Breast Cancer Prediction Using Traditional and Ensemble Machine Learning Techniques

BY

Md. Ziad Hosen
ID: 191-15-2590
AND

Deepanita Baidya
ID: 191-15-2453

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

Supervised By

Amit Chakraborty Chhoton
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised By

Al Amin Biswas
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This project titled "**Breast Cancer Prediction Using Traditional and Ensemble Machine Learning Techniques**", was submitted by **Md. Ziad Hosen**, **Deepanita Baidya** and ID No: 191-15-2590, 191-15-2453 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 25-01-2023

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Atiqur Rahman
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Shayla Sharmin

Internal Examiner

Shayla Sharmin 25.1.23
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Farid 25-01-23

External Examiner

Dr. Dewan Md Farid
Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Amit Chakraborty Chhoton, Lecturer (Senior Scale), Department of CSE,** Daffodil International University. We 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:

Amit

Amit Chakraborty Chhoton
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised by:

23.01.2022

Al Amin Biswas
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

Zied

Md. Ziad Hosen
ID: -191-15-2590
Department of CSE
Daffodil International University

Deepa

Deepanita Baidya
ID: -191-15-2453
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing in making us possible to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Amit Chakraborty Chhoton, Lecturer (Senior Scale)**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance continual encouragement, constant and energetic supervision, constructive criticism valuable advice r and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, Head of**,** the Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire course mate at Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

Abstract

Physical diseases like breast cancer have been on the rise recently.The majority of women are affected by breast cancer. The ratio of normal to diseased areas and the pace of unchecked tissue growth are used to quantify the illness. Breast cancer detection and prediction have been the subject of several research in the past. We have identified a few excellent chances to develop the methodology. We suggest employing efficient algorithm models to forecast dangers and raise early awareness. Our suggested approach is suited for straightforward breast cancer forecasts and is simple to apply in the actual world. We have used two dataset and Kaggle website hosted the dataset. Decision tree (DT), Random Forest (RF), Logistic Regression (LR), K-Nearest Classifier (KNN), and other classifiers have all been integrated in our model. Test accuracy for the Random Forest Classifier was 97.36% and 97.81% which was good performance for datasets A and B. We are getting better accuracy for the Logistic Regression was 98.54% using Dataset B. Other algorithms, Decision Tree tested accurate to 96.49%. In order to defend the performances, we also employed a variety of ensemble models. We used Bagging, Boosting, and Voting algorithms. To assign the optimal parameters to each classifier, we employed hyper-parameter tweaking. The experimental investigation reviewed the results of previous recent studies and found that RFBO and LRGD performed best, with 98.24% and 99.27% accuracy being the highest level of accuracy for breast cancer predictions.

Keywords: Prediction, Machine Learning, Algorithms, Ensemble Model, Voting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Multiple tissues are being harmed or developing out of control, which is known as cancer, since the sickness is the worst aspect of our daily lives. Breast cancer is a type of cancer that develops when unregulated tissue or damaged tissue does so. This patient's prevalence is significantly rising. However, finding or recognizing the injured region at the time of diagnosis is the key issue. Machine learning may be the most effective component of a crucial factor in predicting the presence of breast cancer from responsive health datasets by examining various variables and patient diagnosis records. We looked at the patient's diagnosis papers for our work and discovered certain key factors to pinpoint the condition. The dataset dealt with the size and structure of a woman's bodily tissues as well as determining whether or not she had breast cancer. In order to employ machine learning algorithms to recognize the cancer tissue in the body, several different researchers have worked together. However, their method and accuracy were not appropriate nor smooth for predicting breast cancer. We suggest our method to increase the accuracy rate of breast cancer prediction in a woman's body. There are two different kinds of machine learning techniques. One of them is under supervision, while the other is not. Working with labeled data, supervised learning creates outputs from inputs based on examples of input-output pairings. The dataset's training data is used as the working data. Unsupervised learning works with the unlabeled data and creates the model to work with its patterns and information which was not detected previously. Unsupervised learning uses unlabeled data to build models that can make use of previously undetected patterns and information.

