PROJECT TITLE : BREAST CANCER PREDICTION USING MACHINE LEARNING

IN-LAB INTERNSHIP REPORT

Submitted by

M.NISHANTHAN(23IT099)
M.ARUL KAARTHIKEYAN(23IT020)
T.PRASANNA(23IT107)

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

THIAGARAJAR COLLEGE OF ENGINEERING,
MADURAI – 625 015

August 5, 2024 to August 16, 2024

THIAGARAJAR COLLEGE OF ENGINEERING(TCE),
MADURAI- 625 015

BONAFIDE CERTIFICATE

Certified that this In-Lab Internship report **BREAST CANCER PREDICTION USING ML** is bonafide work of Nishanthan(23IT097), Arul Kaarthikeyan(23IT020), Prasanna(23IT107) III sem. Department of Information Technology who carried out the In-lab Internship at TCE between August. 05,2024–August. 16, 2024.

Submitted for Evaluation held at Thiagarajar College of Engineering on August 16,2024.

EXAMINER 1
(Name with Signature)

EXAMINER 2
(Name with Signature)

Table of Contents

Abstract :

This project focuses on predicting Breast cancer using machine learning algorithms to enhance the accuracy and reliability of the predictions.

We employed a variety of algorithms, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), Gradient Boosting (GB), and Logistic Regression (LR), with the goal of optimizing key performance metrics such as accuracy, precision, confusion matrix, F1 score, recall, ROC, and AUC.

The dataset was sourced from the UCI Repository and was further enhanced using data augmentation techniques to increase its size and variability.

The dataset was split into different training and testing configurations: 70% training and 30% testing, 80% training and 20% testing, and combined two algorithms in 7c2 combinations(Hybrid Model) to explore potential improvements in prediction performance.

List of Tables

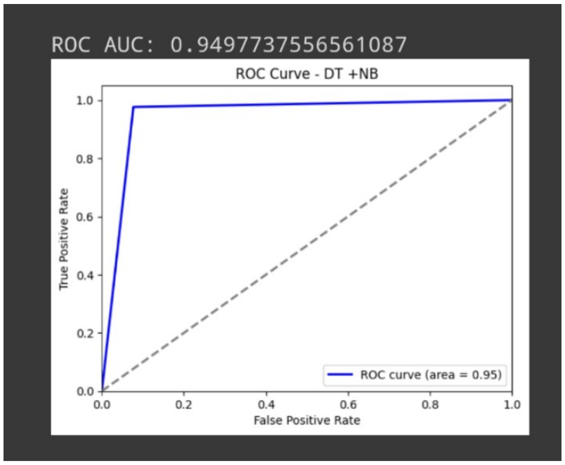
Algorithm	Accuracy	Confusion Matrix	Precision		Recall		F1-Score	
			0	1	0	1	0	1
Logistic Regression	0.83	[[31 3] [14 52]]	0.69	0.95	0.91	0.79	0.78	0.86
Naïve Bayes	0.6	[[31 3] [37 29]]	0.46	0.91	0.91	0.44	0.61	0.59
Support Vector Machine	0.94	[[33 1] [5 61]]	0.87	0.98	0.97	0.92	0.92	0.95
KNN	0.86	[[30 4][10 56]]	0.75	0.93	0.88	0.85	0.81	0.89
Decision Tree	0.94	[[33 1] [5 61]]	0.87	0.98	0.97	0.92	0.92	0.95
Random Forest	0.98	[[65 0][2 83]]	0.97	1	1	0.98	0.98	0.99
Gradient Boosting	0.98	[[33 1][1 65]]	0.97	0.98	0.97	0.98	0.97	0.98

Algorithm	Accuracy	Confusion Matrix	Precision		Recall		F1-Score	
			0	1	0	1	0	1
Logistic Regression	0.8	[[49 16] [14 71]]	0.78	0.82	0.75	0.84	0.77	0.83
Naïve Bayes	0.65	[[57 8] [45 40]]	0.56	0.83	0.88	0.47	0.68	0.6
Support Vector Machine	0.953	[[59 6] [1 84]]	0.98	0.93	0.91	0.99	0.94	0.96
KNN	0.853	[[56 9][13 72]]	0.81	0.89	0.86	0.85	0.84	0.87
Decision Tree	0.9533	[[60 5] [2 83]]	0.97	0.94	0.92	0.98	0.94	0.96
Random Forest	0.96	[[62 3][2 83]]	0.97	0.97	0.95	0.98	0.96	0.97
Gradient Boosting	0.94	[[60 5][4 81]]	0.94	0.94	0.92	0.95	0.93	0.95

