Comparison of Breast Cancer Prediction Using Machine Learning

BY

Md. Mehedi Hasan
ID: 212-15-4092

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Mr. Dewan Mamun Raza
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

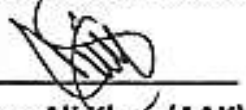
13 JANUARY 2025

APPROVAL

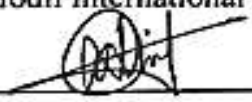
This project titled "**Comparison of Breast cancer prediction using machine learning**", was submitted by **Md. Mehedi Hasan, ID: 212-15-4092** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13 January 2025

BOARD OF EXAMINERS

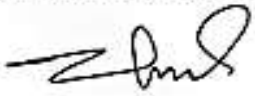
Dr. S.M. Aminul Haque (SMAH) — **Chairman**
Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Md. Abbas Ali Khan (AAK) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Mr. Md. Aynul Hasan Nahid (AHN) — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Md. Zulfiker Mahmud (ZM) — **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University

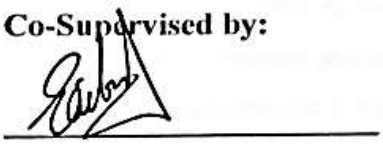
DECLARATION

I hereby declare that this project has been done by me under the supervision of **Mr. Dewan Mamun Raza, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University**. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

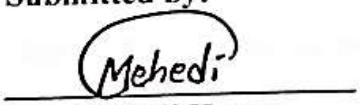
Mr. Dewan Mamun Raza,
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:



Saiful Islam
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:



Md. Mehedi Hasan
ID: 212-15-4092
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

I extend my sincere gratitude to the Almighty for His blessings, enabling us to successfully complete my final year project.

My heartfelt thanks go to **Mr. Dewan Mamun Raza, Assistant Professor in the Department of Computer Science and Engineering at Daffodil International University, Dhaka**. I appreciate my supervisor's deep knowledge and keen interest in the field of "Machine Learning," which guided us through the project. His unwavering patience, scholarly guidance, continual encouragement, energetic supervision, constructive criticism, valuable advice, and diligent review of multiple drafts at various stages have been instrumental in the completion of this project.

I express my gratitude to the Head of the Department of CSE, for his generous assistance in bringing my project to fruition. My thanks also go to the other faculty members and staff of the CSE department at Daffodil International University.

Acknowledgment is due to my fellow course mates at Daffodil International University who participated in discussions during the course of my work.

Lastly, I would like to recognize and appreciate the constant support and patience of my parents.

ABSTRACT

Recent times, breast cancer has seen a concerning rise, affecting a significant proportion of women. To tackle this pressing issue, extensive research efforts have been dedicated to devising effective methodologies for early detection and prediction. Our proposed approach leverages techniques to predict potential risks also promote recent alert of breast cancer. What sets our approach apart is its practical applicability in real-world scenarios, offering a straightforward method for breast cancer prediction. We harnessed the power of four datasets hosted on the Kaggle platform and integrated various classifiers, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), K-Nearest Classifier (KNN), among others, into our model. The results were promising, with the KNN achieving a noteworthy test accuracy of 81.14% for Dataset A, KNN of 97.2% for dataset B, KNN of 98.85% for dataset C and LR of 96.125% for dataset D. Furthermore, Bagging KNN also demonstrated accuracy matching this high standard of 99.42%. To further enhance performance, we implemented a range, including Bagging, Boosting, Stacking and Voting algorithms, optimizing each classifier with the best parameters through hyper-parameter tuning. Through our experimental investigation, we not only contributed to the body of knowledge on breast cancer detection and prediction but also identified the KNNB (K-Nearest Classifier with Bagging) model as the most accurate, achieving an outstanding accuracy rate of 99.42% for breast cancer predictions. This research endeavors to provide invaluable insights into breast cancer management, offering a potential solution for early intervention and ultimately improving patient outcomes.

Keywords: Breast Cancer, Early detection, Prediction methodologies, Classifiers, Datasets (Kaggle), Accuracy rate, Bagging, Boosting, Stacking, Voting algorithms, Hyper-Parameter tuning, KNNB, Patient outcomes, Machine learning models

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Breast cancer, a disease characterized by uncontrolled spread of damaged tissues, poses a significant threat to individuals and society at large. Detecting this type of cancer, which arises from the unregulated growth of breast tissue, is a pressing concern, as its incidence is rapidly on the rise. The primary challenge in addressing this issue is the timely and accurate identification of affected areas during diagnosis. Artificial intelligence (AI) plays a pivotal role in this context, as it leverages private medical information to evaluate various features and patient diagnostic data to detect the presence of cancer in women. In our research, we meticulously examined data and identified key components that prove instrumental in disease determination. The dataset we utilized contained comprehensive information about an individual's body size, composition, and breast cancer status. Many researchers have collaborated on the quest to identify malignant cells in the body using machine learning techniques. However, their methods often lacked practicality and simplicity when it came to predicting breast cancer. As a response, we provide our strategy to increase the possibility that women will correctly identify breast cancer. Our methodology incorporates two main categories of machine learning techniques: supervised learning and unsupervised learning. In supervised learning, models are created based on labeled data, where the algorithm learns from in and out pairs. Alternatively, unseen learning involves constructing models from unlabeled data, enabling the discovery of hidden patterns and information that may not be immediately apparent. This multi-faceted approach aims to improve the diagnosis of breast cancer, ensuring that individuals can receive timely and accurate medical attention.

1.2 MOTIVATION

Breast cancer is an increasingly common health concern that affects a significant portion of the female population, and its prevalence continues to rise. Various factors, including dietary habits and the use of cosmetic creams, contribute to the occurrence of breast cancer.

As per the international study, in 2020, a total of 2,261,419 reported cases of individuals affected by breast cancer, and it resulted in 684,996 recorded deaths [1]. The study also identified several risk factors associated with breast cancer, such as alcohol consumption, higher birth weights, and women who reached menarche at an early age. Numerous research efforts have been made to predict and understand breast cancer better. However, many of these studies have struggled to achieve a high level of accuracy in their predictions. In response to this challenge, we embarked on rigorous research and developed a method that demonstrates the best accurate in forecasting the presence of breast cancer in both patients under regular medical observation and those under suspicion of having the disease. Our innovative approach aims to provide a more reliable and precise means of early detection and diagnosis for breast cancer, thereby contributing to improved healthcare outcomes for individuals at risk of this disease.

1.3 The rationale of the study

We introduce a methodology for prediction disease individuals. Also have recognized the increasing impact of this disease on our society and identified treatment resources. Resource-constrained environment, the cost of cancer diagnosis and symptom analysis is prohibitively high. As researchers, we aim to address this problem through the application of advanced technology and innovative techniques to make breast cancer prediction more accessible and cost-effective. Our goal is to contribute to improved healthcare services and early detection of breast cancer in regions with limited resources.

1.4 Research Objective

1. To examine how dietary patterns, particularly unhealthy diets, contribute to the development and progression of breast cancer diseases.

2. To assess the likelihood and accuracy with which individuals with breast cancer disease can be identified using the proposed model.

3.To determine the model's effectiveness in predicting early signs and risk factors associated with breast cancer diseases.

4. To analyze the primary advantages of the proposed model in terms of accuracy, efficiency, and usability in predicting breast cancer illnesses.

5. To investigate potential practical applications and implications of the study's findings in real-world healthcare settings.

6. To identify and outline the necessary safety measures and protocols to ensure the ethical and safe implementation of the model.

1.5 PROJECT MANAGEMENT AND FINANCE

We offers a cost-effective and practical solution that can be implemented in real-world situations. It can serve as a valuable resource to immediate assessment within our nation. While high-configuration tools may yield the best results and optimal model performance, the use of basic tools is still feasible for implementing the prediction process.

1.6 Report Layout

Chapter 2 explores the relevant research that has already been done by other researchers. Prior to starting our investigation, it is imperative that we carefully review the introduction and motivation. As a result, we expound on the introduction, offering a comprehensive explication of the recommended methodology, and the motivation section, clarifying the reasoning for our prediction. Following the conclusion of the Introduction phase, we turned our attention to pertinent research and gathered internal data for our project. After evaluating the machine learning algorithms, we selected for the methodology portion using our dataset, we concluded which one performed the best. After pre-processing, data testing took place, which finally produced the comparative result—which is the one we were after. We go into great detail about this result in our last part, which is called the conclusion.

CHAPTER 2

BACKGROUND

2.1. PRELIMINARIES

Breast Cancer illness is accurately diagnosed by the application of machine learning algorithms. We examine the application and analysis of patient diagnostic reports in this section. A number of models are included, including RF, LR, DT, KNN, XGB, and GB, and they use algorithms to conduct the investigation. In this part, deep learning methods are applied to better enhance the research. Several of us utilized different models in our research; they are detailed in the corresponding section.

2.2. RELATED WORKS

Several techniques like ML has been developed for categorization, demonstrating their suitability for the task. These classifiers are based on machine learning algorithms employing "tree structures" [2]. Comparisons were made among RF, NB, SVM, and K-NN models in a manner similar to the suggested approach. The most effective classifier, SVM, achieved an impressive precision rate of 97.9% when using a Multilayer Vision model with five levels. Shamrat et al. focused on enhancing reliability in early breast cancer detection using the Wisconsin Breast Cancer Diagnostic dataset (WBCD) through ML-based systems [3]. They employed six supervised classification methods. The investigation revealed that SVM outperformed other methods, achieving the highest classification accuracy at 97.07%, with NB and RF closely behind. Mumine Kaya Keles aimed to predict and identify breast cancer early using non-invasive and painless techniques with data mining algorithms, regardless of tumor size [4]. Their study assessed breast cancer, revealing that RF achieved an average precision of 92.2% using the Information Extraction approach with the Evolution Training data mining software tool. K. Anastraj et al. conducted a comparative study of various machine learning methods using the Wisconsin Breast Cancer (original) datasets, including backpropagation networks, ANN, CNN, and SVM [5]. They employed a deep and convolutional neural network with ALEXNET to

extract and analyze features from benign and cancerous tissues. SVM emerged as the top-performing method with an accuracy rate of 94%. Begüm Erkal and Tülin Erçelebi Ayyldz classified breast cancer using the Breast Cancer Wisconsin (Original) open dataset, utilizing NB, BayesNet, K-NN, SVM, MLP, RF, and LR [6]. Their findings indicated that BayesNet was the most accurate classification method, achieving an accuracy rating of 97.13%. Ch. Shravya et al. examined prediction models developed through LR, SVM, and KNN using data retrieved from the UCI Repository [7]. Their aim was to create prediction models that could forecast actual diseases accurately labeled technique. Their analysis showed that SVM achieved the highest accuracy, with a rate of 92.7%. This comprehensive analysis highlights a variety of machine learning approaches used in breast cancer prediction, emphasizing their performance and potential for early disease detection and treatment. It represents a valuable contribution to the healthcare field and holds the promise of improving outcomes for breast cancer patients.

2.3. COMPARATIVE ANALYSIS AND SUMMARY

The analysis showed in Table 2.1.

Table 2.1: Comparative analysis

Authors	Methodology	Dataset	Results
Shamrat et al.	Machine Learning Algorithms	Breast cancer Wisconsin dataset taken from UCI	97.07%, with NB and RF
Mumine Kaya Keles	Machine Learning Algorithms	Breast cancer Wisconsin dataset taken from UCI	Bagging, IBk, Random Committee, Random Forest, and SimpleCART algorithms were the most successful algorithms, with over 90% accuracy.
K. Anastraj et al	Machine Learning Algorithms	Breast cancer Wisconsin dataset taken from UCI	RF 92.2%
Begüm Erkal and Tülin Erçelebi Ayyldz	Machine Learning Algorithms	The dataset, available on Kaggle.	SVM 94%
Our proposed method	Machine Learning Algorithms and Ensemble classifiers	The dataset, available on Kaggle.	KNNB with 99.42%

2.4. SCOPE OF THE PROBLEM

The task at hand included rationalizing and expediting the process of diagnosing breast cancer illness. Our goal was to get the highest accuracy possible with our recommended model, taking into account the many machine learning research that are associated with it. Our objective was to implement the concept by using basic technologies to simplify the diagnosis of coronary artery disease and so minimize complexity, even if there was not much room for improvement in the present approach.

2.5 CHALLENGES

We obtained the dataset from Kaggle [8-11], and it proved to be highly accessible and user-friendly. Upon concluding the data collection, we meticulously examined the dataset for any missing values. We decided to exclude two anonymous columns that didn't serve any useful purpose. Our attention to detail ensured that our dataset is exceptionally accurate and complete.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 RESEARCH SUBJECT AND INSTRUMENT

To maximize the accuracy of the dataset, we applied many hybrid models and algorithms. This process needed a number of tools, including efficient setup tools that used the greatest GPUs on the market. Python was the primary programming language utilized, and Jupyter Notebook, Google Colab, and Anaconda were among the tools used. These browser-based platforms made it possible for Python applications to be developed and executed without any issues.

3.2 DATA COLLECTION PROCEDURE

The data was stored in kaggle nearly prepared for utilization. Dataset A has 4024 rows,16 columns, dataset B contains 569 columns and 32 rows, dataset C contains 683 columns and 11 rows and dataset D contains 4000 columns and 10 rows. The columns “Status”, “diagnosis”, “Class” and “Classification” are responsible for categorizing breast cancer prevalence. Each attribute in the dataset played a vital role in breast cancer prediction, where patients were classified into two groups. The datasets were then divided into two subsets: one for testing (20%) and the other for training (80%).

