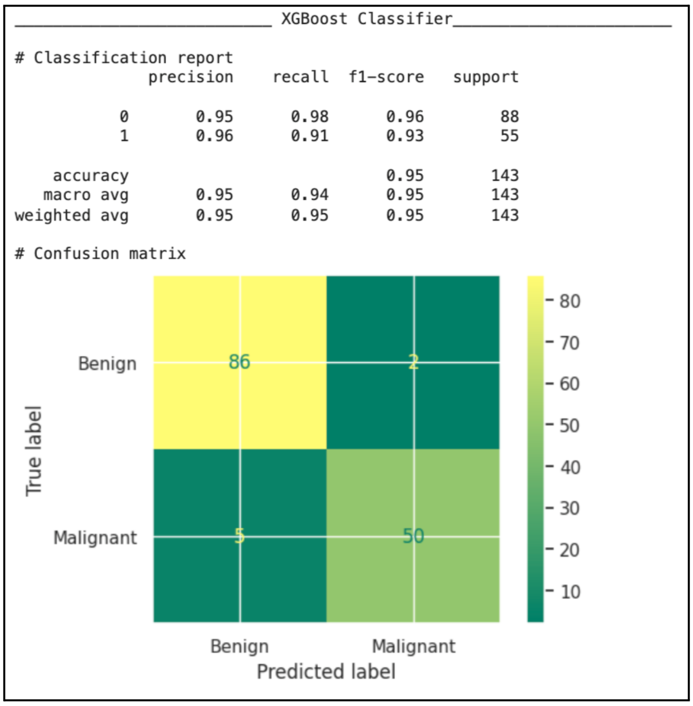
FN: 8 positive cases are misclassified (wrong negative predictions)

3. Boosting:

Boosting focuses on sequentially training models where each subsequent model corrects the errors of its predecessor. It assigns higher weights to misclassified instances, forcing the algorithm to pay more attention to the previously misclassified data points.

The **XGBoost** (eXtreme Gradient Boosting) is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler and weaker models.





```
___________________________ XGBoost Classifier_______________________

# Classification report
              precision    recall  f1-score   support

           0       0.95      0.98      0.96        88
           1       0.96      0.91      0.93        55

    accuracy                           0.95       143
   macro avg       0.95      0.94      0.95       143
weighted avg       0.95      0.95      0.95       143

# Confusion matrix
```

Final Accuracy of Boosting classifier along with it's Confusion matrix

Important:

FN: 5 positive cases are misclassified (wrong negative predictions)

