Cloud Computing Project

COCSC 15

Submitted by:
Shikha 2021UCS1531
Isha 2021UCS1552
Ananya 2021UCS1573

Battling Breast Cancer: Science, Strength, and Hope

Table of Contents

01
Introduction
Background and Domain

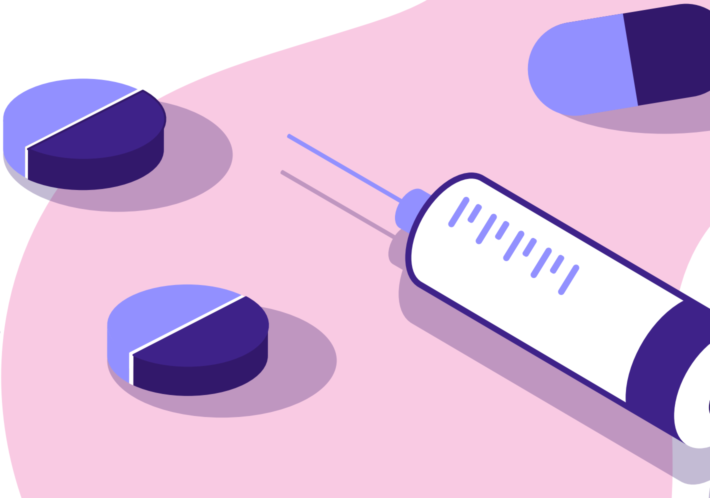
02
Research Gap
Advancements and Future Scope

03
Motivation
Our Objective & Motivation to choose this project

04
Contribution
Usage of different ML Algos for detection

05
Literacy Review
Comparison of various research papers on the basis of approach, performance and limitations

06
Sources Cited
List of Research articles used for the project



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.

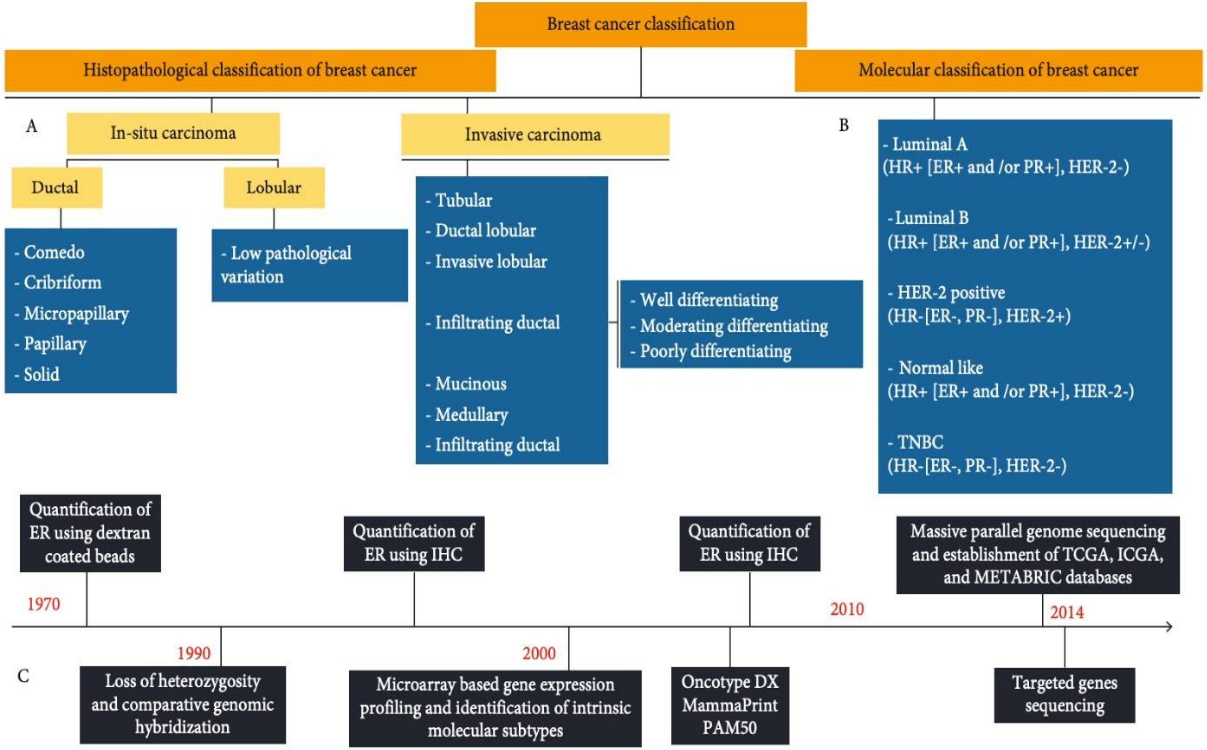
2. **Global Impact:** In 2018, the International Agency for Research on Cancer (IARC) reported *2.3 million* new breast cancer cases worldwide, accounting for *11.7%* of all cancer diagnoses, with a *6.9%* mortality rate [2]. This highlights breast cancer's global reach and severity.