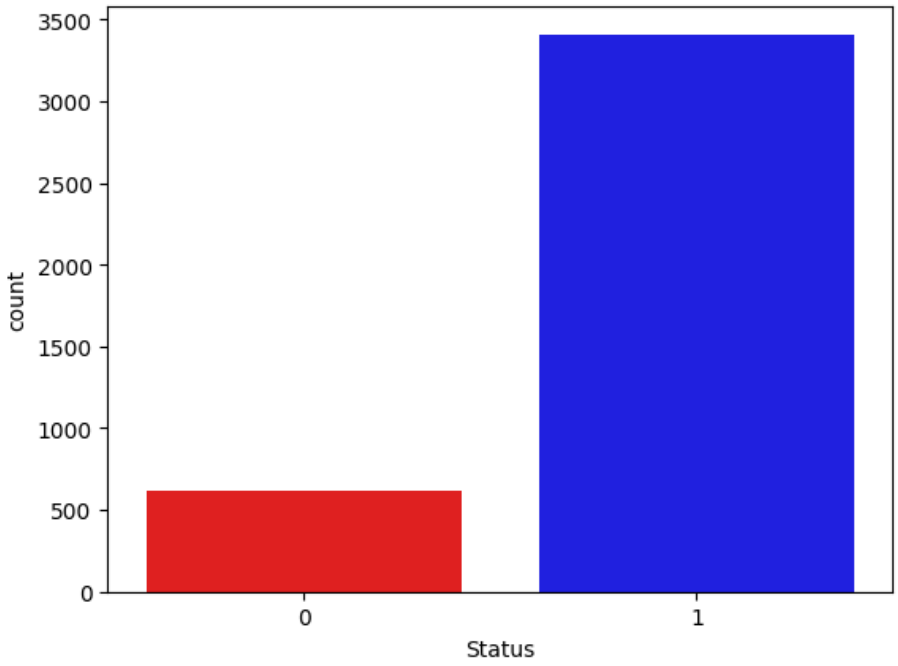


Figure 3.1: Count plot of Dataset A

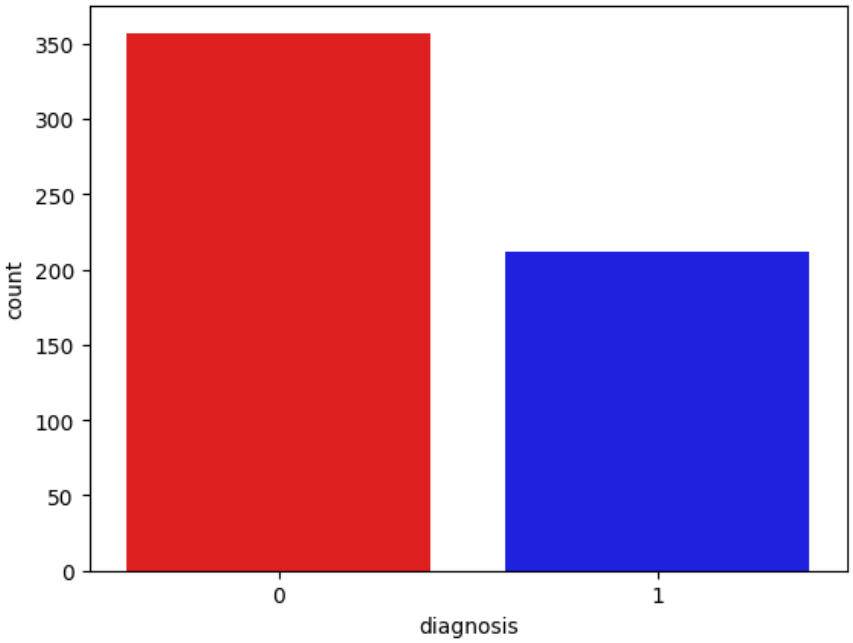


Figure 3.2: Count plot of Dataset B

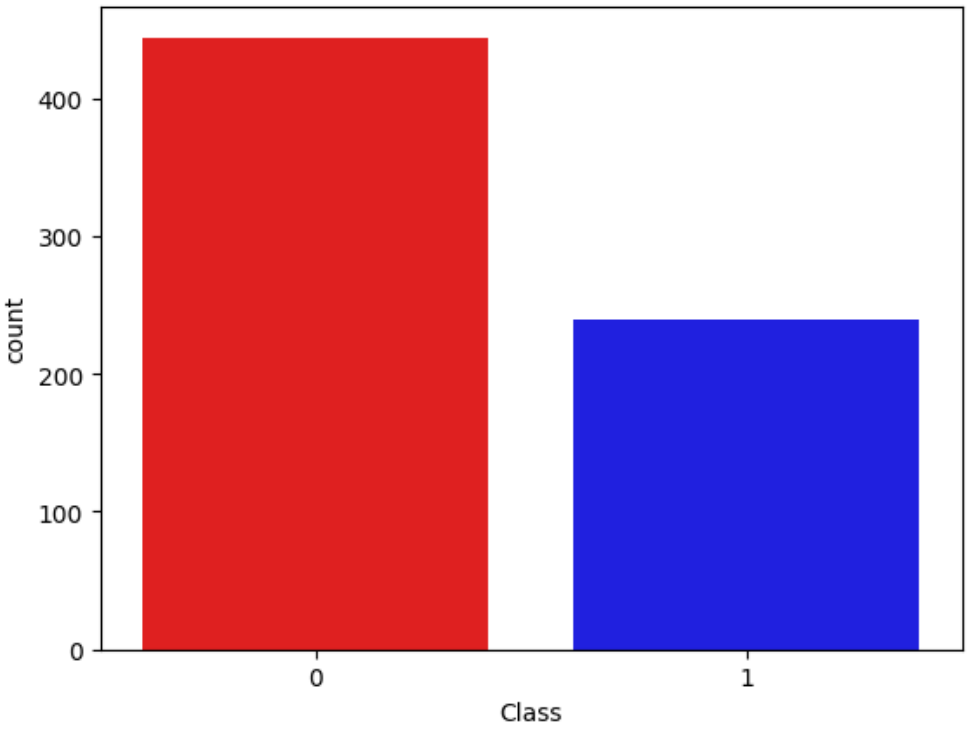


Figure 3.3: Count plot of Dataset C

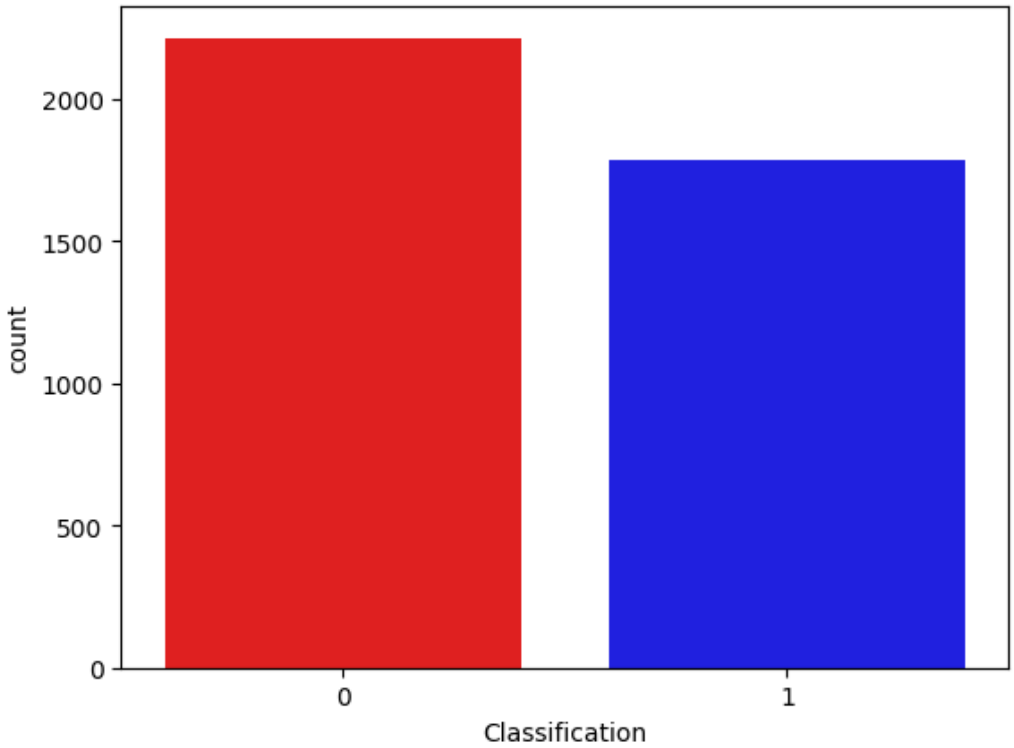


Figure 3.4: Count plot of Dataset D

3.2.1 Categorical Data Encoding

The process of converting nominal values from categorical data is known as the "data encoding system of categorical method". The category encoding approach becomes crucial in machine learning since input and output are often numerical. Using the categorical data encryption approach in our research, which involved columns like target column, was crucial.

3.2.2 Missing Value Imputation

Imputation is the process of substituting values from another dataset with research-derived values to fill in the blanks or missing data. That being said, the fact that our dataset had two null values is notable and promising. We have replaced 0 in null values to handle them.

3.2.3 Handling Imbalanced Data

Through the process of incrementally adding new samples to the dataset, the original data's category distribution is altered. The representation of minority data is improved when the entire dataset is used as input. We used ADASYN technique to balance the imbalance dataset Hungarian.

3.2.4 Feature Scaling

It is a technique for homogenizing a range of dissimilar independent datasets. The Standard Scalar measure scales all relevant data to the interval [0, 1] when there are no local variables. However, in the event that local variables are available, it scales the relevant data to the range [-1, 1].

3.3 STATISTICAL ANALYSIS

We harnessed the power of four datasets hosted on the Kaggle platform and integrated various classifiers, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), K-Nearest Classifier (KNN), among others, into our model. The results were promising, with the KNN achieving a noteworthy test accuracy of 81.14% for Dataset A, KNN of 97.2% for dataset B, KNN of 98.85% for dataset C and LR of 96.125% for

dataset D. Furthermore, Bagging KNN also demonstrated accuracy matching this high standard of 99.42%.

3.4 Proposed Methodology

Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), K-Neighbors Classifier (KNN), XGB Classifier (XGB) and Gradient Boosting Classifier (GB) were used in our proposed model.

3.4.1 Traditional Classifiers

LOGISTIC REGRESSION

Modeling the likelihood of a binary result using one or more predictor variables, logistic regression is a method of statistics used for binary classification. Typically coded as 0 or 1, logistic regression predicts the chance that an instance belongs to a certain category, as opposed to linear regression's prediction of a continuous result. This is accomplished by converting the linear sum of the given input variables into an amount between 0 and 1 using the logistic function, commonly referred to as the sigmoid function. The model provides coefficients that indicate the influence of each predictor and calculates the likelihood that the dependent event will occur [12-13]. In order to predict the existence of diseases, consumer behavior, and other things, logistic regression is often utilized in the social sciences, medical, and machine learning domains.

RANDOM FOREST

For problems involving regression and classification, an ensemble learning technique called a random forest is commonly employed. It functions by building many decision trees in training and producing the mean prediction (regression) or mode of grouping (classification) of each individual tree. The "forest" is constructed using a process known as bagging (Bootstrap Aggregating), in which a decision tree is taught on each of the dataset's many subsets that are produced via replacement. This method averages out variances and biases from individual trees, which helps to decrease overfitting and increase the model's generalization [14-15]. The usefulness of random forests in producing precise

forecasts without requiring substantial parameter adjustment is well-known, as is their robustness and capacity to handle big datasets with increased dimensionality.

GRADIENT BOOSTING

For regression and classification applications, gradient boosting is a potent machine learning approach that creates an ensemble of weak learners, usually decision trees. In order to fix the mistakes produced by earlier models, iteratively adding new models while concentrating on the data points that are most difficult to forecast correctly is how it operates. Every new model is trained to gradually minimize the total prediction error by reducing the residual errors of the aggregate ensemble. Because gradient boosting uses a sequential technique, it may produce very precise prediction models. This makes it very useful for managing complicated datasets and identifying minute trends. To avoid overfitting and maximize performance, nevertheless, meticulous hyperparameter djustment is necessary [16-17].

K-NEAREST

A straightforward, non-parametric, and user-friendly technique for classification and regression problems in machine learning is the k-Nearest Neighbors (k-NN) classifier. It functions by finding the k instances in the training dataset that are the closest to a given input (neighbors), and then predicting the output based on the average value (for regression) or majority class (for classification) of these neighbors. Distance measures like Euclidean distance are commonly used to quantify similarity. Even though k-NN is straightforward, it can be effective, particularly for short datasets or in situations where the decision border is extremely irregular. It is, however, sensitive to the choice of k and the distance measure and computationally costly for big datasets. For data to perform better, feature selection and data scaling must be done correctly [18].

XGB CLASSIFIER

The XGBoost (Extreme Gradient Boosting) classifier is a powerful and efficient implementation of the gradient boosting framework, specifically designed for supervised

learning tasks. It excels in predictive performance by combining an ensemble of weak prediction models, typically decision trees, to create a robust predictive model. XGBoost offers several key advantages, including handling missing values, incorporating regularization to prevent overfitting, and leveraging advanced tree learning algorithms that optimize speed and performance. Its scalability and flexibility make it a popular choice for various applications, from structured/tabular data to more complex domains, offering superior accuracy and efficiency in comparison to many traditional machine learning algorithms [19].

3.4.2 ENSEMBLE METHODS OF MACHINE LEARNING

For bagging, boosting, stacking, voting, and random subspace, we used ensemble models [20].

BAGGING

Bootstrap Aggregating, or Bagging, is a strategy for ensemble learning that aims to increase the precision and resilience of machine learning models. It entails using several subsets of the training data produced by bootstrap sampling—random sampling with replacement—to train numerous base models, usually decision trees. Every model undergoes separate training, and for classification tasks, the predictions are aggregated by a majority vote, and for regression tasks, through averaging. Bagging, as opposed to using a single model, produces more consistent and dependable performance by combining the predictions of many models, which lowers variance and helps prevent overfitting. One of the most well-known uses of this technique is Random Forest, which expands bagging by adding more randomization to the characteristics chosen for each split in the trees [21].

BOOSTING

An effective machine learning method that builds a strong overall model by combining several weak classifiers is called a boosting ensemble classifier. Boosting is based on the sequential training of classifiers, where each new model learns from the mistakes made by the prior ones. This is usually accomplished by giving misclassified cases larger weights, which incentivizes the subsequent classifier in the series to fix the errors. AdaBoost and

Gradient Boosting are two popular boosting methods that iteratively modify the weights of the classifiers and training data. Boosting improves the final classifier's accuracy and resilience by combining the predictions of all the models, frequently beating the performance of individual models, especially in situations with complicated patterns and noisy data [22].

STACKING

An effective machine learning method that mixes many base models to increase prediction accuracy is the stacking ensemble classifier. This approach involves training several learning algorithms on the same dataset so they can each generate unique predictions. The final prediction is then produced by a second-level meta-model using these predictions as input attributes. By leveraging the strengths and compensating for the weaknesses of the individual base models, stacking often results in superior performance compared to any single model. This approach effectively reduces overfitting and bias, making it particularly useful in complex tasks where no single model performs optimally across all scenarios [23].

VOTING

A machine learning model called a voting ensemble classifier aggregates the predictions of several different classifiers to increase overall resilience and accuracy. The way the ensemble functions is by using a voting mechanism, which can be either soft or hard voting, to aggregate the predictions of its component models. Hard voting selects the class with the majority of votes as the final forecast, with each classifier voting for a particular class label. During the soft voting process, the class with the greatest average probability is selected by averaging the projected probabilities of each class from all classifiers. This approach leverages the strengths of diverse models, reducing the likelihood of errors and increasing the performance, especially in complex and noisy datasets. By pooling the decisions of several models, a voting ensemble can achieve better generalization compared to any individual model alone [24-25].

3.4.3 Flow chart of proposed methodology:

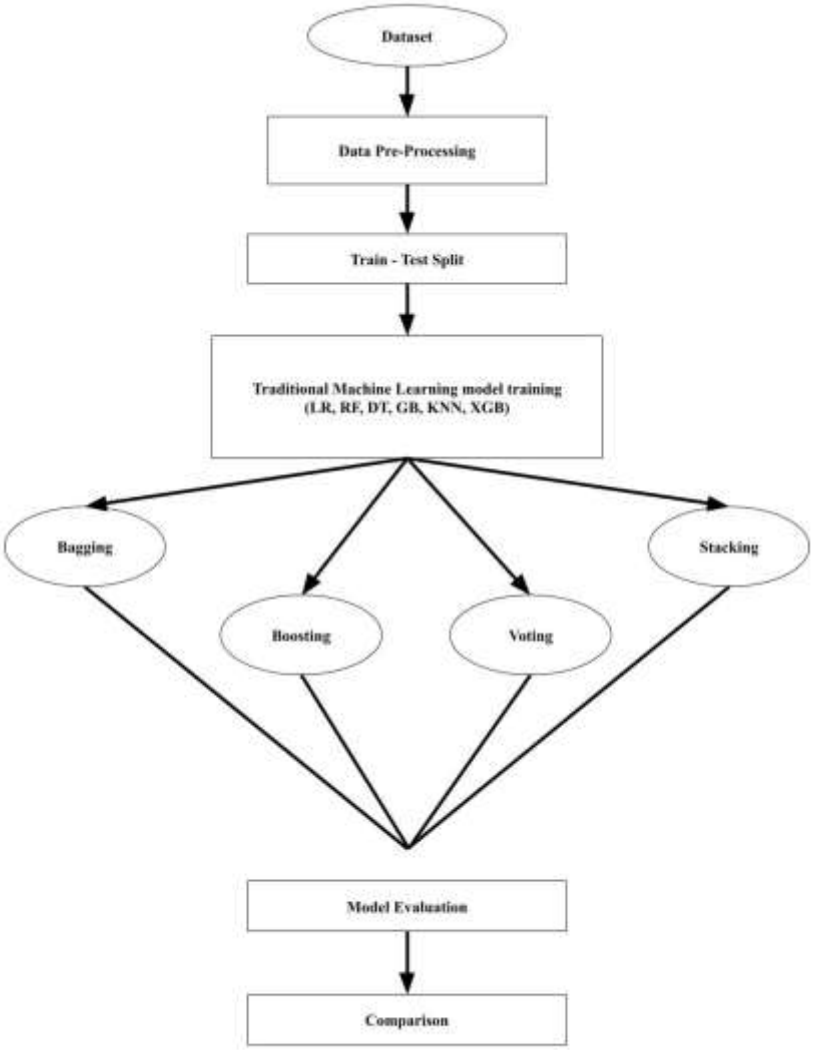


Figure 3.5: Methodology of Breast Cancer Disease Prediction

The research utilizes four datasets sourced from Kaggle. First, the dataset we selected was put into use. By locating and updating any inaccurate or missing numbers, we guaranteed the integrity of the data. Next, we employed a range of algorithms and assessed their performance. It employs a variety of algorithm models, including Gradient Boosting (GB), K-Neighbors Classifier (KNN), XGB Classifier (XGB), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT), to predict breast cancer disease risks. To enhance prediction accuracy, the study also incorporates ensemble models such as Bagging, Boosting, Stacking, and Voting [26].