1.2 Motivation

Breast cancer is becoming more prevalent since it affects the majority of women, and the prevalence is rising daily. The cause of breast cancer includes eating habits, cosmetic cream use, and other factors. According to a study conducted by the World Cancer Research Fund International, there were 2261419 afflicted individuals in 2020. The death has the number 684996 [1]. Additionally, they claimed that alcohol, higher birth weights, and adults who had reached the age of eight were signs of breast cancer. A few studies have been conducted to predict cancer. We do several research on the prognosis of breast cancer. The majority of them don't have greater accuracy. As a result, we become more determined, and eventually, we discovered our method's highest accuracy. We have developed a method to forecast the presence of breast cancer in suspicion or regular patients.

1.3 The rationale of the study

In this research, we put out a methodology to forecast breast cancer in humans. Recently, we have observed that this cancer is beginning to harm our society. But we also observed that there is a shortage of knowledge and diagnostic tools. Cancer detection and symptom analysis are expensive in our developing nation. In our work as researchers, we are attempting to use machine learning to address the issue

1.4 Research Questions

1) How are the algorithms in this suggested model functioning?
2) What is the likelihood that someone with breast cancer will survive?
3) How can the early diagnosis of breast cancer be predicted?
4) What advantages does our suggested model have?
5) What potential applications of this work exist in the actual world?
6) What is the project's projected future?
7) What safety measures are required for this work?
8) How can we assess our breast cancer prediction model?

1.5 Expected Output

Breast cancer is affecting people today. Additionally, nobody is certain if she is impacted or not. We are recommending the best approach for predicting or identifying the condition by looking at the diagnosis report. Our approach can discover breast cancer patients, enhance decision-making, and precisely evaluate the effect. It may quantify life quality while also analyzing connected issues. It can raise people's awareness of the condition of breast cancer. The suggested model can assess the illness in the smallest amount of time.

1.6 Project Management and Finance

Our suggested model is economical and useful in everyday life. The evaluation of breast cancer may be a useful asset for our country. To apply the prediction process in real life, common tools are required. The greatest results and seamless operation of our model will result from the usage of high-configuration tools. However, it is still possible if we utilize simple tools.

1.7 Report Layout

The relevant study done by the earlier researchers is covered in Chapter 2. Before beginning the investigation, we need to examine the introduction and motive. As a consequence, we talk about the introduction, which may explain the suggested approach in depth, and the motivation portion, which can explain the forecast. After finishing the Introduction section, we concentrated on relevant research and gathered internal data for our work. In our methodology section, we have chosen machine learning algorithms, applied them to our dataset, and then determined which one is the best. Following the pre-processing phase, we tested the data, and at last, we obtained our desired result, which we may refer to as the comparison one. That was explained in our last part which is known as the conclusion. That was discussed in our final section, which is referred to as the conclusion.

CHAPTER 2

BACKGROUND

2.1. Preliminaries

To determine the exact layout of breast cancer, machine learning techniques are applied. In this section, we try to examine the investigations connected to the evaluation examination of the patient's diagnosis report. These models use computations like Decision Tree, Gradient Boosting, Adaboost, Random Forest, and Logistic Regression. Deep learning models are put into practice in this section to play out the exploration. Several researchers who used several models in their study are mentioned in the section.

2.2. Related works

We have used a few machine learning classifiers to categorize breast cancer, and they are appropriate for the job we are proposing. To execute decision models, machine learning algorithms that are based on decision tree models are known as "tree structures" [1] [2]. similarly proposed a comparison between Random Forest, Naïve Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (K-NN) and they found the SVM is the best classifier with an accuracy of 97.9% compared with K-NN, RF and NB, they are based on Multilayer perception with 5 layers and 10 times cross validation using MLP. In this study the author F. M. Javed Mehedi Shamrat et al.[3] focused on the enhancement of the accuracy value using Wisconsin Breast Cancer Diagnostic dataset (WBCD) by applying ML-based system for the early prediction of breast cancer disease. Six supervised classification techniques are used which are: SVM, NB, KNN, RF, DT, and LR. According to the analysis of breast cancer prediction performance, SVM had the highest performance and the highest classification accuracy (97.07%). While NB and RF have attained the second-highest prediction accuracy. In this paper, the author Mumine Kaya Keles [4] was to predict and detect breast cancer early even if the tumour size is petite with non-invasive and painless methods that use data mining classification algorithms. The effectiveness of data mining techniques in the detection of breast cancer was examined in this study using the Weka data mining software and an antenna dataset. The