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.

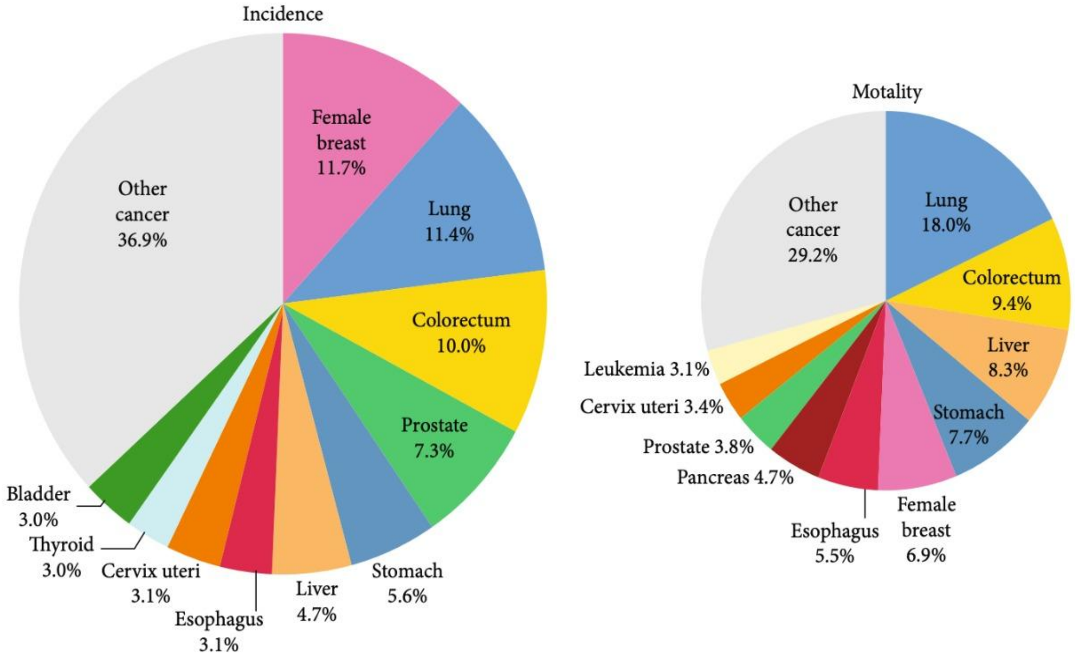
Breast Cancer: Not a single Disease


Breast cancer classification
Histopathological classification of breast cancer
Molecular classification of breast cancer
A
In-situ carcinoma
Ductal
Lobular
- Comedo
- Cribriform
- Micropapillary
- Papillary
- Solid
- Low pathological variation
Invasive carcinoma
- Tubular
- Ductal lobular
- Invasive lobular
- Infiltrating ductal
- Mucinous
- Medullary
- Infiltrating ductal
- Well differentiating
- Moderating differentiating
- Poorly differentiating
B
- Luminal A (HR+ [ER+ and /or PR+], HER-2-)
-Luminal B (HR+ [ER+ and /or PR+], HER-2+/-)
- HER-2 positive (HR-[ER-, PR-], HER-2+)
- Normal like (HR+ [ER+ and /or PR+], HER-2-)
- TNBC (HR-[ER-, PR-], HER-2-)
C
Quantification of ER using dextran coated beads
1970
1990
Loss of heterozygosity and comparative genomic hybridization
Quantification of ER using IHC
2000
Microarray based gene expression profiling and identification of intrinsic molecular subtypes
Quantification of ER using IHC
Oncotype DX MammaPrint PAM50
2010
Massive parallel genome sequencing and establishment of TCGA, ICGA, and METABRIC databases
2014
Targeted genes sequencing

Source: Research Article Global Increase in Breast Cancer Incidence: Risk Factors and Preventive [Measureshttps://shorturl.at/cvX67](https://shorturl.at/cvX67)

Breast cancer is not a single disease but rather a complex and diverse group of diseases. This diversity arises from variations in *genetic*, *molecular*, and *cellular* characteristics within breast tumors. These differences lead to various subtypes of breast cancer, each with distinct biological properties and clinical behaviors.

Distribution Of Cancer Incidences:



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

Domain:

- Breast cancer falls under the *healthcare* domain, which encompasses medical conditions, diagnoses, and treatments.
- Breast cancer stands out as one of the most aggressive types of cancer, with a notable mortality rate.(6.9%)
- Our goal is to use ML to address key concerns in breast cancer **detection.**