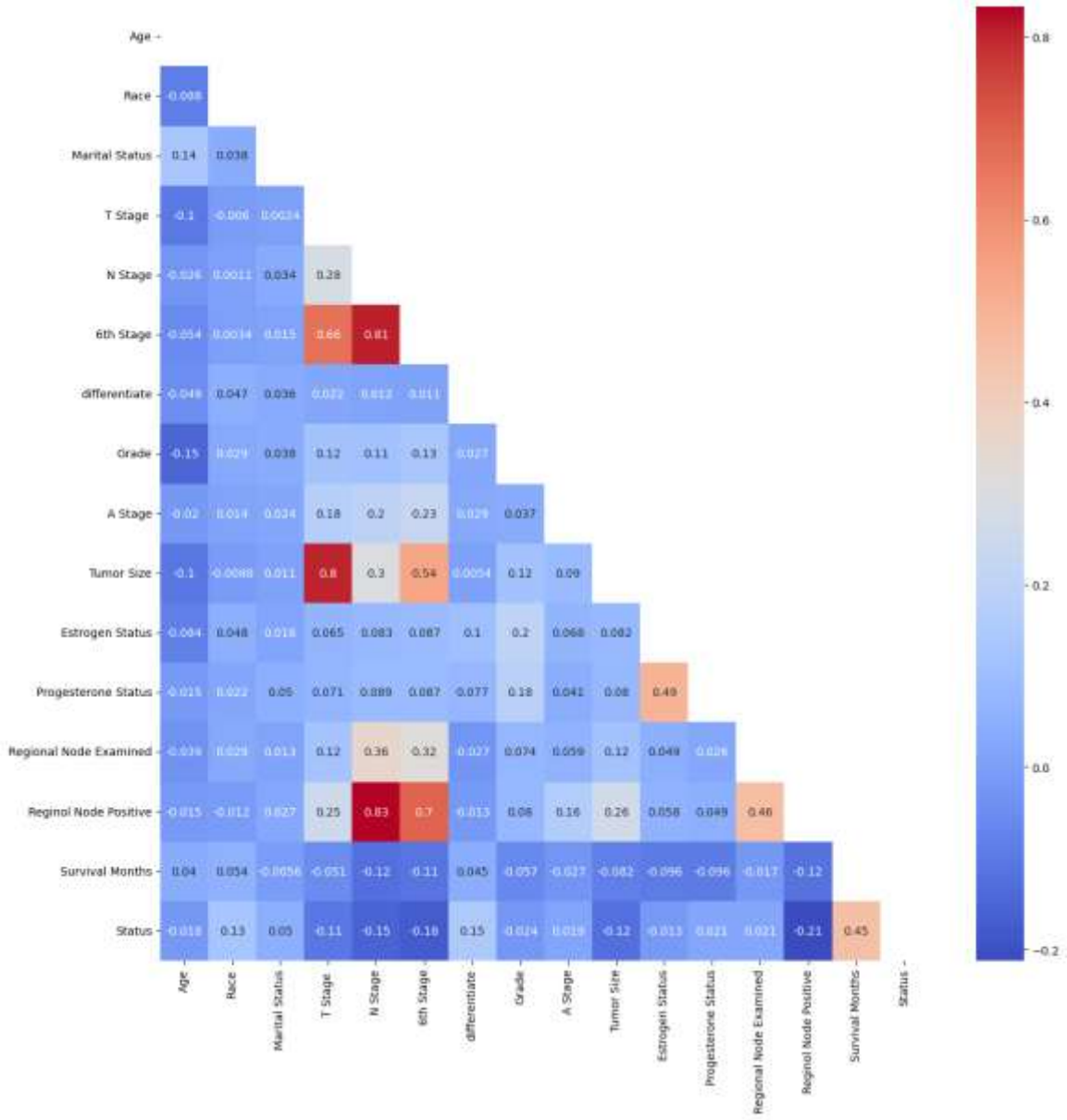


Figure 3.6: Correlated Features of Dataset A

A correlation subplot illustrates the finding of interdependencies between variables that change how they relate to one other. The likelihood that one may be anticipated from the other increases with the degree of interdependence between the variables. This methodology denotes a more profound comprehension of the dataset, augmenting our capacity to discern pivotal elements [27-28]. Every connected quality that was associated with the anticipated characteristic "target" was displayed in Fig. 3.6, 3.7, 3.8 and 3.9 for dataset A, dataset B, dataset C and dataset D.

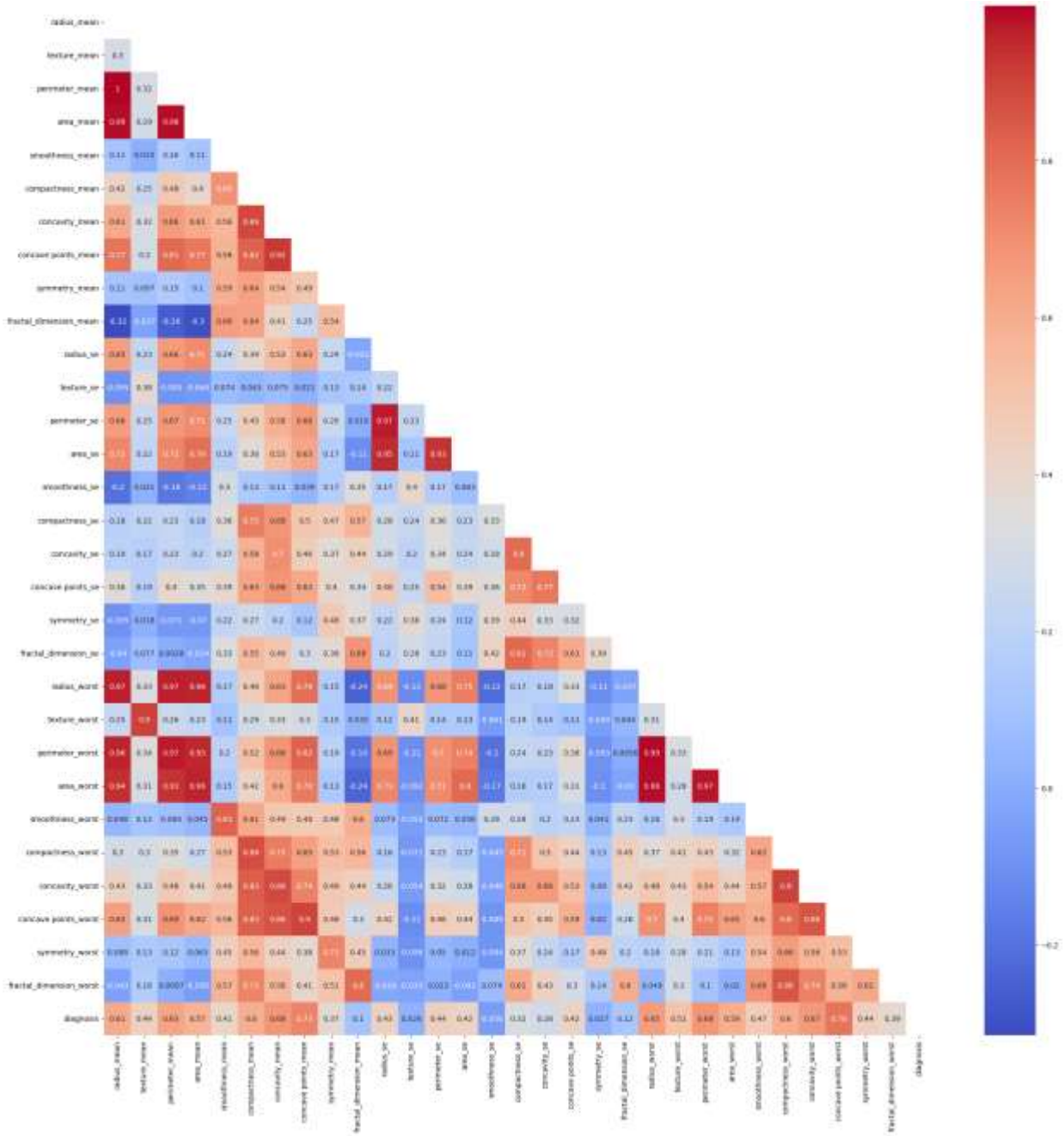


Figure 3.7: Correlated Features of Dataset B

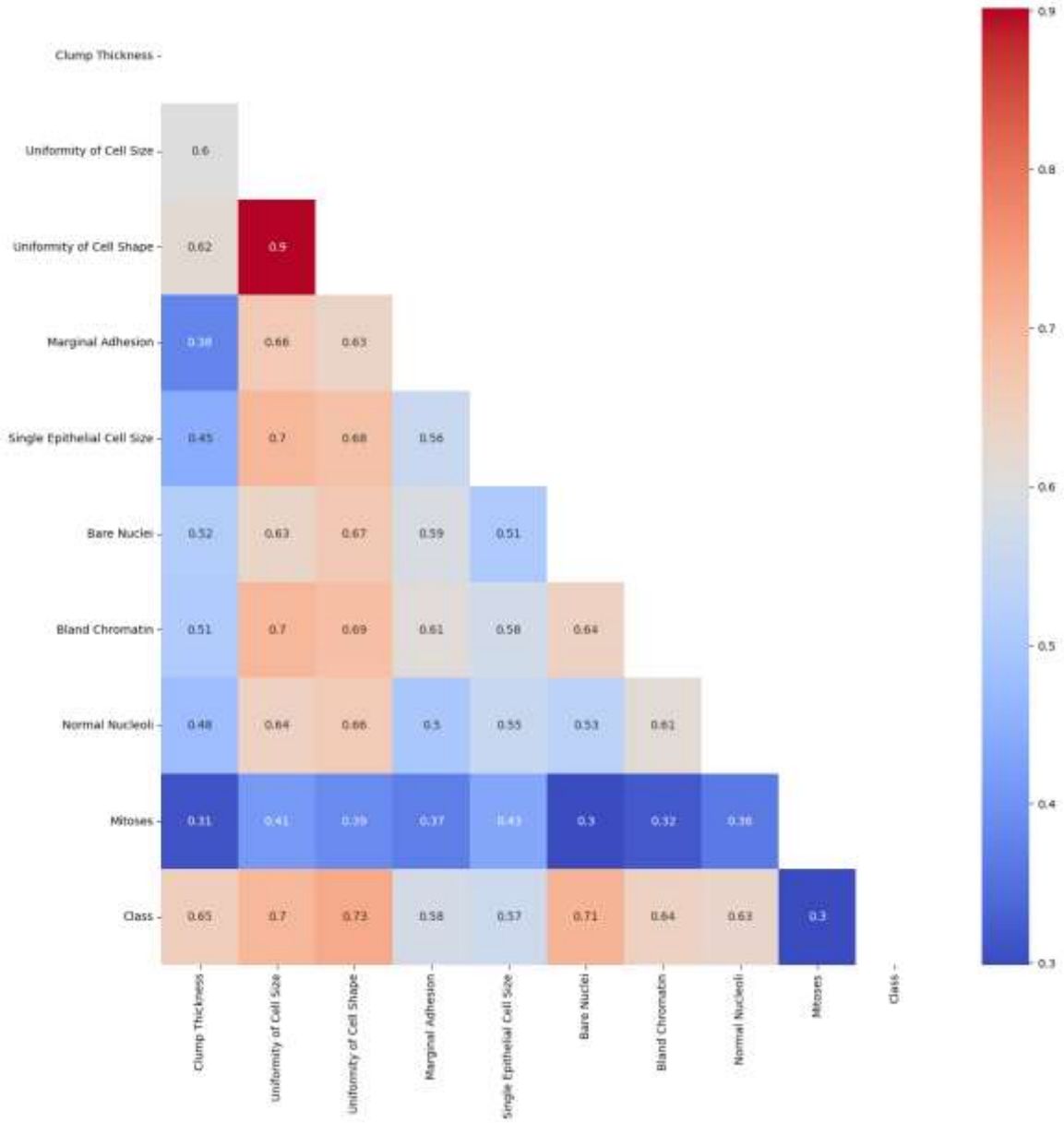


Figure 3.8: Correlated Features of Dataset C

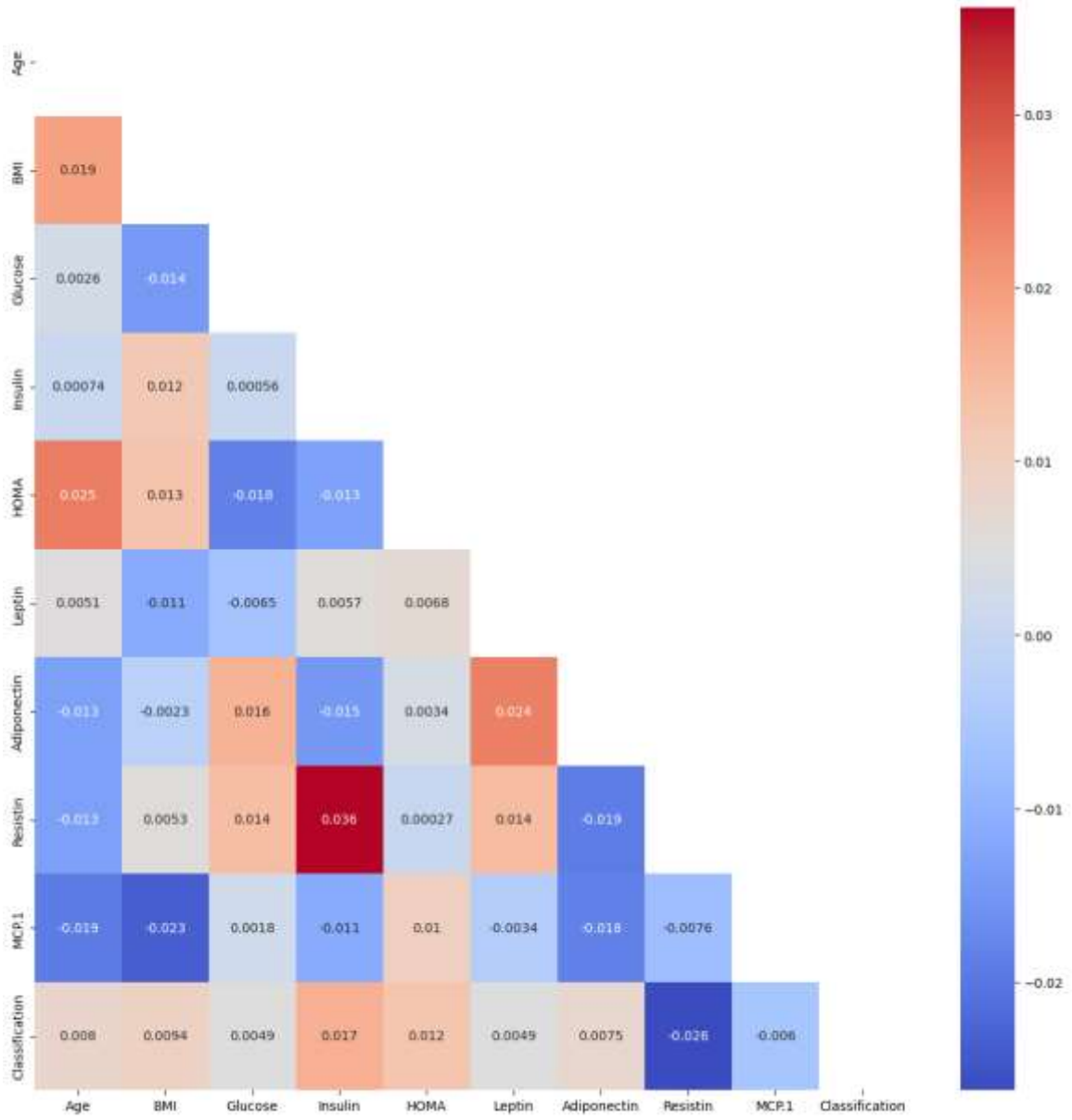


Figure 3.9: Correlated Features of Dataset D

3.5 IMPLEMENTATION REQUIREMENTS

Effectively apply, reliable datasets. Data cleaning is a crucial step to ensure that our model operates smoothly. The dataset underwent several filtering procedures to remove inconsistencies and outliers. We utilized the Standard Scaler Transform to standardize the data, converting categorical data into numerical values. The dataset was then divided into 80% for training and 20% for testing.