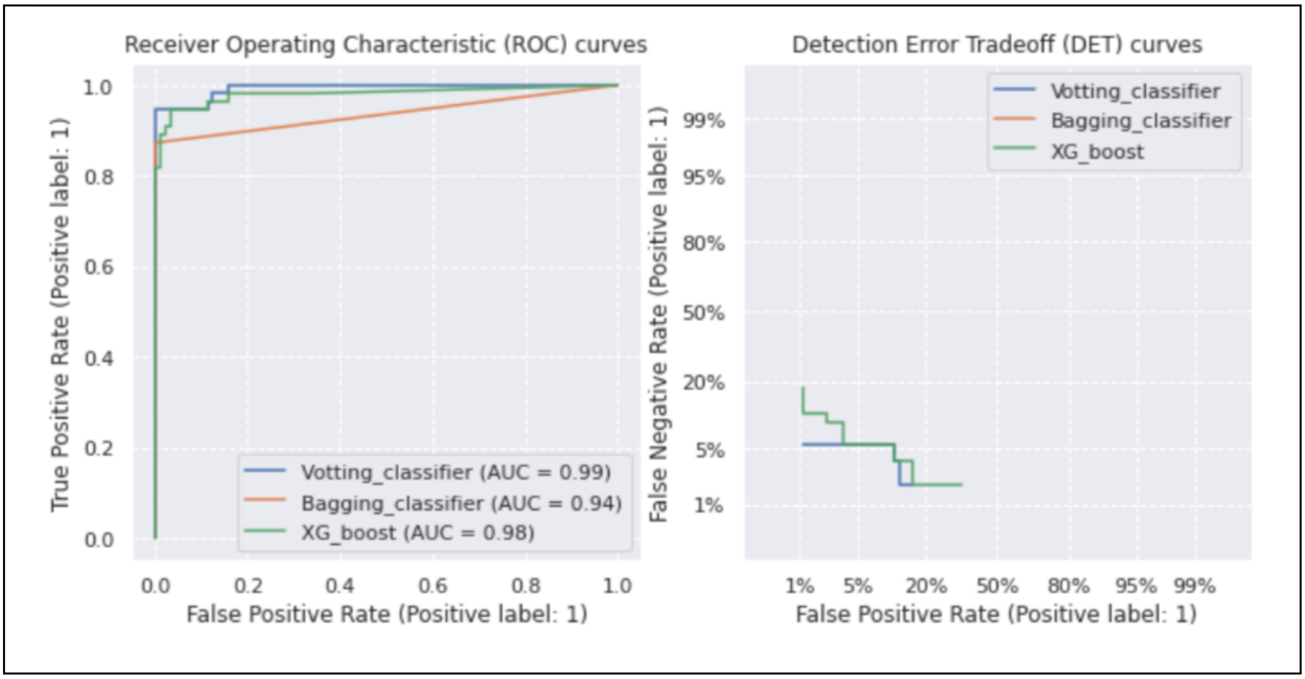
<u>RESULT</u>

Some of Our Observations:

comparing F1 Score:

- our proposed ensemble learning models achieved F1 score accuracies of 94%, 90% and 92% respectively.
- Voting Classifier performs better than XG Boost with an accuracy of 94%

Since our project is medical diagnosis we need to give maximum importance to Type II error in statistics(False Negative).False Negative is that the truth is positive, but the test predicts a negative. The person is sick, but the test inaccurately reports that they are not.To know the performance over False negative rate we will compare our ensemble models with confusion matrix ,ROC and DET curves.



Comparing ROC Curve:

- While comparing ROC Curves we found **Voting classifier** lies at ideal point that is top left corner and has larger area under the curve (AUC) which is 0.99 compared to XG Boost which has 0.98 and Bagging classifier with 0.94

Comparing DET Curve:

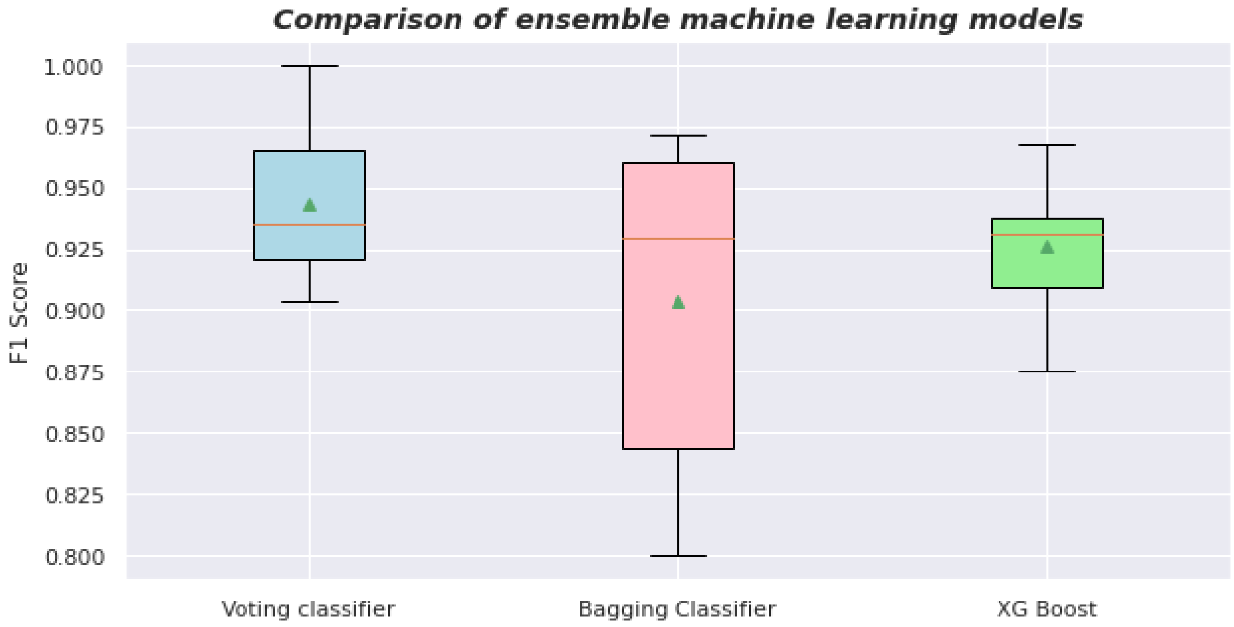
- The DET Curve has distinct advantages over the standard ROC type curve for presenting performance results where tradeoffs of two error types are involved. Here we can observe **Voting Classifier** has lesser error tradeoff compared to XG Boost.