2,300,000

Diagnosed cases worldwide with a staggering **685,000** deaths globally.

Problem Statement : 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.

Motivation

Our project is driven by these **key motivations:**

1. Increasing Accuracy: Our primary objective is to enhance the accuracy of breast cancer detection, thereby improving patient outcomes and reducing false diagnoses.

2. Saving Lives:We are motivated by the potential to save lives through early breast cancer detection and improved treatment strategies.

3. Global Impact: Breast cancer is a global issue with an extremely high mortality rate and we are committed to contributing to a worldwide solution.

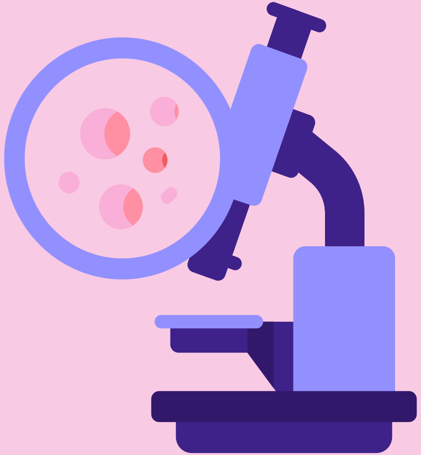
4. Health Equity: Addressing disparities in breast cancer outcomes is a priority, ensuring equitable access to quality care for all.

5. Complexity Challenge:We are motivated to unravel the complexity of breast cancer, advancing diagnosis and treatment by using various machine learning models that aid in our research and detection.

6. Empowerment: Our project empowers individuals by providing a user-friendly platform for breast cancer analysis and informed decision-making.

7. Healthcare Contribution: We aim to make a meaningful contribution to healthcare by leveraging technology and research to enhance breast cancer detection, treatment, and awareness.

Research Gap



Technology-Enhanced Approach:

Combining various machine learning (ML) techniques to improve accuracy and reliability involves using different ML approaches in tandem. Here's how it can be done:

Ensemble Learning:

Ensemble methods combine predictions from multiple ML models to make a final prediction. Common ensemble techniques include:

- **Voting:** Different ML models "vote" on the final prediction, and the most common prediction is chosen.
- **Bagging:** Bootstrap aggregating (bagging) trains multiple instances of the same model on different subsets of the data and combines their predictions.
- **Boosting:** Boosting focuses on training models sequentially, with each model correcting the mistakes of the previous one.