Subsequently, we implemented various machine learning algorithms and assessed their performance. To achieve best predictive score, applied techniques such as Bagging, Stacking, Boosting, and Voting. The results from these techniques were carefully evaluated. We further validated our results through hyperparameter tuning.

To complete the learning process, data analysis was conducted. This involved fitting forecasting algorithms and applying model learning techniques. By using ensemble methods such as bagging, boosting, stacking, and voting, we aimed to maximize predictive accuracy. We selected the best-performing model for implementation based on the results.

This comprehensive process ensures that our proposed model can effectively predict breast cancer and contribute to early diagnosis and intervention. It involves a thorough data analysis, model evaluation, and optimization to achieve the best results.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 EXPERIMENTAL SETUP

The supervised learning technique used in this work is based on training and assessment. Using the training datasets, a classification model is constructed as part of the procedure. The result is then obtained by applying the developed model to the testing dataset. In the sections that follow, the machine-learning algorithm will be covered in more detail.

4.2 EXPERIMENTAL RESULT & ANALYSIS

It was now essential to evaluate the models that were already in place. We used performance assessment metrics to gauge our suggested model's effectiveness. These techniques use data that hasn't been seen to assess overall performance. This section contains an analysis report that was created using our targeted coronary artery disease dataset and the experimental outputs of machine learning models. First, the dataset we selected was put into use. By locating and updating any inaccurate or missing numbers, we guaranteed the integrity of the data. Next, we employed a range of algorithms and assessed their performance. For classic methods like Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), K-Neighbors Classifier (KNN), XGB Classifier (XGB), and Gradient Boosting Classifier (GB), confusion matrices were produced. We also experimented with several ensemble methods, evaluating their effectiveness using confusion matrices.

ACCURACY

The percentage of correct predictions generated from testing data is covered by the measure. Comparing the expected results with the actual measurements, based on a single variable, yields the accuracy. This basic measurement methodology is one of the easiest ways to evaluate a model because it mostly concentrates on deliberate mistakes. Making sure that models are accurate is an essential part of what I do.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Precision

The accuracy, or the percentage of positively predicted observations that really occurred, is examined in this section. Precision is the actual percentage of all cases that were correctly predicted to be true. It's crucial to remember that a high recall rate might be deceiving for any kind of model.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall

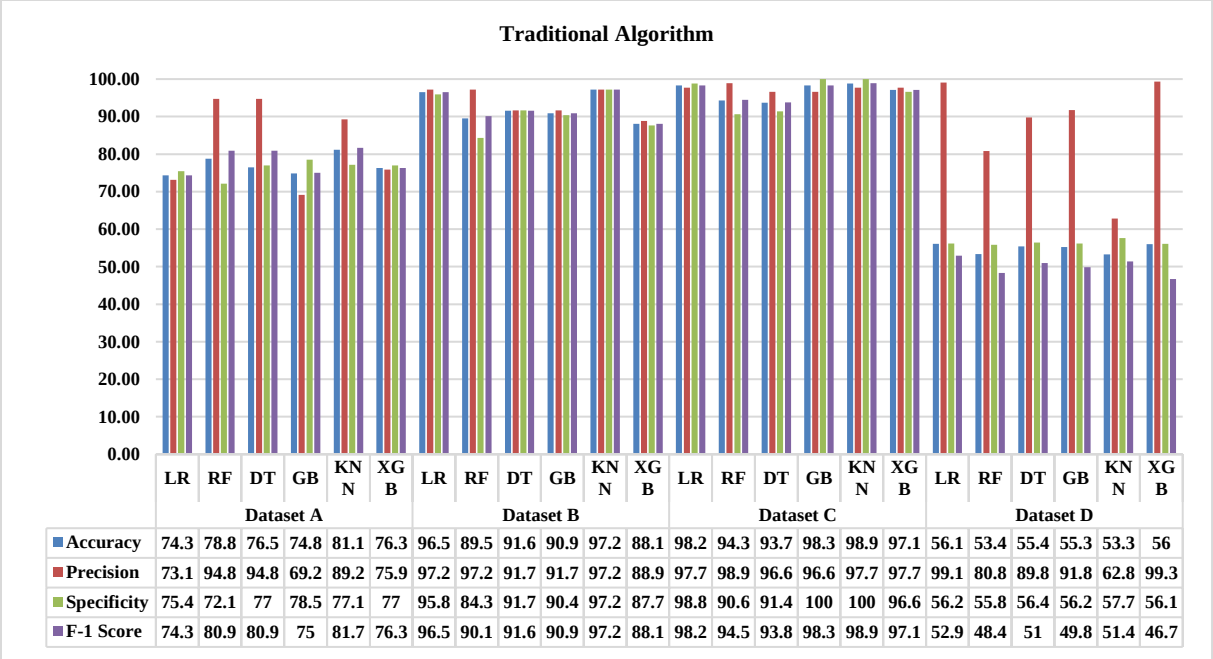
This has to do with the expectedly positive ratio of data from a model. It's important to remember, though, that great accuracy can occasionally be deceptive. In this context, recall—typically defined as the ratio of all positive labels to expected positives—is an important parameter.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

F-1 Score

Recall and accuracy ratios are two important metrics that are highlighted in this section along with the harmonic and accuracy techniques of recall. It is mentioned that the model could not function properly if the mean of the harmonic measurements is low.

$$F - 1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$



	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN	XGB
	Dataset A						Dataset B						Dataset C						Dataset D					
Accuracy	74.3	78.8	76.5	74.8	81.1	76.3	96.5	89.5	91.6	90.9	97.2	88.1	98.2	94.3	93.7	98.3	98.9	97.1	56.1	53.4	55.4	55.3	53.3	56
Precision	73.1	94.8	94.8	69.2	89.2	75.9	97.2	97.2	91.7	91.7	97.2	88.9	97.7	98.9	96.6	96.6	97.7	97.7	99.1	80.8	89.8	91.8	62.8	99.3
Specificity	75.4	72.1	77	78.5	77.1	77	95.8	84.3	91.7	90.4	97.2	87.7	98.8	90.6	91.4	100	100	96.6	56.2	55.8	56.4	56.2	57.7	56.1
F-1 Score	74.3	80.9	80.9	75	81.7	76.3	96.5	90.1	91.6	90.9	97.2	88.1	98.2	94.5	93.8	98.3	98.9	97.1	52.9	48.4	51	49.8	51.4	46.7

Figure 4.1: Results of Traditional Classifiers

The evaluation of traditional algorithms for breast cancer prediction across four datasets (A, B, C, and D) reveals significant differences in performance metrics such as accuracy, precision, specificity, and F-1 score. In Dataset A, algorithms like KNN (81.14%) and RF (78.79%) outperformed others in terms of accuracy, with KNN achieving a high precision (89.24%) and RF showing a notable F-1 score (80.88%). For Dataset B, KNN stood out with the highest accuracy (97.2%) and precision, indicating strong prediction performance, while LR also performed exceptionally well with an accuracy of 96.50%. Dataset C showed very high accuracies across the board, with KNN (98.85%) and Gradient Boosting (GB, 98.28%) achieving the best results, particularly excelling in specificity and F-1 scores. However, in Dataset D, the algorithms struggled, with relatively low accuracies, such as 56.12% for LR and 53.37% for RF, despite LR and XGB showing high precision (99.1% and 99.33% respectively). These results highlight the varying strengths of algorithms across different datasets, with certain models excelling in specific scenarios, such as KNN and LR in high-accuracy environments, while others like RF and XGB performed better in more challenging cases with lower accuracy but higher precision.

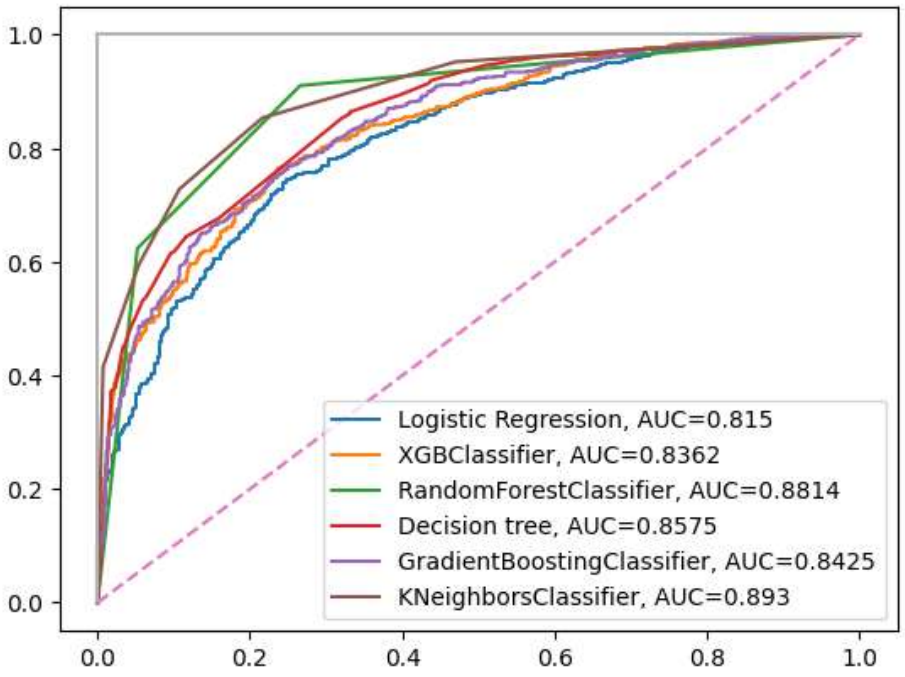


Figure 4.2: Analysis of Traditional algorithms for A dataset (AUC-ROC)

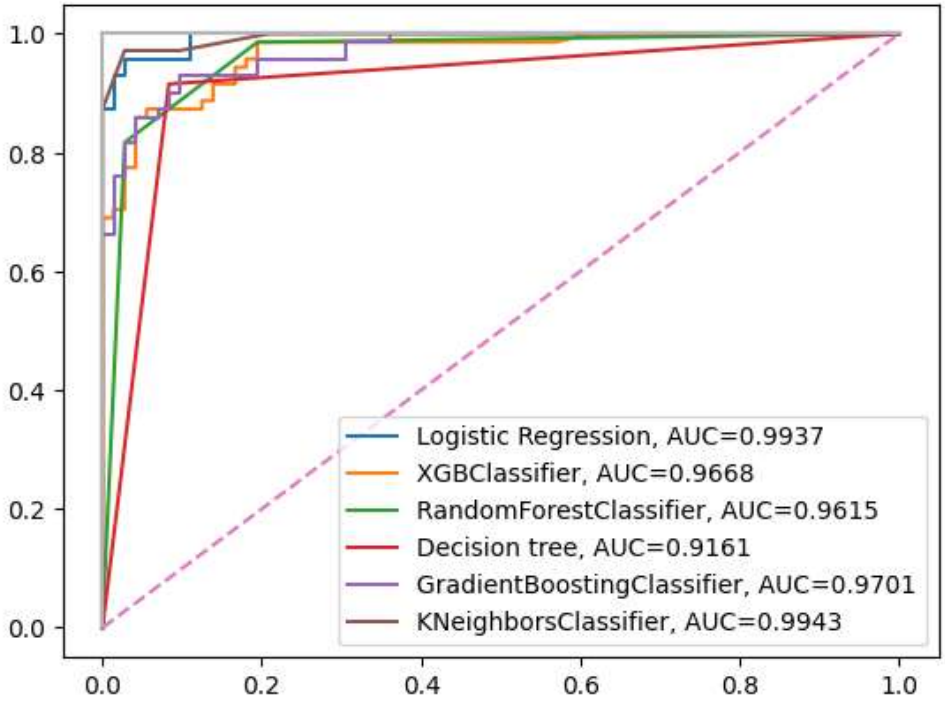


Figure 4.3: Analysis of Traditional algorithms for B dataset (AUC-ROC)

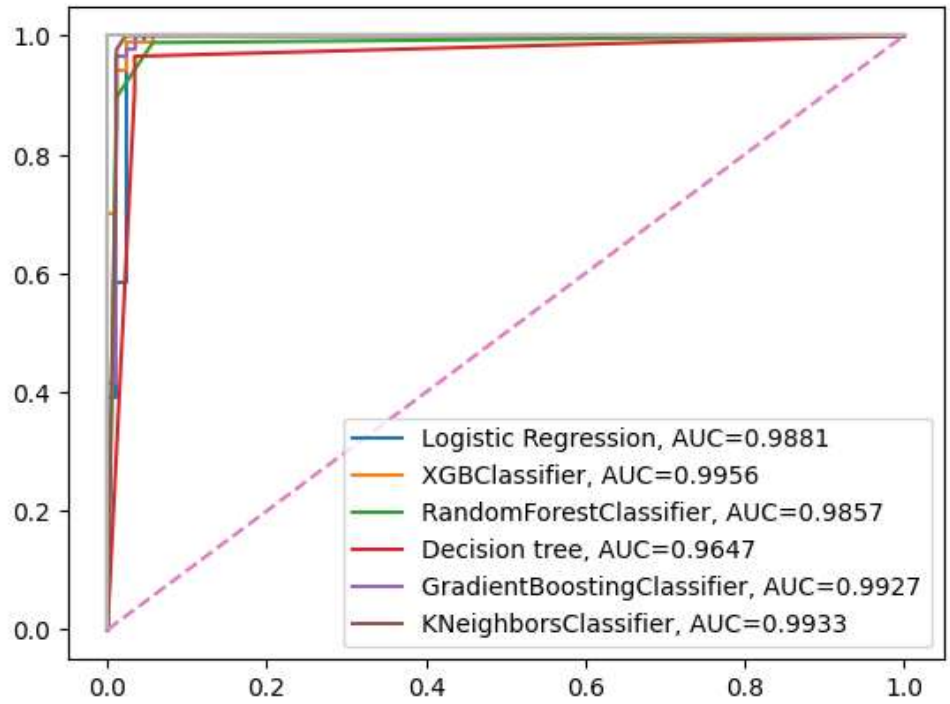


Figure 4.4: Analysis of Traditional algorithms for C dataset (AUC-ROC)