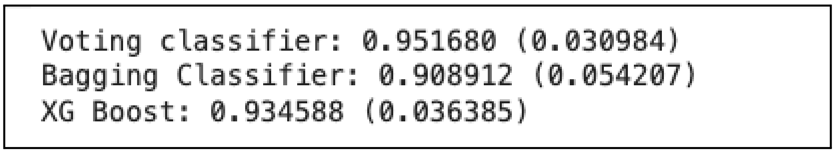
Comparing confusion matrix:

- Voting classifier has **less number of False Negatives** compared to XG Boost and Bagging Classifier

3 FN(voting classifier) < 5 FN(boosting classifier<8 FN (bagging classifier)

Model Comparisons:





```
Voting classifier: 0.951680 (0.030984)
Bagging Classifier: 0.908912 (0.054207)
XG Boost: 0.934588 (0.036385)
```

Hence, experimental results show that **Voting Classifier(soft voting)** was the most powerful prediction model than other ensemble machine learning techniques for Breast Cancer dataset.

ML with a touch of cloud

Data Ingestion:

- Uploading our breast cancer dataset to a cloud-based storage service like Amazon S3 (in AWS) or Azure Blob Storage (in Azure).

Data Cleaning and Transformation:

- Preprocessing service such as AWS Data Pipeline or Azure Data Factory. These services offer data transformation capabilities.
- Tasks like removing duplicates, handling missing values, and standardizing data formats would be taken care of.

Feature Extraction:

- For our image data, we'll use cloud-based services to extract meaningful features from images. Convolutional Neural Networks (CNNs) or pre-trained models can be applied for feature extraction. Such as AWS SageMaker or Google AI Platform

Benefit of cloud in the Project

1. **Data Storage:** Collecting and storing a large dataset of medical images and associated patient information is a critical component of a breast cancer detection system.
2. **Scalable Computing:** Efficiently train machine learning models.
3. **Model Deployment:** Make the model accessible to healthcare professionals.
4. **Collaboration:** Facilitate remote collaboration and image sharing.
5. **Data Security:** Data privacy and compliance with healthcare regulations.

Deployment of project on Cloud

- Deploying a Breast Cancer Prediction Model on Azure.
- Azure Advantages : Scalability, AI/ML services, Integration capabilities

1. Choosing VM type : OS selection (Ubuntu)Configuration.
2. Dependencies Installation: Installing Python, the necessary libraries (e.g., scikit-learn), and any other dependencies.
3. Uploading data cloud storage service Azure.
4. Azure Blob Storage :Model Deployment
5. Deploy trained model : Containerization (e.g. Docker)
6. Networking & Security.
7. Configure network, security, and SSL : Monitoring & Logging
8. Summarize Azure deployment process and benefits

Sources Cited

- Research Article Global Increase in Breast Cancer Incidence: Risk Factors and Preventive : *Measureshttps://shorturl.at/cvX67*

- Blueprint for cancer research:Critical gaps and Opportunities: *https://shorturl.at/cyHKZ*

- Comparative analysis of breast cancer detection using machine learning and biosensors: *https://www.sciencedirect.com/science/article/pii/S2667102621000887?ref=pdf_download&fr=RR-2&rr=8116e4c29c5b1b64*

- Breast cancer detection based on thermographic images using machine learning and deep learning algorithm: *https://shorturl.at/fzRZ1*

- Screening Mammography Breast Cancer Detection: *https://paperswithcode.com/paper/screening-mammography-breast-cancer-detection*