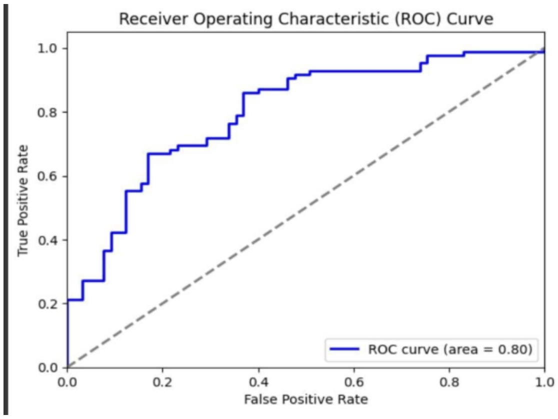
Algorithm	Accuracy	Confusion Matrix	Precision		F1-Score	
			0	1	0	1
RF AND GB	0.94	[[59 6] [3 82]]	0.95	0.93	0.93	0.95
RF AND LR	0.98	[[62 3][0 85]]	1	0.97	0.98	0.98
RF AND SVM	0.89	[[51 14] [3 82]]	0.94	0.85	0.86	0.91
RF AND KNN	0.87	[[46 19] [1 84]]	0.98	0.82	0.82	0.89
DT AND RF	0.95	[[60 5] [2 83]]	0.97	0.94	0.94	0.96
NB AND RF	0.65	[[39 26] [26 59]]	0.6	0.69	0.6	0.69
GB AND LR	0.94	[[60 5] [4 81]]	0.94	0.94	0.93	0.95
GB AND SVM	0.94	[[60 5] [4 81]]	0.94	0.94	0.93	0.95
GB AND KNN	0.94	[[60 5] [4 81]]	0.94	0.94	0.93	0.95
GB AND DT	0.94	[[59 6] [3 82]]	0.95	0.93	0.93	0.95
GB AND NB	0.94	[[60 5] [4 81]]	0.94	0.94	0.93	0.95
LR AND SVM	0.86	[[59 6] [15 70]]	0.8	0.92	0.85	0.87
LR AND KNN	0.81	[[50 15] [14 71]]	0.78	0.83	0.78	0.83
LR AND DT	0.71	[[37 28] [17 68]]	0.69	0.71	0.62	0.75

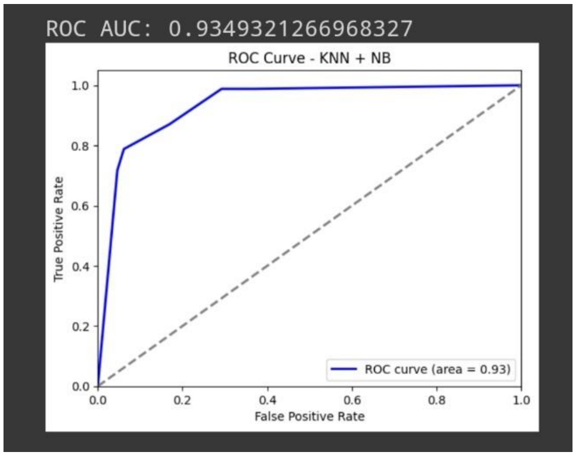
LR AND NB	0.86	[[59 6][15 70]]	0.8	0.92	0.85	0.87
SVM AND KNN	0.91	[[52 13] [2 83]]	0.96	0.86	0.87	0.92
SVM AND DT	0.89	[[51 14] [3 82]]	0.94	0.85	0.86	0.91
SVM AND NB	0.94	[[57 8] [1 84]]	0.98	0.91	0.93	0.95
KNN AND DT	0.87	[[46 19] [1 84]]	0.98	0.82	0.82	0.89
KNN AND NB	0.85	[[54 11] [11 74]]	0.83	0.87	0.83	0.87
DT AND NB	0.95	[[60 5] [2 83]]	0.97	0.94	0.94	0.96