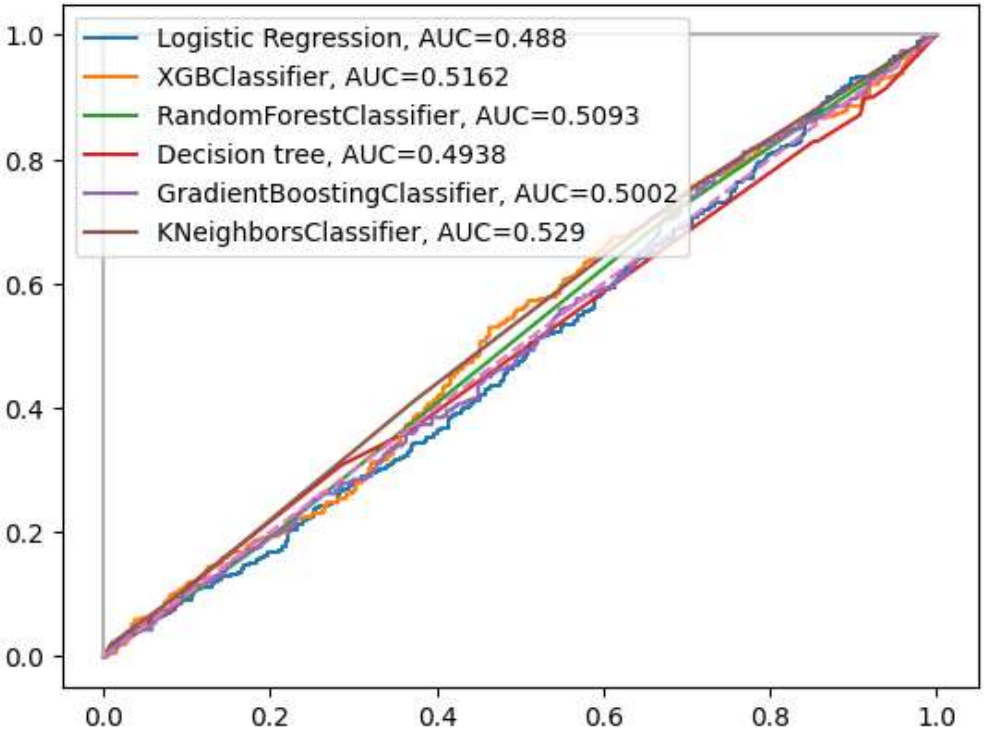
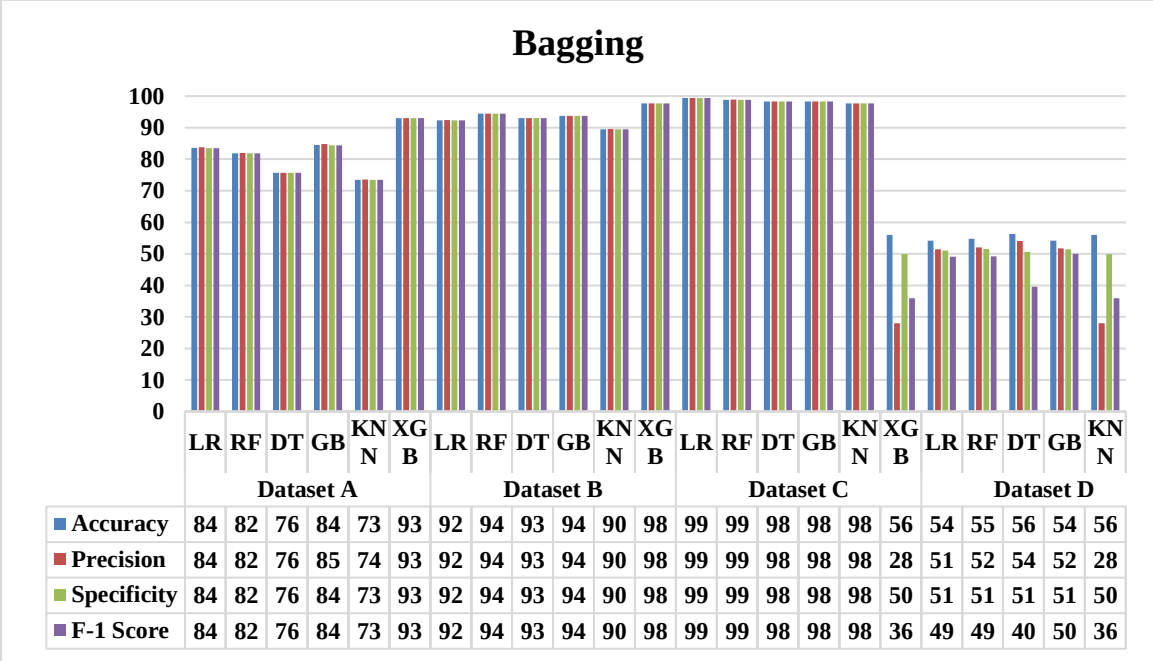


Figure 4.5: Analysis of Traditional algorithms for D dataset (AUC-ROC)



	Dataset A						Dataset B						Dataset C						Dataset D				
	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN
Accuracy	84	82	76	84	73	93	92	94	93	94	90	98	99	99	98	98	98	56	54	55	56	54	56
Precision	84	82	76	85	74	93	92	94	93	94	90	98	99	99	98	98	98	28	51	52	54	52	28
Specificity	84	82	76	84	73	93	92	94	93	94	90	98	99	99	98	98	98	50	51	51	51	51	50
F-1 Score	84	82	76	84	73	93	92	94	93	94	90	98	99	99	98	98	98	36	49	49	40	50	36

Figure 4.6: Experimental Results of Bagging

The evaluation of various bagging algorithms for breast cancer prediction across multiple datasets shows varied performance, with Dataset C generally yielding the highest accuracy and Dataset D the lowest. Random Forest (RF) consistently performs well, especially in Dataset C, achieving a near-perfect accuracy of 99.42%, indicating its strong ability to handle complex data. Decision Trees (DT) and Gradient Boosting (GB) also perform robustly, with DT reaching 98.85% accuracy in Dataset C. In contrast, Dataset D shows much lower predictive performance, particularly for Logistic Regression (LR) and XGBoost (XGB), both of which exhibit poor precision and specificity, indicating challenges in distinguishing between classes. The performance in Dataset A is moderate, with K-Nearest Neighbors (KNN) achieving the highest accuracy (84.46%), while Dataset B also performs well, especially with DT achieving the highest accuracy at 94.4%. Overall, the results suggest that the choice of algorithm and the nature of the dataset significantly affect the prediction accuracy for breast cancer, with ensemble methods like RF and KNN generally performing better across different datasets.

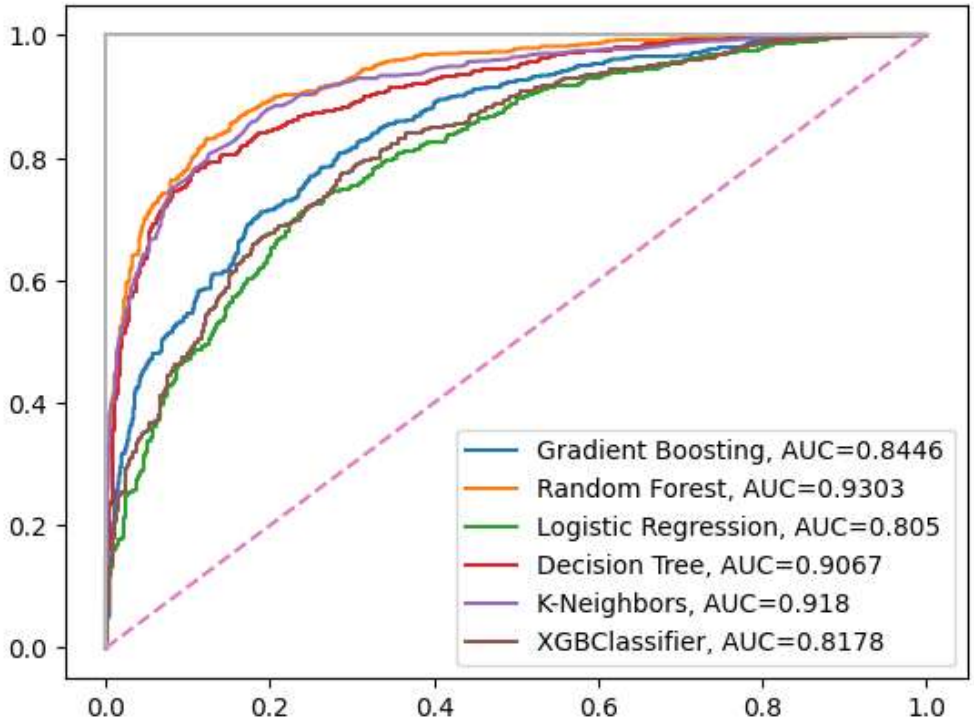


Figure 4.7: Analysis of Bagging for A dataset (AUC-ROC)

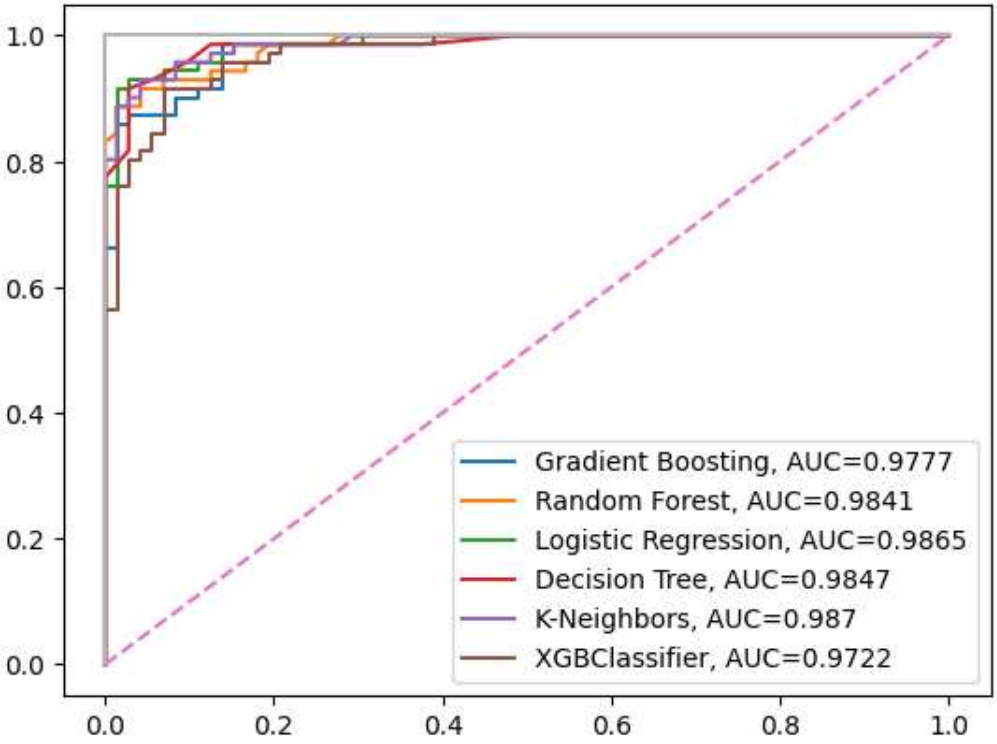


Figure 4.8: Analysis of Bagging for B dataset (AUC-ROC)

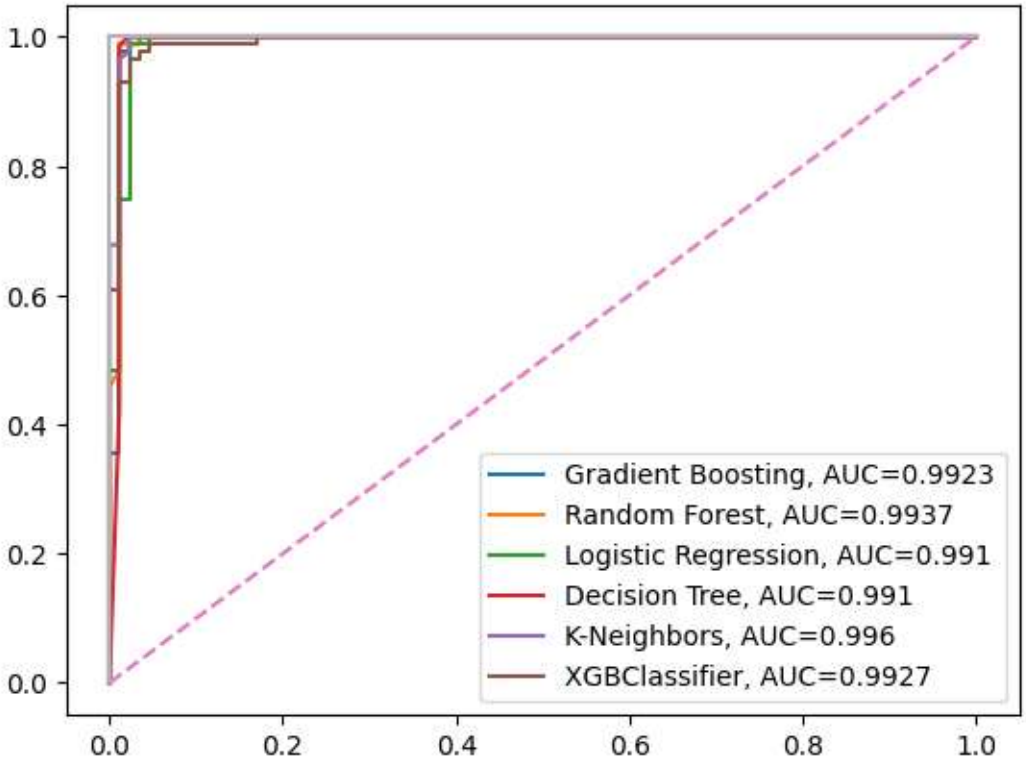


Figure 4.9: Analysis of Bagging for C dataset (AUC-ROC)

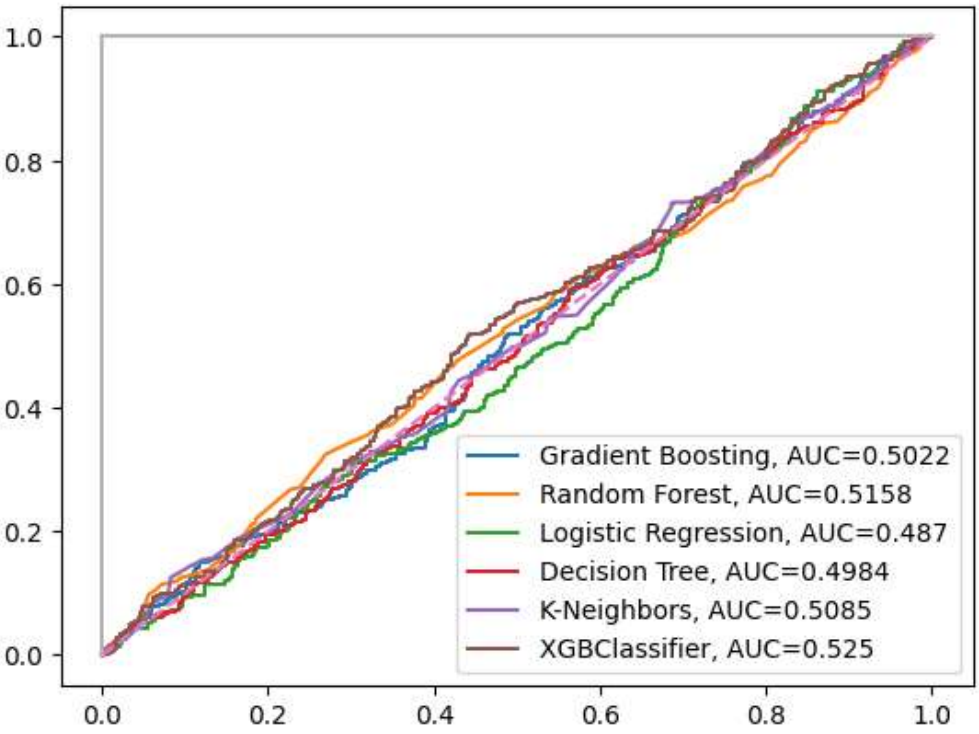
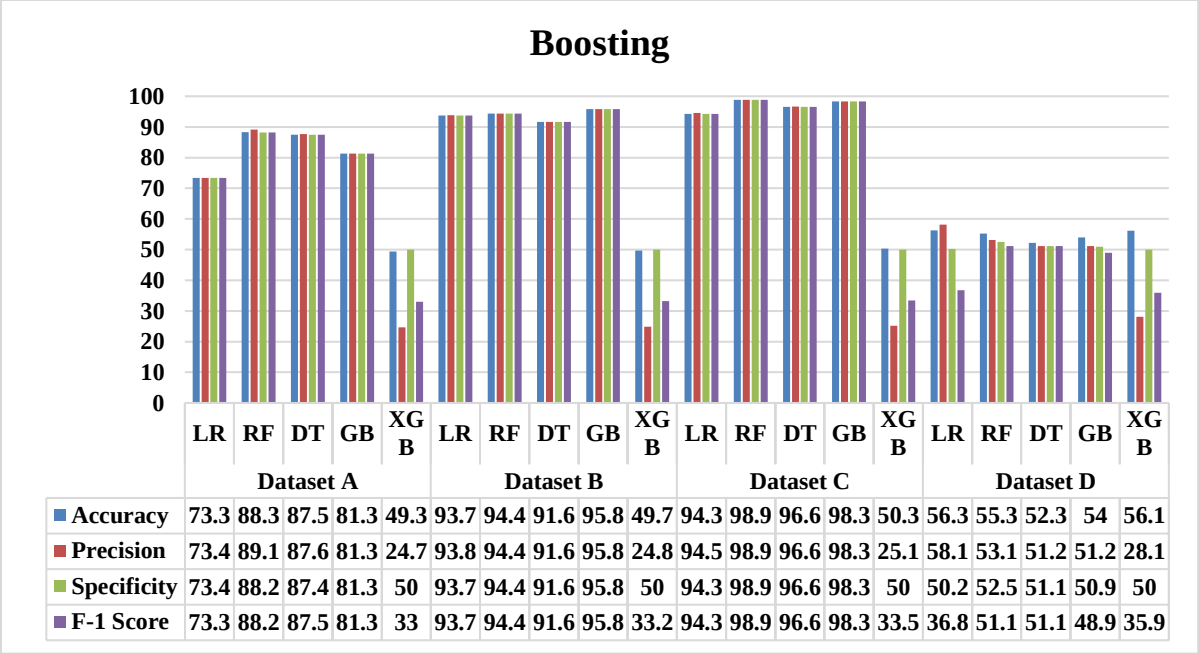