Contribution

Contributions Using Different ***ML Algorithms:***

1. Logistic Regression: Develop logistic regression models for binary classification of breast cancer cases. Fine-tune regularization parameters to balance model complexity and accuracy.

2. Neural Networks (Deep Learning): Experiment with deep neural networks, including convolutional neural networks for image-based breast cancer detection. Fine-tune architectures, layers, and activation functions for optimal performance.

3. Decision Trees, KNN, SVC: used for bagging, boosting in the project as the models.

3. Model Evaluation: Evaluate the performance of each algorithm using appropriate metrics like accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices.

4. Interpretability: Ensure that the predictions of ML models, especially complex ones like deep neural networks, can be interpreted and understood by healthcare professionals.

Features Selection

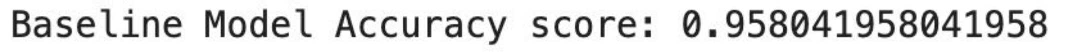
Our selected features range from **‘radius_mean’** to **‘fractal_dimension_worst’**.

- 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.**

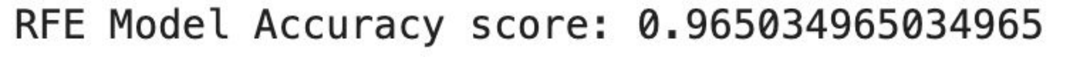
Observations:

Before applying RFE: The accuracy of the model obtained is *95.8%*

After applying RFE: The accuracy of the model obtained is *96.5%*



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



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

Implementation and Contents

Implementation of our project!