List of Figures

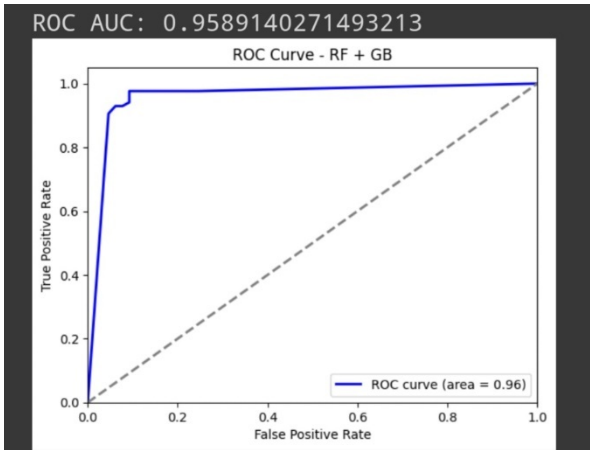


```
ROC AUC: 0.9497737556561087
```

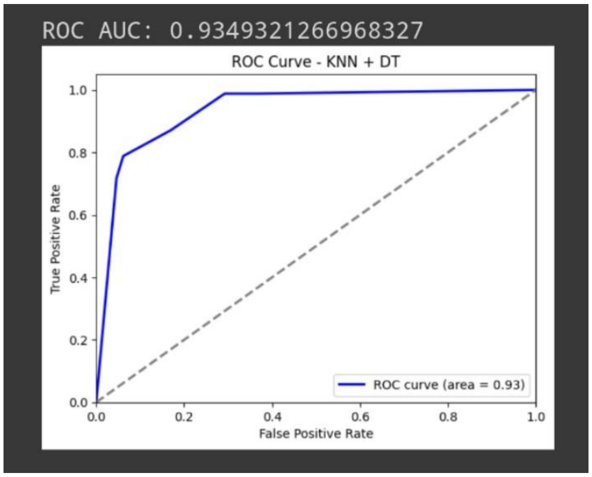




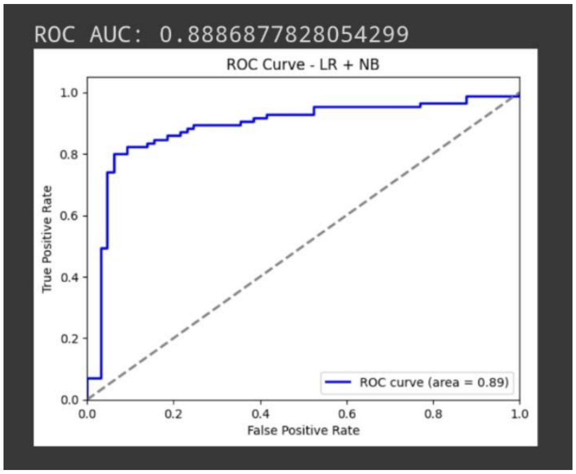
```
ROC AUC: 0.9349321266968327
```



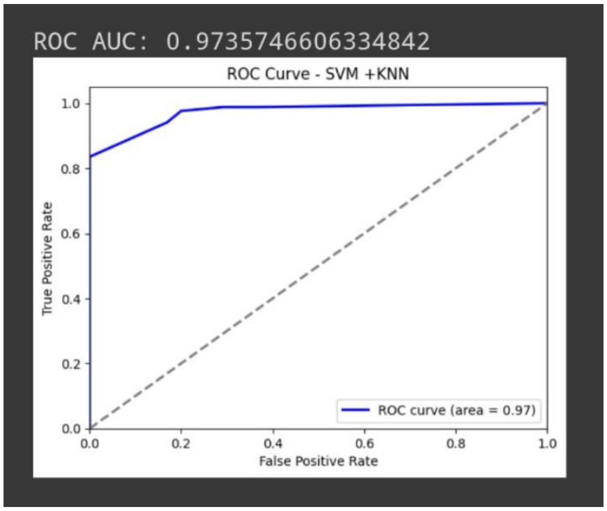
```
ROC AUC: 0.9589140271493213
```



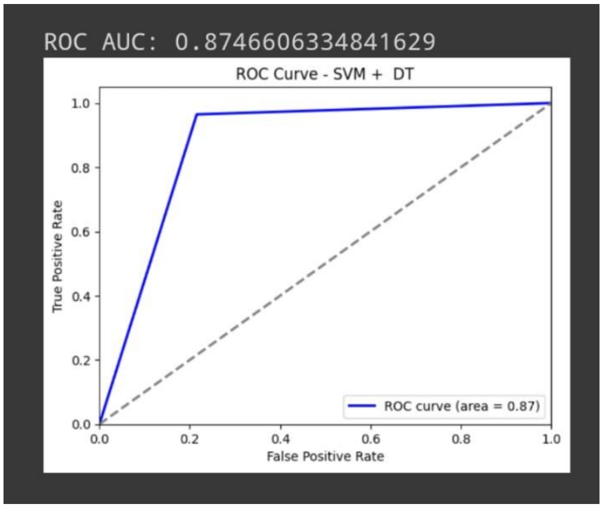
```
ROC AUC: 0.9349321266968327
```



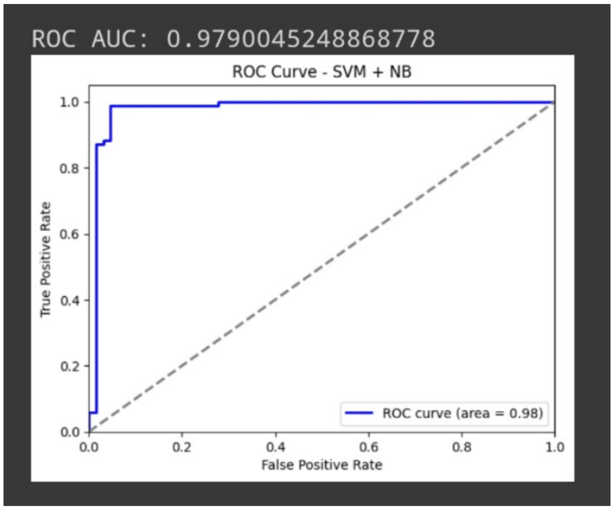
```
ROC AUC: 0.8886877828054299
```



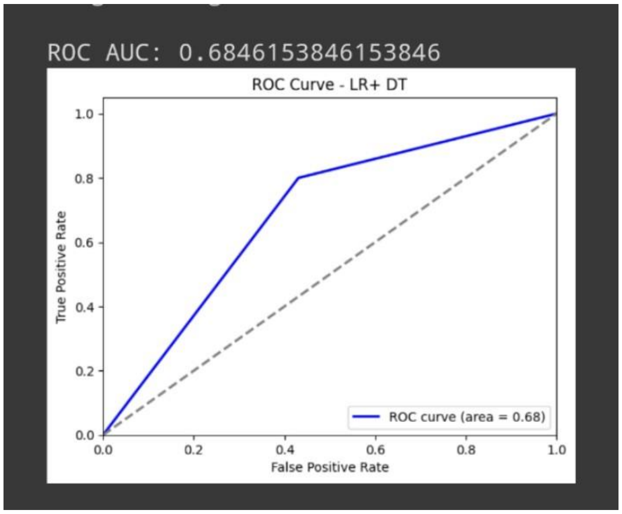
```
ROC AUC: 0.9735746606334842
```



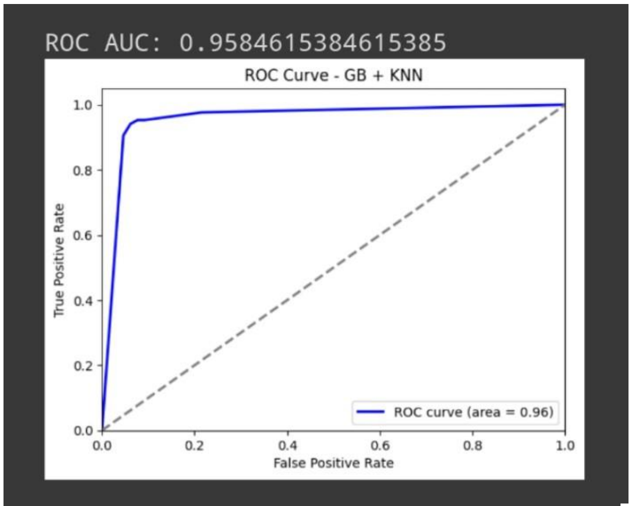
```
ROC AUC: 0.8746606334841629
```



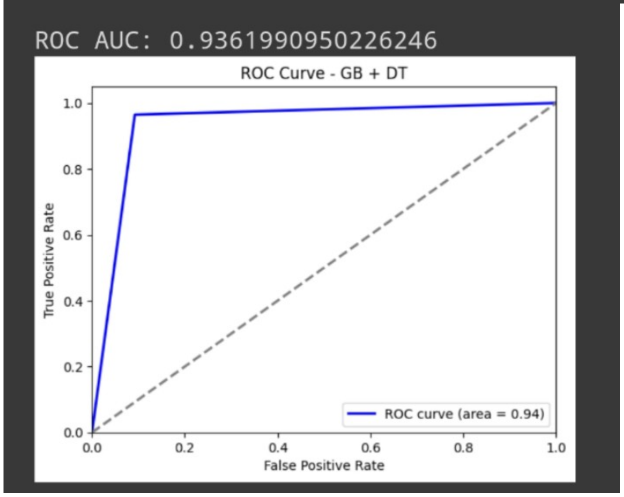
```
ROC AUC: 0.9790045248868778
```



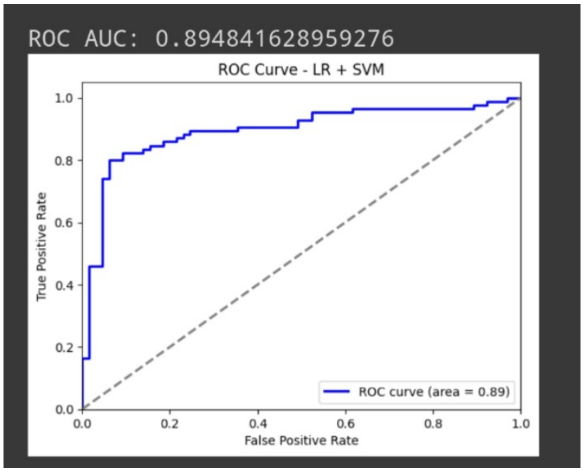
```
ROC AUC: 0.6846153846153846
```



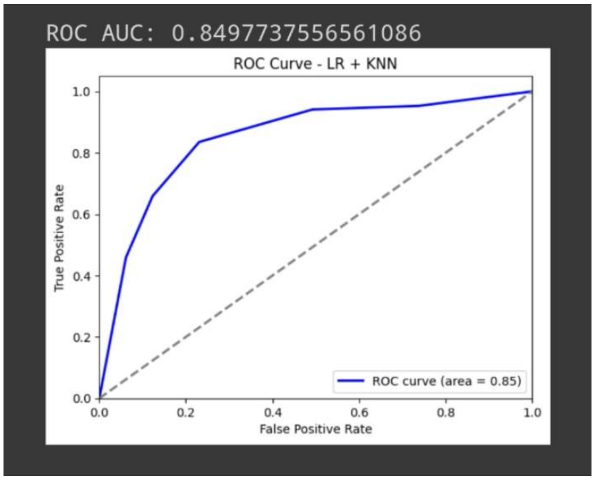
```
ROC AUC: 0.9584615384615385
```



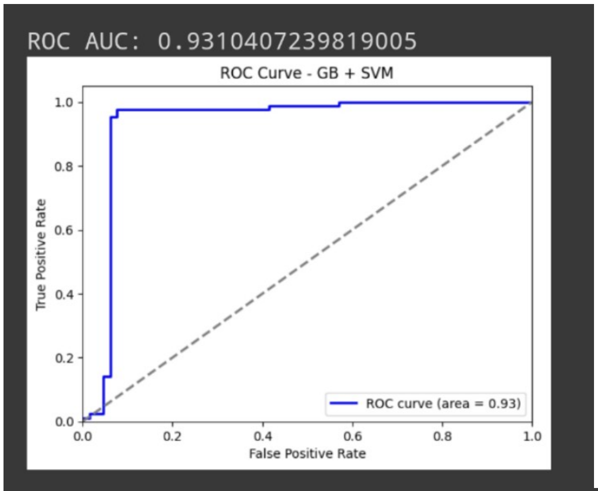
```
ROC AUC: 0.9361990950226246
```



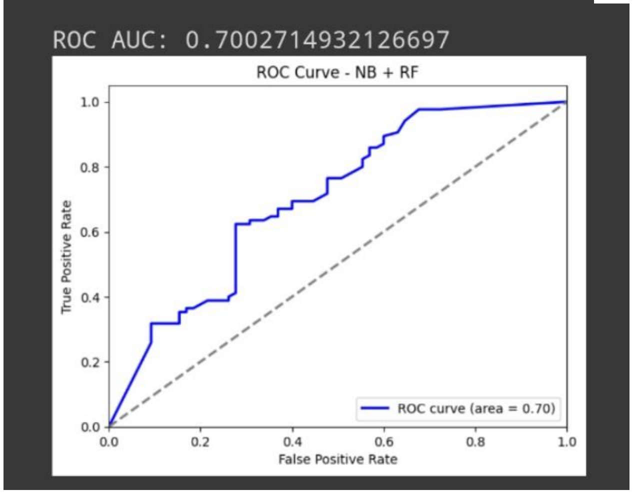
```
ROC AUC: 0.894841628959276
```