Figure 4.10: Analysis of Bagging for D dataset (AUC-ROC)



	LR	RF	DT	GB	XGB	LR	RF	DT	GB	XGB	LR	RF	DT	GB	XGB	LR	RF	DT	GB	XGB
	Dataset A					Dataset B					Dataset C					Dataset D				
Accuracy	73.3	88.3	87.5	81.3	49.3	93.7	94.4	91.6	95.8	49.7	94.3	98.9	96.6	98.3	50.3	56.3	55.3	52.3	54	56.1
Precision	73.4	89.1	87.6	81.3	24.7	93.8	94.4	91.6	95.8	24.8	94.5	98.9	96.6	98.3	25.1	58.1	53.1	51.2	51.2	28.1
Specificity	73.4	88.2	87.4	81.3	50	93.7	94.4	91.6	95.8	50	94.3	98.9	96.6	98.3	50	50.2	52.5	51.1	50.9	50
F-1 Score	73.3	88.2	87.5	81.3	33	93.7	94.4	91.6	95.8	33.2	94.3	98.9	96.6	98.3	33.5	36.8	51.1	51.1	48.9	35.9

Figure 4.11: Experimental Results of Boosting

The evaluations of boosting algorithms, specifically Gradient Boosting (GB) and XGBoost (XGB), for breast cancer prediction across four datasets reveal varying levels of performance. In Dataset A, GB achieved moderate accuracy (81.29%) and balanced precision, specificity, and F1-score, while XGB underperformed significantly with low accuracy (49.33%) and precision (24.66%), indicating poor predictive capability. In Dataset B, GB excelled with the highest metrics across all measures (95.8%), outperforming XGB, which again had low accuracy and performance (49.65%). In Dataset C, GB delivered near-optimal performance with accuracy and precision exceeding 98%, while XGB remained inefficient with around 50% accuracy. In contrast, Dataset D showed low performance for both GB (54% accuracy) and XGB (56.12%), reflecting difficulties in handling this dataset. Overall, GB consistently outperformed XGB in predictive accuracy and balanced metrics, but both struggled with Dataset D, indicating the complexity of breast cancer prediction in certain cases.

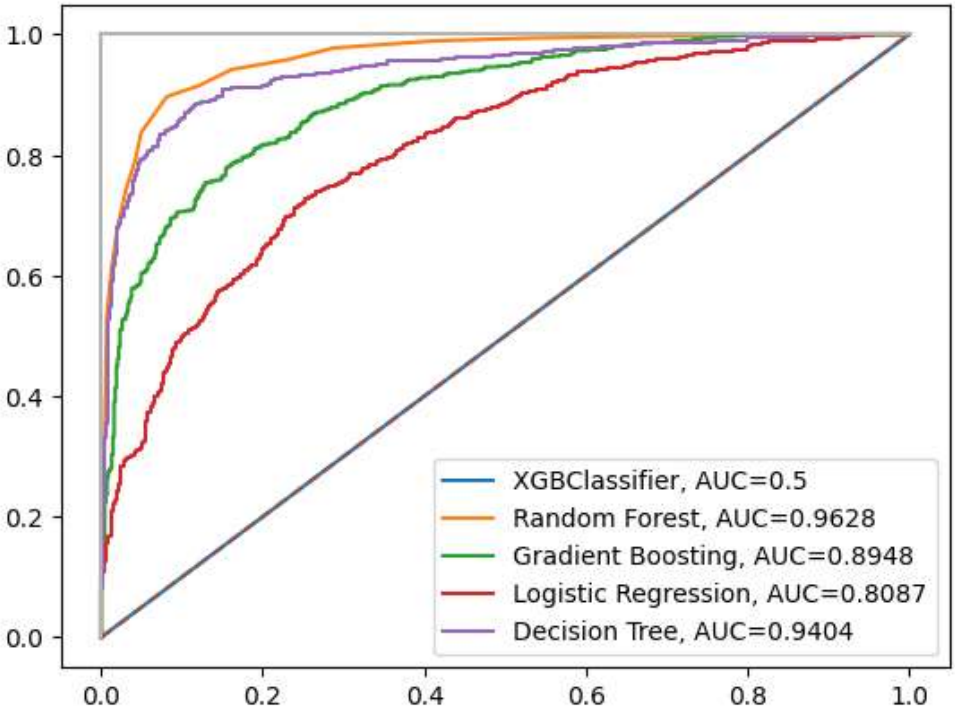


Figure 4.12: Analysis of Boosting for A dataset (AUC-ROC)

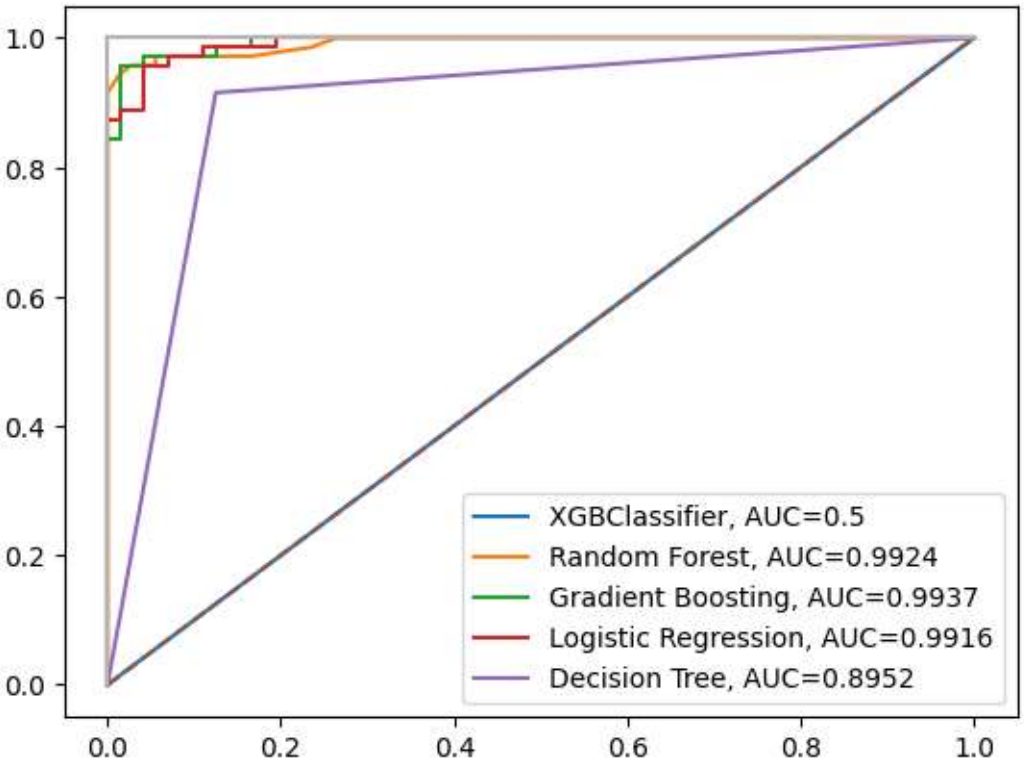


Figure 4.13: Analysis of Boosting for B dataset (AUC-ROC)

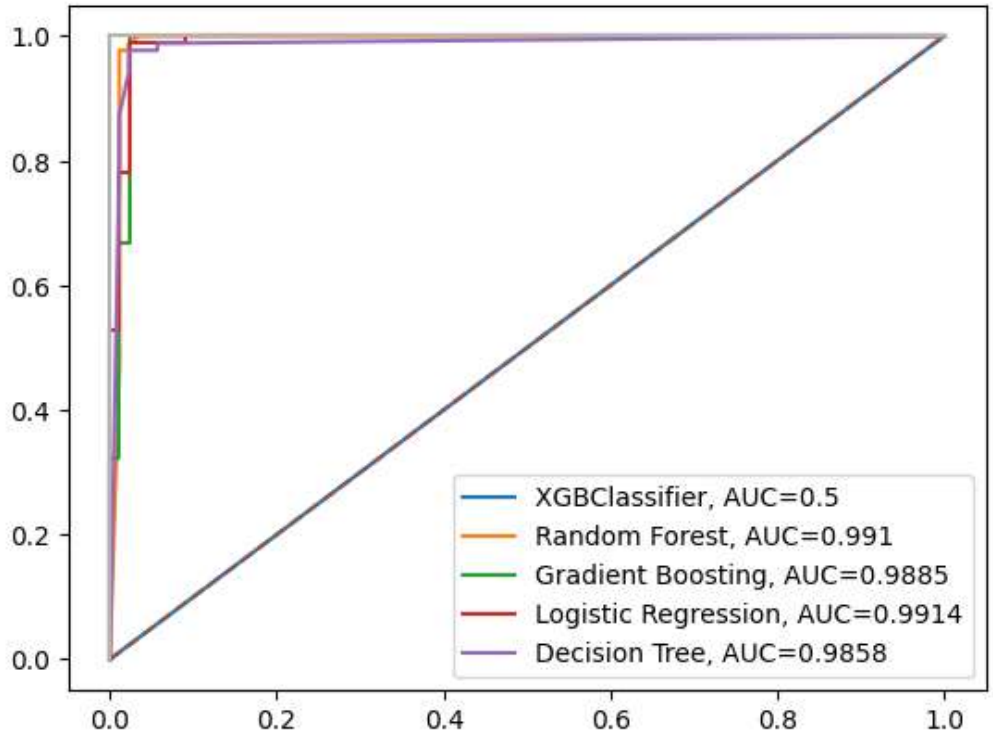


Figure 4.14: Analysis of Boosting for C dataset (AUC-ROC)

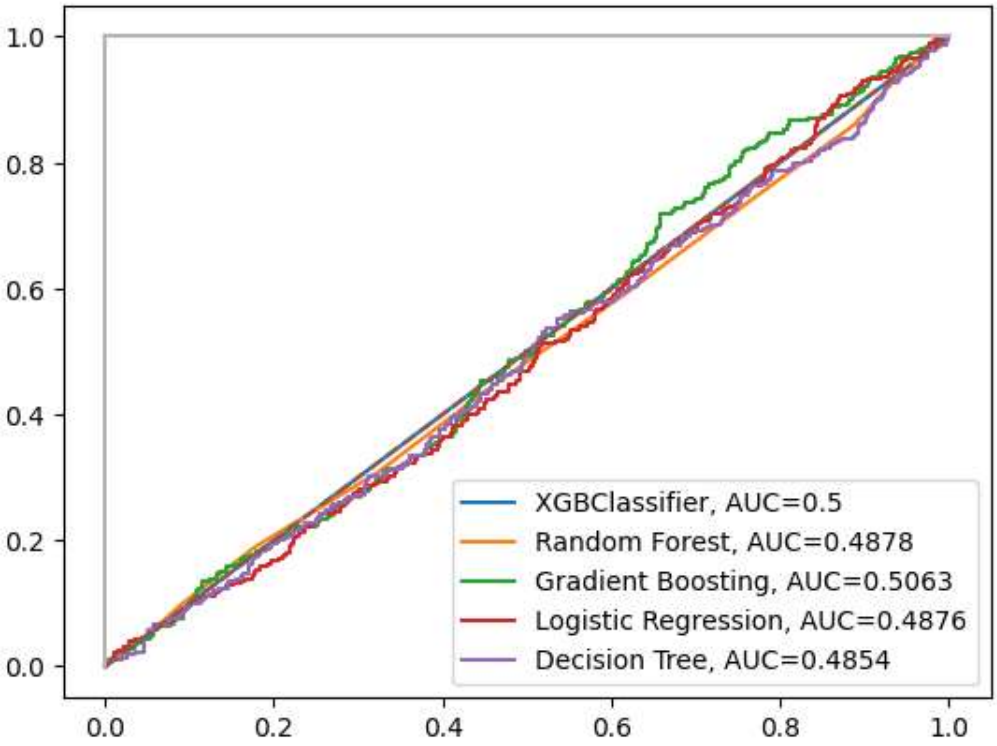
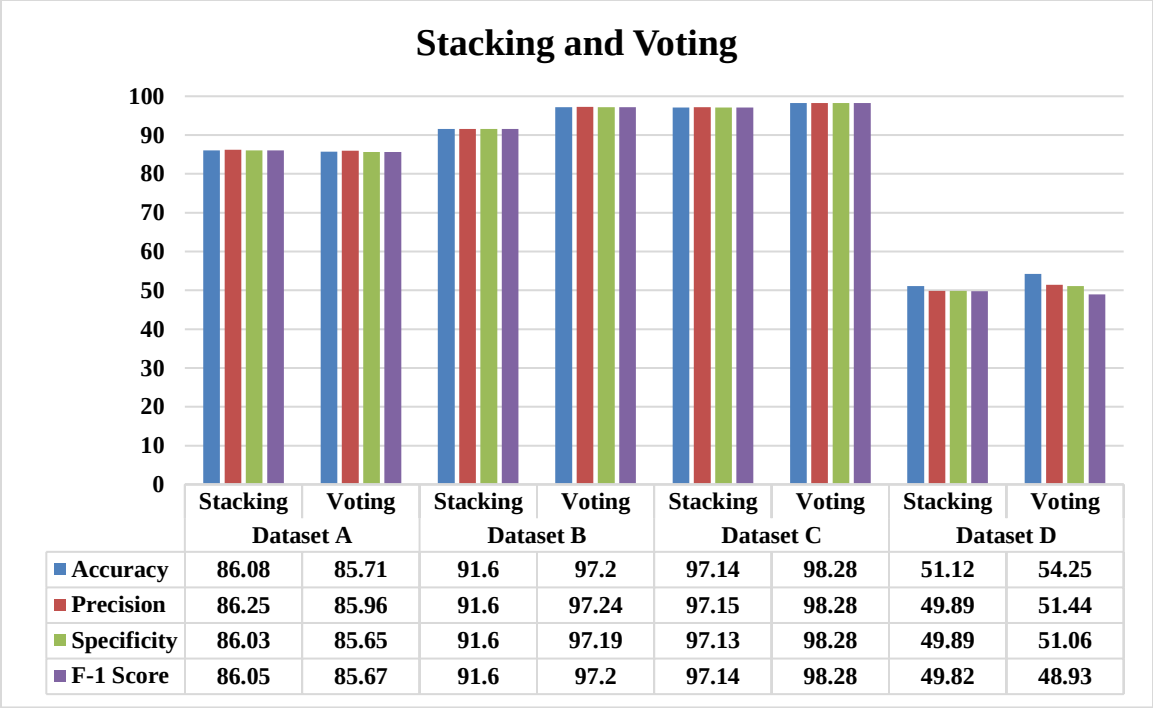


Figure 4.15: Analysis of Boosting for D dataset (AUC-ROC)



	Stacking	Voting	Stacking	Voting	Stacking	Voting	Stacking	Voting
	Dataset A		Dataset B		Dataset C		Dataset D	
Accuracy	86.08	85.71	91.6	97.2	97.14	98.28	51.12	54.25
Precision	86.25	85.96	91.6	97.24	97.15	98.28	49.89	51.44
Specificity	86.03	85.65	91.6	97.19	97.13	98.28	49.89	51.06
F-1 Score	86.05	85.67	91.6	97.2	97.14	98.28	49.82	48.93

Figure 4.16: Analysis of Stacking and Voting (AUC-ROC)

The evaluation of stacking and voting algorithms for breast cancer prediction across four datasets (A, B, C, and D) highlights the varying performance of these ensemble methods. In Dataset A, stacking achieved an accuracy of 86.08% with a balanced performance in precision (86.25%), specificity (86.03%), and F1-score (86.05%), closely followed by the voting algorithm with slightly lower accuracy at 85.71%. For Dataset B, voting outperformed stacking with a significantly higher accuracy of 97.2%, precision, and F1-score (all 97.2%), whereas stacking showed good, though lower, consistency at 91.6% for all metrics. In Dataset C, both algorithms performed exceptionally well, with voting achieving the highest scores across the board (98.28%) and stacking close behind with 97.14%. However, Dataset D showed much lower predictive power for both models, with stacking having a notably poor performance, particularly in specificity (49.89%) and F1-score (49.82%), while voting performed slightly better with an accuracy of 54.25% and an F1-score of 48.93%. Overall, voting consistently outperformed stacking in the more accurate datasets, especially in higher-quality datasets like B and C.