Table of contents

1. Importing Packages

2. Reading Data

3. Understanding and preparing data

3.1 Data Dimension

3.2 Cleaning data

3.3 Encoding the target variable

3.4 Statistical Summary

3.5 Exploratory Data Analysis

3.6 Feature Selection

4. Ensemble Learning Techniques

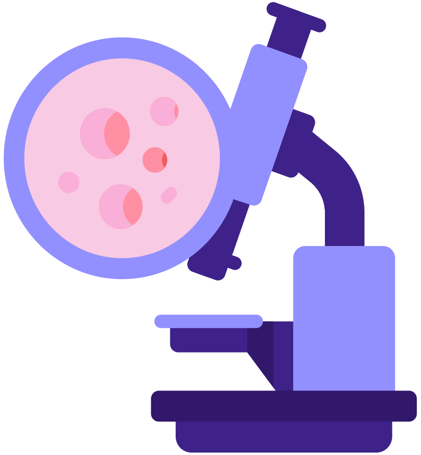
4.1 Voting Classifier

4.2 Bagging Classifier

4.3 XGBM

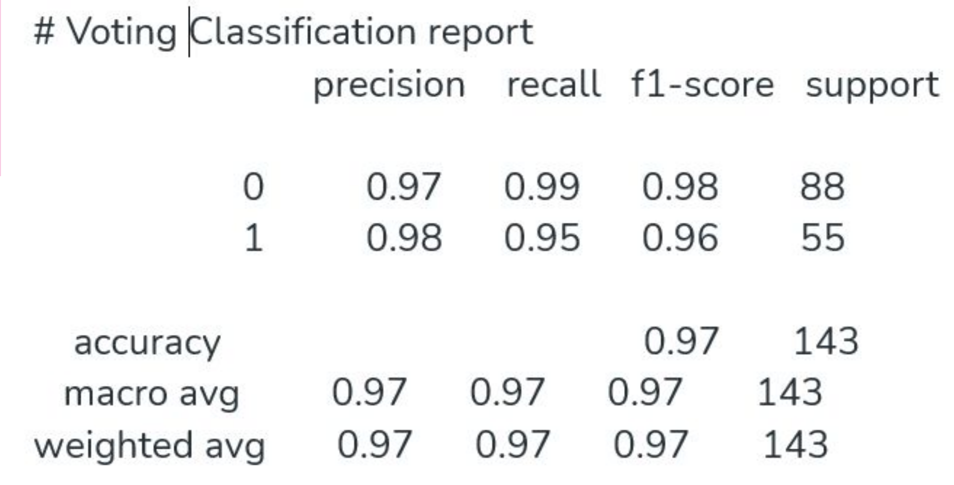
5. Model Comparison

6. Summary

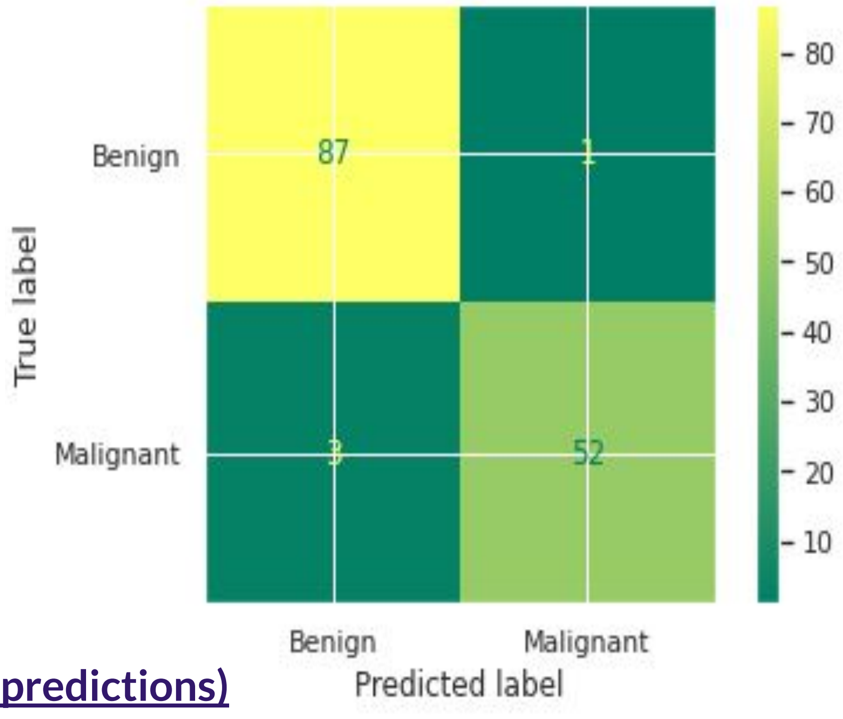


Result

Voting Classifier

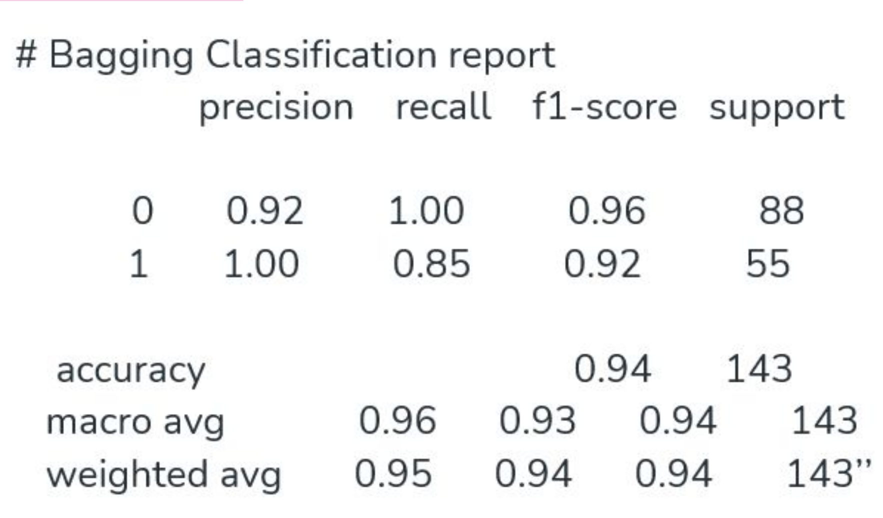


```
# Voting Classification report
              precision    recall  f1-score   support

           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
```