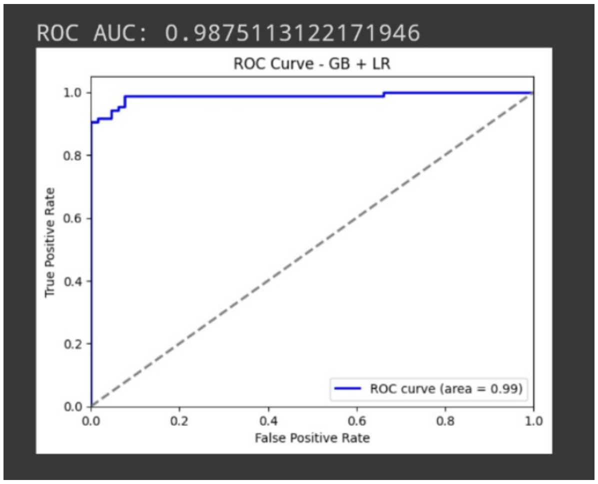
```
ROC AUC: 0.8497737556561086
```



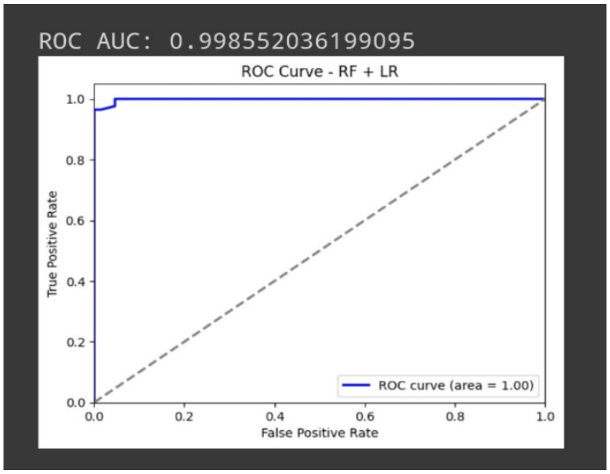
```
ROC AUC: 0.9310407239819005
```



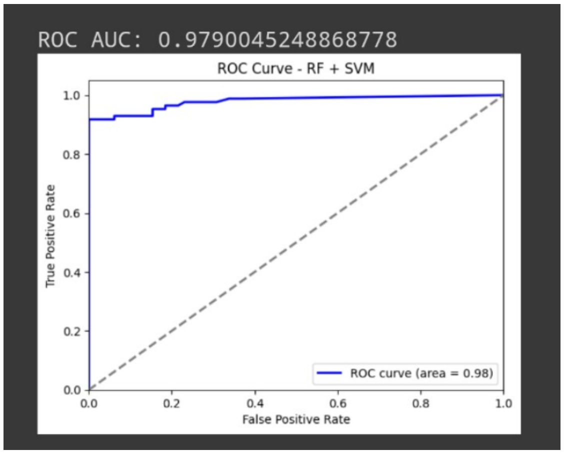
```
ROC AUC: 0.7002714932126697
```



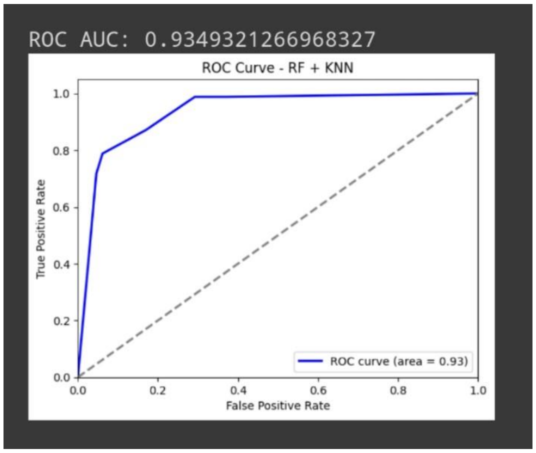
```
ROC AUC: 0.9875113122171946
```



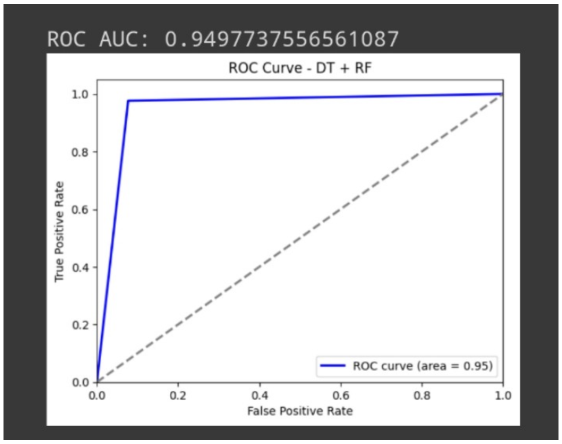
```
ROC AUC: 0.998552036199095
```



```
ROC AUC: 0.9790045248868778
```



```
ROC AUC: 0.9349321266968327
```



```
ROC AUC: 0.9497737556561087
```

List of Abbreviations

- ML : Machine Learning
- AI : Artificial Intelligence
- DT : Decision Tress
- NB : Naïve Bayes
- NN : Neural Network
- KNN: K-Nearest Neighbors
- RF : Random Forest
- SVM: Support Vector Machine
- LR : Logistic Regression

- PCA: Principal Component Analysis
- ROC: Receiver Operating Characteristic
- AUC: Area Under the Curve
- TP : True Positive
- FP : False Positive
- TN : True Negative
- FN : False Negative

Introduction :