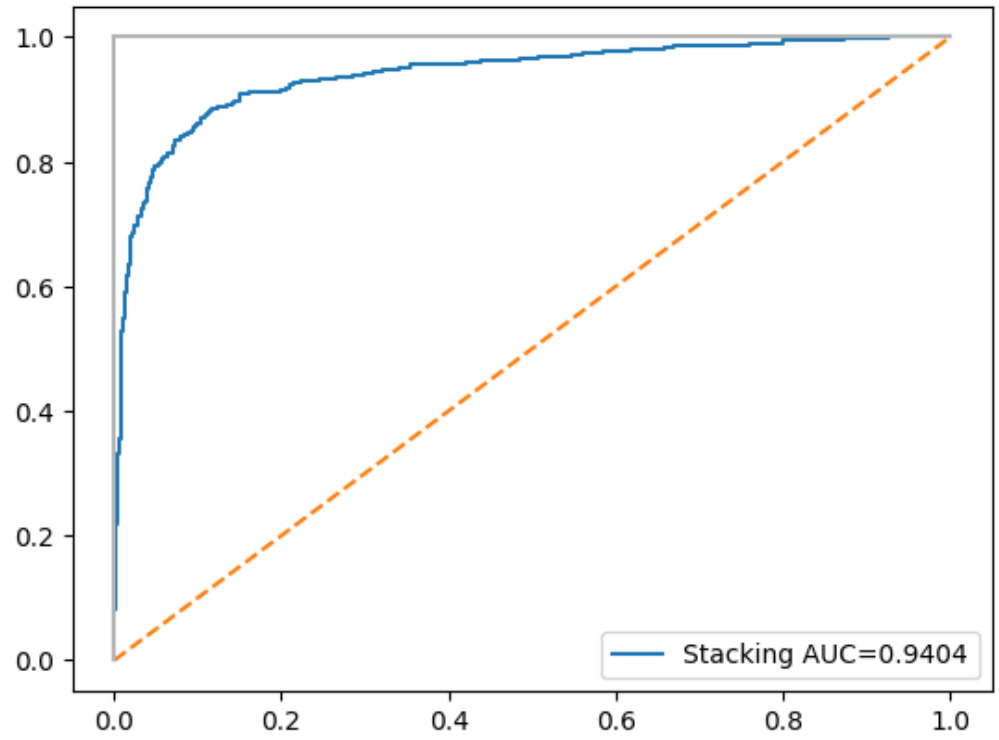


Figure 4.17: Analysis of Stacking for A dataset (AUC-ROC)

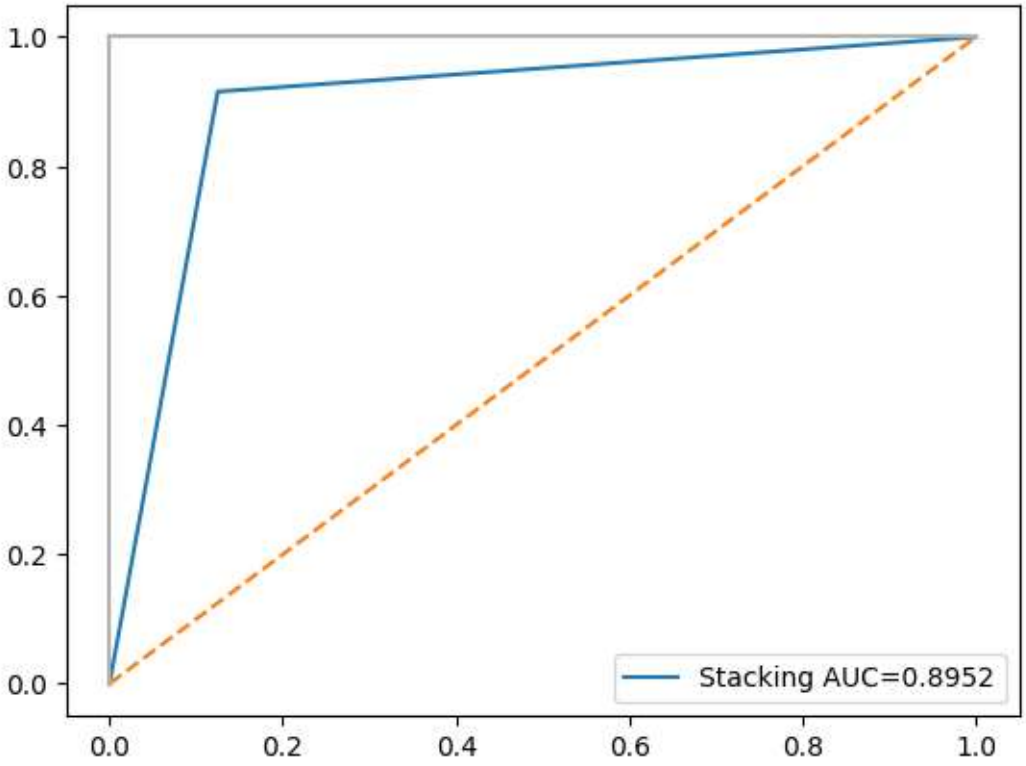


Figure 4.18: Analysis of Stacking for B dataset (AUC-ROC)

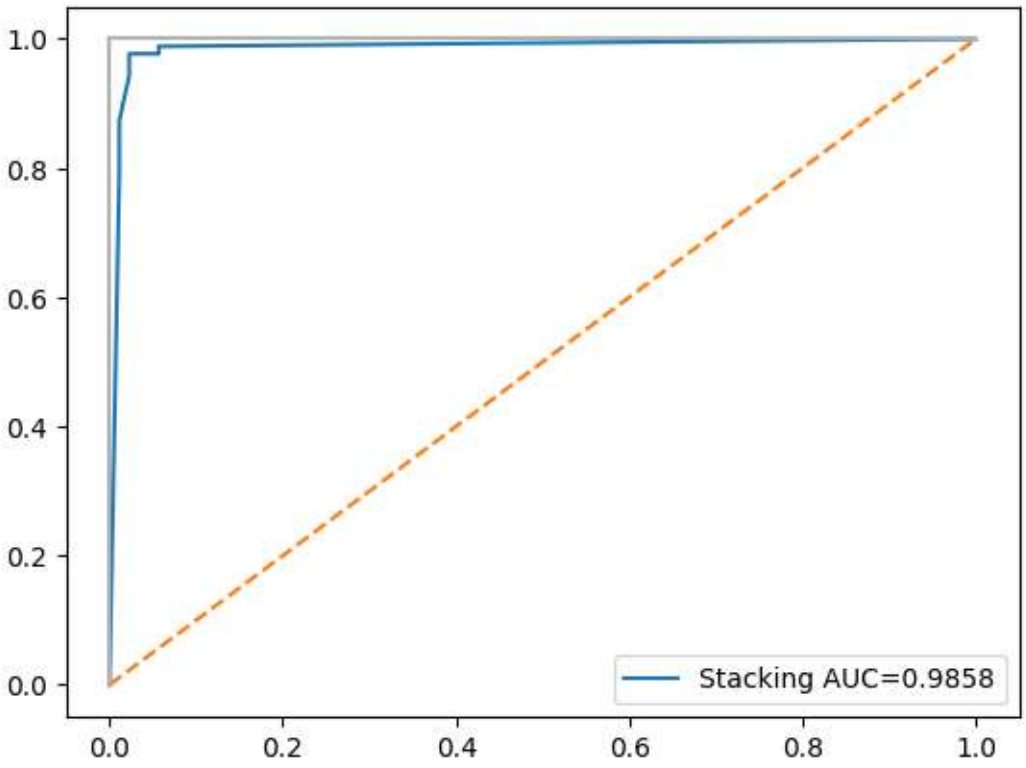


Figure 4.19: Analysis of Stacking for C dataset (AUC-ROC)

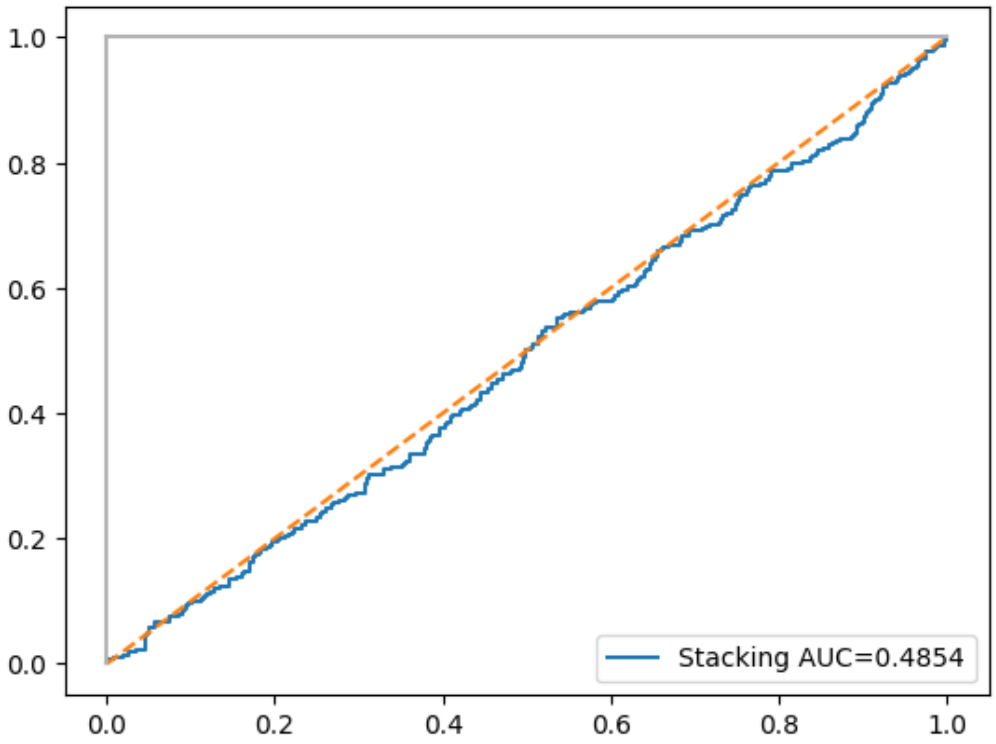


Figure 4.20: Analysis of Stacking for D dataset (AUC-ROC)

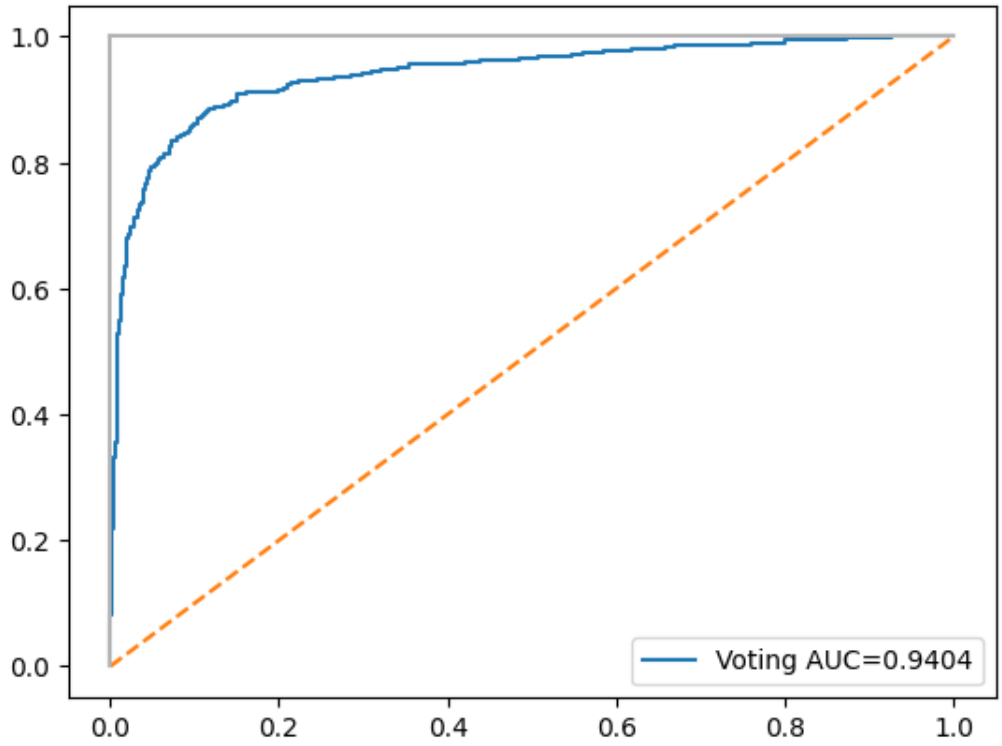


Figure 4.21: Analysis of Voting for A dataset (AUC-ROC)

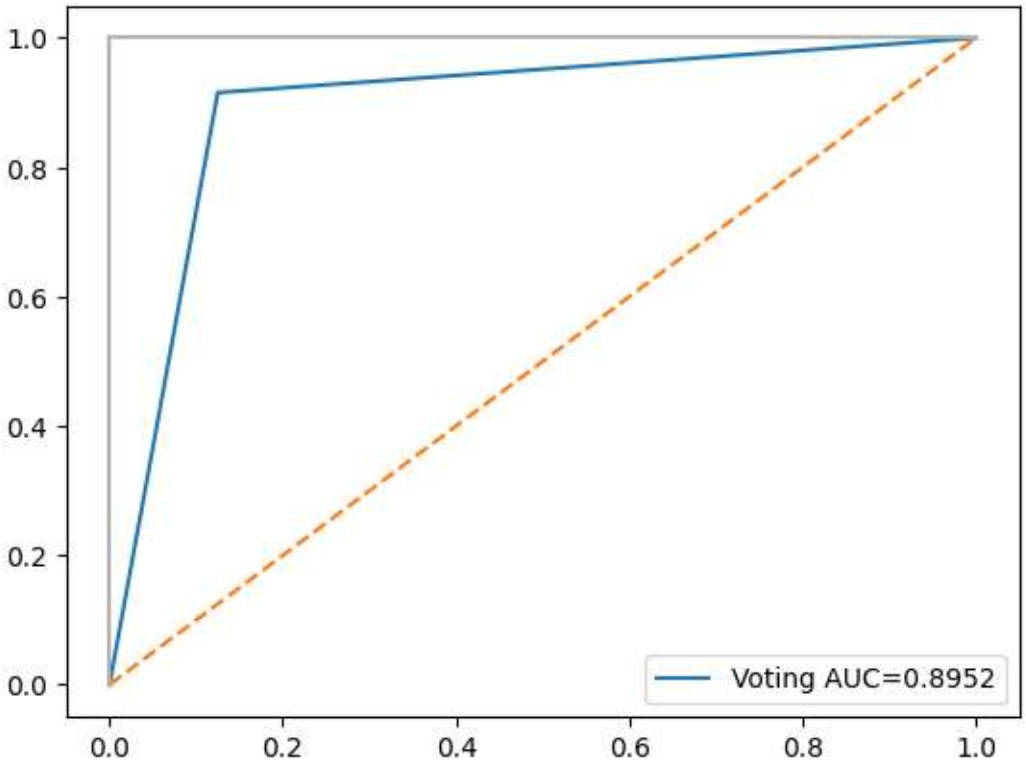


Figure 4.22: Analysis of Voting for B dataset (AUC-ROC)