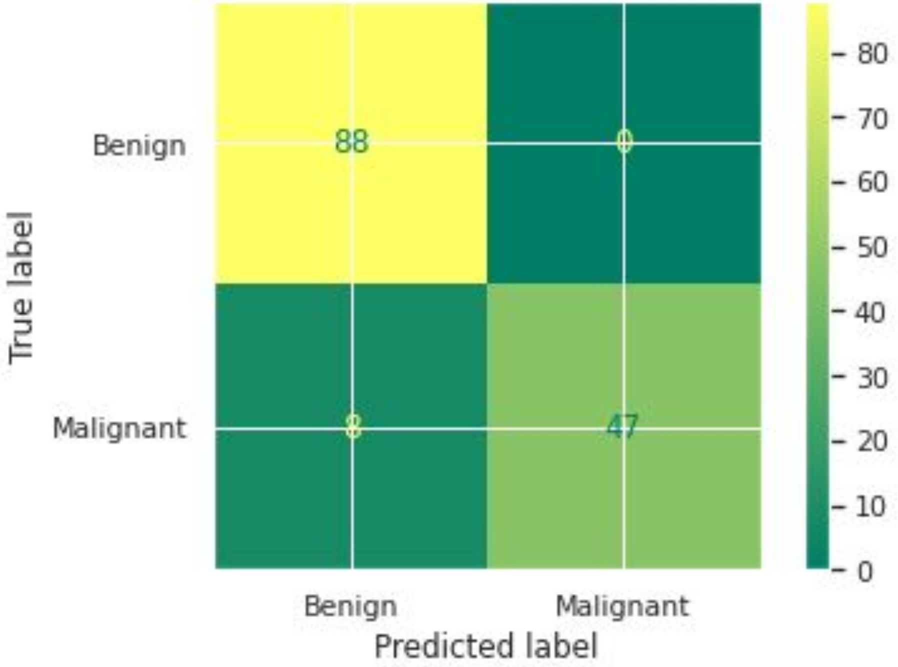


FN: 3 positive cases are misclassified (wrong negative predictions)

Bagging Classifier

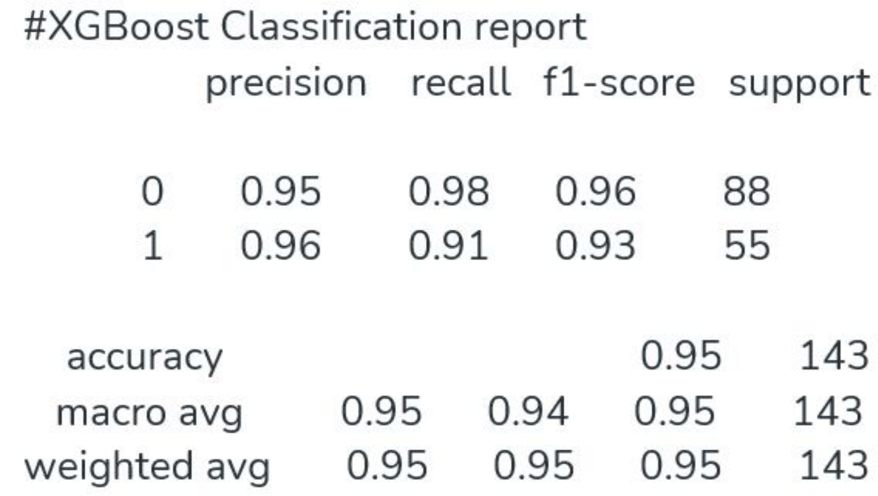


```
# Bagging Classification report
              precision    recall  f1-score   support

       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"
```