Breast cancer prediction is a critical task in medical diagnostics, where early detection significantly increases the chances of successful treatment.

The project aims to leverage machine learning algorithms to improve prediction accuracy and other key performance metrics.

By experimenting with various algorithms, data splits, and augmentations, the goal is to find the most effective model for predicting Breast cancer.

Background :

Industry Context

The healthcare industry has witnessed significant advancements in recent years, particularly in the application of technology to improve patient outcomes. The rise of machine learning and data analytics has revolutionized medical diagnostics, offering new tools for early disease detection, including Breast cancer. Breast cancer prediction models have gained traction as a valuable resource for healthcare providers, allowing for more accurate diagnosis and personalized treatment plans. As medical data becomes increasingly complex, the demand for sophisticated, reliable predictive models continues to grow.

Development of Breast cancer Prediction Models

This project was initiated to address the critical need for accurate and reliable Breast cancer prediction tools. Using a combination of machine learning algorithms such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), Gradient Boosting (GB), and Logistic Regression (LR), the project focuses on maximizing prediction accuracy and robustness. The dataset from the UCI Repository was augmented to enhance model performance, ensuring that the prediction models are trained on a comprehensive and diverse set of data. Additionally, to ground the project in the current scientific landscape, we reviewed and analyzed ten recent breast Breast cancer prediction studies from Google Scholar. These studies provided valuable insights into the strengths and limitations of existing models, guiding the development and refinement of our approach

Competitive Landscape

In the rapidly evolving field of medical diagnostics, the development of Breast cancer prediction models is highly competitive. Various institutions and research groups are exploring different machine learning techniques to improve prediction accuracy and clinical applicability. This project differentiates itself by integrating multiple algorithms and leveraging data augmentation to create a model that not only performs well across various metrics like accuracy, precision, and recall but also offers insights into the disease's progression. By focusing on a broad array of performance indicators, this project aims to set a new standard in the field of Breast cancer prediction, contributing to the larger goal of enhancing patient care through technological innovation.

Objectives :

Improve Accuracy: Enhance the overall accuracy of the Breast cancer prediction models.

Optimize Key Metrics: Focus on improving precision, recall, F1 score, ROC, and AUC.

Model Comparison: Evaluate the performance of individual algorithms and their combinations.

Data Augmentation: Increase the dataset size using augmentation techniques to improve model robustness.