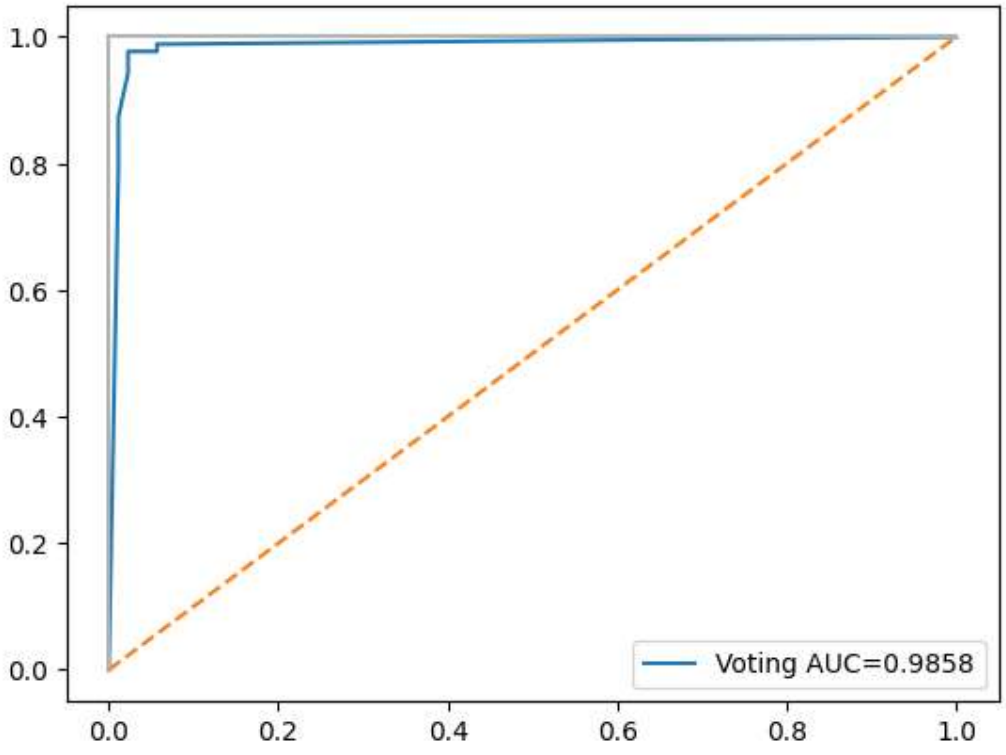


Figure 4.23: Analysis of Voting for C dataset (AUC-ROC)

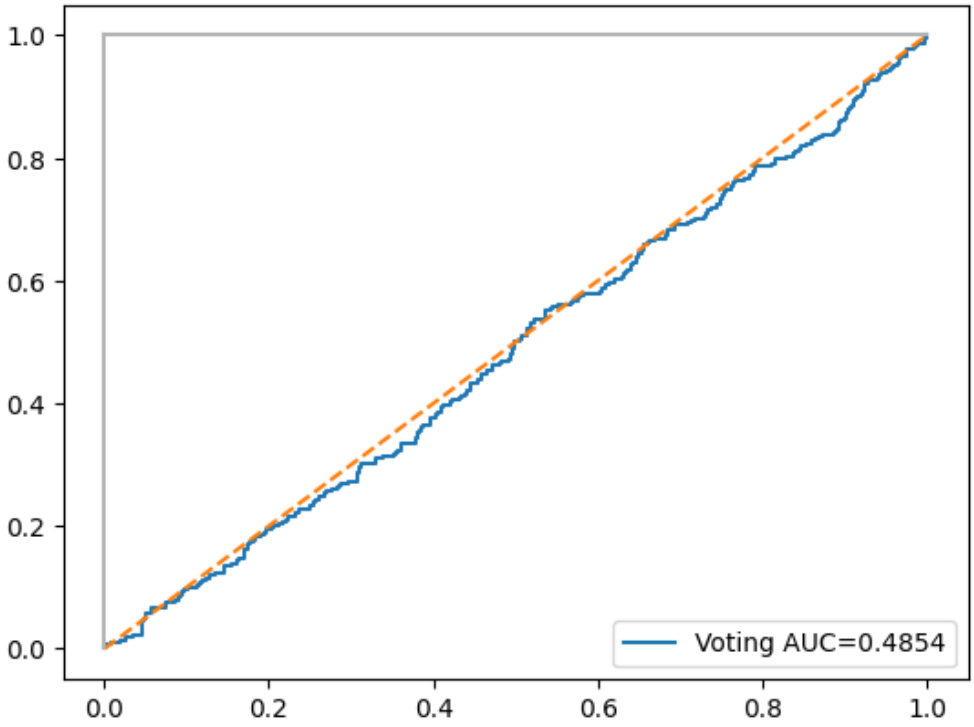


Figure 4.24: Analysis of Voting for D dataset (AUC-ROC)

4.3 DISCUSSION

The experimental investigation utilizing four Kaggle-hosted datasets revealed promising advancements in breast cancer prediction, particularly through the integration of various classifiers such as Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbors (KNN). Among these, KNN stood out, achieving impressive test accuracies of 81.14%, 97.2%, and 98.85% for Datasets A, B, and C, respectively, while Logistic Regression performed notably well with 96.125% for Dataset D. Further optimization through ensemble methods such as Bagging, Boosting, Stacking, and Voting, combined with hyper-parameter tuning, elevated model performance, with the Bagging KNN (KNNB) model emerging as the most accurate, attaining an exceptional accuracy rate of 99.42%. This underscores KNNB's potential as a highly effective tool for early breast cancer detection. The study's findings contribute valuable insights to breast cancer prediction and management, offering a pathway for improving early intervention strategies and patient outcomes.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

The strategy we propose has many benefits, both in terms of the economy and society. Our model is designed to explore and detect the essential features of a patient with breast cancer disease using a real-world dataset. By providing information regarding the frequency of breast cancer illness and suggesting preventative strategies, this project has advantages for society. By use of precise diagnosis and routine examinations, early intervention may be suggested, increasing people's awareness of possible health hazards. Our method simplifies and accurately predicts disease, as seen by its quicker processing times and lower compilation needs. Our approach employs cutting-edge diagnostic techniques and comprehensive data analysis to identify the underlying causes of breast cancer disease. We hope that society will accept and apply this suggested course of action.

5.2 IMPACT ON ENVIRONMENT

Our suggested paradigm's streamlined diagnostic processes are especially useful in distant areas. By applying our proven technique, we can significantly reduce the time and complexity required to detect breast cancer disease. The approach is uncomplicated and has no negative effects, thus the environment gains from its simplicity. Patients may now receive evaluations for breast cancer illness without having to fly to large cities. The patient's diagnostic report and the prediction model, which also forecasts potential outcomes, might work well together. The inexpensive cost of identifying breast cancer disease allays worries over the cost of in-person treatments. It is simple enough for people of all skill levels to use. Our methodology helps to enhance the social and economic environments by making it clear whether a patient has breast cancer disease. We are optimistic that our approach will constitute a major breakthrough in medical scientific technology if it is put into practice.

5.3 ETHICAL ASPECTS

It's critical to set up ethical protections prior to system deployment in order to stop sensitive data, such as diagnostic reports and personal information, from being disclosed. Future research projects and the diagnosis and treatment of breast cancer illness may benefit from the practical uses of our suggested methodology. We recognize that the problem at hand has worldwide ramifications, impacting not just certain places or regions but the entire earth. The proposed approach enables people with breast cancer illness or others who are aware of the problem to predict how it could influence their life.

5.4 SUSTAINABILITY PLAN

We are optimistic that our proposed model may be easily included into the global breast cancer disease diagnostic and research technologies. We think that women who are at risk of breast cancer disease would benefit most from our suggested strategy, which would help them predict how likely it is that they would have the ailment. We are motivated and prepared to offer our support to rural areas if given the resources and chances to put it into practice. We believe that the paradigm we have proposed will be both practical and long-lasting, making a significant worldwide contribution to the progress of breast cancer disease research and diagnostics.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 SUMMARY OF THE STUDY

In this interesting paper, we use algorithms to assess the effect rate on people. Our methodology allows us to accurately predict the future, even though people may wrongly think they should be on the lookout for breast cancer disease. People may determine if they are going to be impacted or not by using the diagnostic approach that our prediction system employs. Our approach makes it easier to identify different phases of breast cancer illness quickly, and it can also be useful for diagnostic authorities. We have employed a variety of widely used algorithms that are simple to use, need minimal training, and have good accuracy.

6.2 CONCLUSION

The modern society we live in is open to everybody and has sophisticated technology. The recommended technology is quick and easy to use, making it accessible to a worldwide audience. We want to make the process of predicting breast cancer disease as easy as possible so that everyone may take advantage of our state-of-the-art algorithms. We pledge to guarantee the idea's feasibility, and in the future, we'll add more features and focus on subjects that are more commonly discussed. This objective is stated rather effectively.

6.3 IMPLICATION FOR FURTHER STUDY

Death is a natural part of being human, and different illnesses have an impact on our day-to-day living. breast cancer illness is a problem that affects many people, yet new developments in treatment and diagnostic tools provide chances for recovery. Since we live in a developing nation, we have access to more sophisticated and accurate medical technology. These modern tools have made the process of diagnosing breast cancer disease simpler and more efficient. We hope that by attempting something different, more

individuals would choose to support their community as we have. We wish to introduce additional algorithms in the future after working on a few to increase performance.

6.4 LIMITATIONS

Because we are mortal, we are susceptible to a wide range of illnesses. Many health issues affect our everyday lives, and although some of us may have access to necessary recovery tools and efficient treatments, many others still struggle with ailments like breast cancer disease. Living in a developing nation, however, has allowed us to benefit from advances in diagnostic and treatment technology, which are always changing to become more precise and dynamic. These developments provide promise for better healthcare results and a higher standard of living for those suffering from a variety of illnesses. We want to continue working on this project if the community accepts our suggested methods. It's crucial to remember that our model is now only being used with a small dataset, thus the assessment may differ from other research approaches.

REFERENCE

[1] "World Cancer Research Fund International" Last Accessed: March 22, 2022. Available: https://www.wcrf.org/cancer-trends/breast-cancer-statistics/

[2] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: theory and practice," Neurocomputing, vol. 415, pp. 295–316, 2020.

[3] Shamrat, F.J.M., Raihan, M.A., Rahman, A.S., Mahmud, I. and Akter, R., 2020. An analysis on breast disease prediction using machine learning approaches. International Journal of Scientific & Technology Research, 9(02), pp.2450-2455.

[4] Keleş, M.K., 2019. Breast cancer prediction and detection using data mining classification algorithms: a comparative study. Tehnički vjesnik, 26(1), pp.149-155.

[5] Anastraj, K., Chakravarthy, T., Sriram, K., Collge, A.S.P. and Poondi, T., 2019. Breast cancer detection either benign or malignant tumor using deep convulionalneural network with machine learning techniques. Adalya Journal, 8, pp.77-83.

[6] Erkal, B. and Ayyıldız, T.E., 2021, November. Using Machine Learning Methods in Early Diagnosis of Breast Cancer. In 2021 Medical Technologies Congress (TIPTEKNO) (pp. 1-3). IEEE.

[7] Shravya, C., Pravalika, K. and Subhani, S., 2019. Prediction of breast cancer using supervised machine learning techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(6), pp.1106-1110.

[8] "Breast Cancer Dataset", Accessed: December 29, 2021, Available: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

[9] "Wisconsin Breast Cancer Database", Accessed: December 29, 2021, Available: https://www.kaggle.com/datasets/roustekbio/breast-cancer-csv.

[10] "Breast Cancer Prediction", Accessed: December 29, 2020, Available: https://www.kaggle.com/datasets/adhyanmaji31/breast-cancer-prediction/data

[11] "Breast Cancer Coimbra", Accessed: December 29, 2023, Available: https://www.kaggle.com/datasets/atom1991/breast-cancer-coimbra.

[12] L. Mary Gladence, M. Karthi, V. Maria Anu. "A statistical Comparison of Logistic Regression and Different Bayes Classification Methods for Machine Learning" ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, Vol -10, No-14, August 2015

[13] "Logistic Regression for Machine Learning", Accessed: August 6, 2021, Available: https://www.capitalone.com /tech/machine-learning/what-is-logistic-regression/

[14] Ghosh, Pronab, Asif Karim, Syeda Tanjila Atik, Saima Afrin, and Mohd Saifuzzaman. "Expert cancer model using supervised algorithms with a LASSO selection approach." International Journal of Electrical and Computer Engineering (IJECE) 11, no. 3 (2021): 2631

[15] Aunik Hasan Mridul, Md. Jahidul Islam, Mushfiqur Rahman, Mohammad Jahangir Alam, Asifuzzaman Asif. "A Machine Learning-Based Traditional and Ensemble Technique for Predicting

Breast Cancer", In December, 2022. Conference: 22th International Conference on Hybrid Intelligent Systems (HIS 2022) online, 2022At: Auburn, Washington, USA

[16] Shravya, C., Pravalika, K. and Subhani, S., 2019. Prediction of breast cancer using supervised machine learning techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(6), pp.1106-1110.

[17] Merouane, E. and Said, A., 2022. Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers. International Journal of Advanced Computer Science and Applications, 13(2).

[18] Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Alamin Dhaly, Md Nour Hossain. "A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease". In April 2022 The 10th International Conference on Computer and Communications Management in Japan

[19] Pasha, Maruf, and Meherwar Fatima. "Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection." J. Softw. 12, no.12 (2017): 923-933

[20] emmens, Aurélie, and Christophe Croux. "Bagging and boosting classification trees to predict churn." Journal of Marketing Research43, no. 2 (2006): 276-286

[21] Wang, Yizhen, Somesh Jha, and Kamalika Chaudhuri. "Analyzing the robustness of nearest neighbors to adversarial examples." In International Conference on Machine Learning, pp. 51335142. PMLR, 2018.

[22] Drucker, Harris, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. "Boosting and other ensemble methods." Neural Computation 6, no. 6 (1994): 1289-1301.

[23] L. Mary Gladence, M. Karthi, V. Maria Anu. "A statistical Comparison of Logistic Regression and Different Bayes Classification Methods for Machine Learning" ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, Vol -10, No-14, August 2015

[24] Sharma, Ajay, and Anil Suryawanshi. "A novel method for detecting spam email using KNN classification with spearman correlation as distance measure." International Journal of Computer Applications 136, no. 6 (2016): 28-35

[25] Pasha, Maruf, and Meherwar Fatima. "Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection." J. Softw. 12, no.12 (2017): 923-933.

[26] Islam, Rakibul, Abhijit Reddy Beeravolu, Md Al Habib Islam, Asif Karim, Sami Azam, and Sanzida Akter Mukti. "A Performance Based Study on Deep Learning Algorithms in the Efficient Prediction of Heart Disease." In 2021 2nd International Informatics and Software Engineering Conference (IISEC), pp. 1-6. IEEE, 2021

[27] "What is Correlation in Machine Learning?", Accessed: August 6, 2020, Available: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47

[28] "What is Correlation in Machine Learning?", Accessed: November 8, 2021, Available: https://medium.com/analytics /what-is-correlation

Breast cancer

ORIGINALITY REPORT

24%	21%	14%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	thesai.org Internet Source	7%
3	Submitted to Daffodil International University Student Paper	2%
4	Submitted to Loomis-Chaffee High School Student Paper	1%
5	Submitted to Kaplan Professional Student Paper	<1%
6	hrcak.srce.hr Internet Source	<1%
7	ebin.pub Internet Source	<1%
8	journals.plos.org Internet Source	<1%
9	Submitted to Athlone Institute of Technology Student Paper	<1%