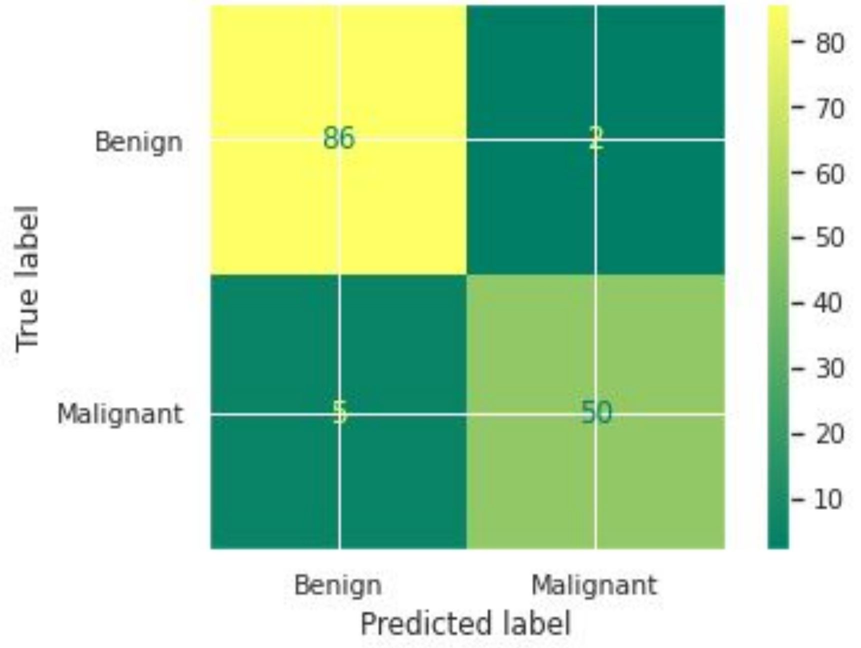


FN: 8 positive cases are misclassified (wrong negative predictions)

XGBoost Classifier



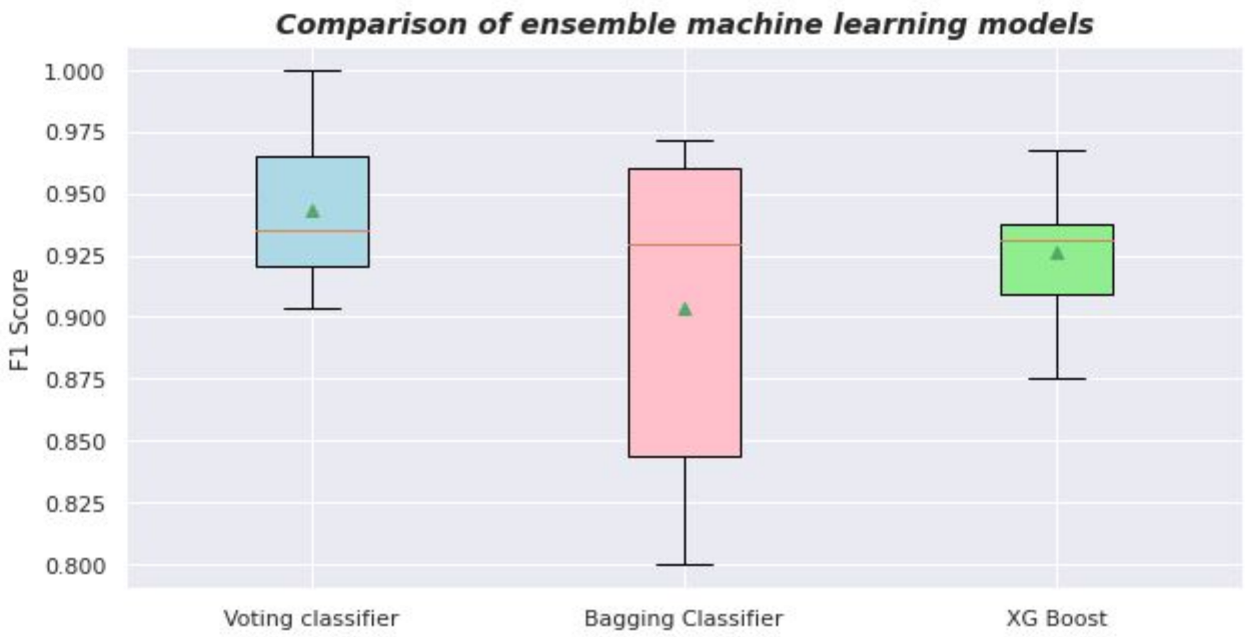
```
#XGBoost Classification report
              precision    recall  f1-score   support

           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
```