Data Split Strategies: Test different training and testing data splits (70-30%, 80-20%) to identify the most effective approach.

Algorithm Combinations: Explore 7c2(21) algorithm combinations to identify potential synergies between models.

Purpose of the Work :

Enhance Predictive Accuracy : Utilize machine learning techniques to create a robust model for Breast cancer prediction, which could potentially aid in early detection and improve patient outcomes.

Explore Data-Driven Solutions: Investigate how different machine learning algorithms perform on Breast cancer data and determine the most effective combinations and configurations.

Contribute to Medical Research: Provide insights and methodologies that can be applied to real-world medical scenarios, potentially contributing to advancements in Breast cancer diagnostics.

Optimize Resources: Develop a machine learning model that can efficiently process medical data with high accuracy, minimizing the need for excessive computational resources.

Problem Formulation :

Inconsistent Model Performance: There is a challenge in achieving consistent performance across different machine learning models due to variations in the data and algorithm sensitivity.

Data Limitations: The original dataset from the UCI Repository may not be large or varied enough to train a highly accurate model, necessitating the use of data augmentation techniques.

Overfitting and Generalization: Balancing the complexity of models to avoid overfitting while maintaining generalizability to new, unseen data is a critical issue**.**

Metric Optimization: Achieving optimal results across multiple performance metrics (accuracy, precision, F1 score, etc.) can be difficult, especially when different metrics may conflict.

Combination of Algorithms: Identifying effective combinations of algorithms (7c2) and understanding their synergistic effects on Breast cancer prediction is complex and requires extensive experimentation.

Methodology :

Data Source: The dataset was obtained from the UCI Repository, a well-known source of machine learning datasets.

Data Augmentation: To enhance the dataset's size and variability, data augmentation techniques were applied.

Data Splitting: The dataset was split into three configurations:

70-30% Split: 70% of the data was used for training and 30% for testing.

80-20% Split: 80% of the data was used for training and 20% for testing.

7c2(21) Combinations(Hybrid Model): Combinations of two algorithms were tested to identify potential performance improvements.
Model Evaluation: Each model and combination was evaluated using metrics such as accuracy, precision, confusion matrix, F1 score, recall, ROC, and AUC.

Results and Discussion :

Performance Analysis: The project observed variations in model performance based on different data splits, augmentation, and algorithm combinations. The ROC and AUC metrics, along with accuracy and precision, provided insights into the effectiveness of each approach.

Model Comparisons: The combined algorithm approach (7c2) showed potential in improving certain metrics over single algorithms.

Challenges: Issues related to overfitting, model complexity, and the need for balanced datasets were identified and addressed.

Conclusion :

The study successfully demonstrated the potential of machine learning algorithms in Breast cancer prediction.

By comparing different models, data augmentation, and data split strategies, we identified the most promising approaches for improving prediction accuracy and other key metrics.

Future work will focus on refining these models and exploring additional algorithm combinations to further enhance performance.

Future Enhancements :

Incorporate More Data: Continue expanding the dataset with more diverse samples and advanced augmentation techniques.

Advanced Techniques: Explore more advanced machine learning techniques such as deep learning and ensemble methods.

Real-World Application: Apply the model to real-world medical data for validation and potential clinical use.

Image Data set : Use the images as a Data set to predict .

Appendix :

A.Data Sources :

UCI Repository: The dataset was sourced from the UCI Machine Learning Repository, containing essential breast Breast cancer data for model training and testing.

Google Scholar: Reviewed ten recent breast Breast cancer prediction papers to understand current methodologies and best practices.

B.Data Augmentation Techniques :

Feature Scaling: Standardization and normalization ensured equal contribution from all features.

Data Splitting: Split into 70%/30% and 80%/20% training/testing sets, with an additional combination of algorithms.

C. Machine Learning Algorithms Applied :

KNN, SVM, RF, NB, DT, GB, and LR to improve classification and accuracy.

D. Performance Metrics : Evaluated models using accuracy, precision, recall, F1 score, confusion matrix, ROC curve, and AUC.E.

E.Technical Specifications :

Development Environment: Code was run on Google Colab using Python, with libraries including NumPy, Scikit-Learn, Matplotlib, and Pandas.

Hardware: Utilized Google Colab's virtual environment with adequate computational resources.

F. Review of Literature : Summarized key methodologies and findings from ten Breast cancer studies, with properly formatted references.