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

Model Comparisons:

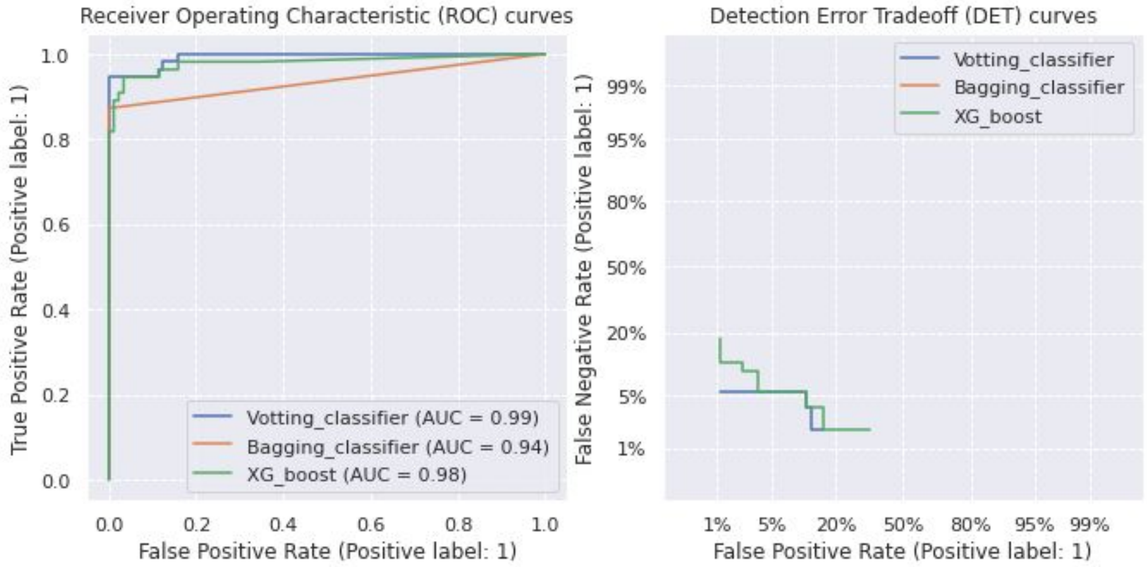
Between Voting, Bagging, Boosting classifiers



Observation:

our proposed ensemble learning models achieved F1 score accuracies of 94%, 90% and 92% respectively. **Voting Classifier** performs the best out of the 3 models!

Plotting ROC and DET Curves



- 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
- 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.

Enhancing Model with Cloud Computing

1. Scalability: Cloud computing enables flexible resource scaling to handle varying workloads.

2. Secure Data Storage: Robust cloud storage ensures secure storage of patient data and model checkpoints with encryption and access controls.

3. Remote Access: Cloud technology allows remote model access and data collaboration.

4. Cost Control: Tools for cost optimization, like instance sizing and reserved instances, help manage expenses.

5. Reliability: Load balancing and disaster recovery ensure high availability.

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 comply 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