10-fold cross-validation was used to obtain the most authentic results using the Knowledge Extraction based on Evolutionary Learning data mining software tool where Random forest outperformed all the other algorithms giving an average accuracy of 92.2 percent. In this study, K.Anastraj et al.[5] have performed a comparative analysis between different machine learning algorithms which are: back propagation network, artificial neural network (ANN), convolutional neural network (CNN) and support vector machine (SVM) on the Wisconsin Breast Cancer (original) datasets. Deep and convolutional neural network with ALEXNET was used for feature extraction and analysis of the benign and malignant tumor. According to the simulation results, support vector machine is the best approach and had given better results (94%). In this study the authors Begüm Erkal and Tülin Erçelebi Ayyıldız [6] provided the results by using seven different machine learning techniques which are: Naïve Bayes, BayesNet, K-Nearest Neighbor (k-NN), Support Vector Machine (SVM), Multilayer Perceptron (MP), Random Forest (RF), Logistics Regression (LR) on the Breast Cancer Wisconsin (Original) open dataset for the classification of breast cancer. In the experimental results, BayesNet was the best classification method with an accuracy rate of 97.13%. In this paper Ch. Shravya et al.[7] provided relative study on the implementation of models using Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) on the dataset taken from the UCI Repository. With respect to the results of accuracy, precision, sensitivity, specificity and False Positive Rate the efficiency of each algorithm is measured and compared and focused in the advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. The results analysis shows that the combination of multidimensional data with various feature selection, classification, and dimensionality reduction techniques can offer advantageous tools for inference in this field. This study has shown that SVM is the best accuracy of 92.7%. The authors Ertel Merouane et al.[8].

2.3. Comparative Analysis and Summary

The machine learning model is one that is used a lot these days. To locate our respective job, we were required to complete a challenging endeavor. All connected works have poor model results and poor accuracy. To identify the dataset's greatest accuracy of prediction, we had to apply a different machine learning models. We had to deal with running the models on high-end hardware. To get at the categorization rates, we used a few individual computations. By adding pricey GPUs, complicated models might generate lengthy runtime.

2.4. Scope of the Problem

The issue involved making the breast cancer diagnostic process easier and more familiar to women. We attempted to provide the highest accuracy with our suggested model because there are so many works with machine learning that are linked to it. Although we had little room for improvement in the process, we could execute the concept using straightforward techniques to reduce the number of breast cancer diagnoses.

2.5 Challenges

Dataset sourced from Kaggle [9] [22]. The information was very usable and simple to use. We must manually review the dataset for any missing data when the data gathering is complete. Two anonymous columns have been removed since we found no use for them. With this dataset, no one has ever as accurate as we are.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrument

To extract the most accuracy from the dataset, we used a variety of algorithms and hybrid models. We required some tools, such as efficient conFigureuration tools the with best GPUs. Python programming language and technologies including Google Collaboratory, Jupiter Notebook, and Anaconda have been utilized. Through the browser, it enables the authoring and execution of any Python code. All tests were performed on a computer running the 64-bit version of Windows 10 Pro on an AMD Ryzen 5 3600 6-core processor clocked at 3.59 GHz with 8 GB of RAM.

3.2 Data Collection Procedure

The dataset was virtually ready for implementation when it was downloaded from Kaggle. The sizes of the dataset A column and rows are 32 and 569, respectively. There is one additional dataset B, and there are 10 rows and 683 columns. The frequency of breast cancer is categorized in the diagnostic and Class column. Every characteristic was crucial for predicting breast cancer. Malignant and Beginning conditions are used to categorize patients. Here, Malignant stands in for M, and Begin for B. These values have been converted using nominal values. There, 0 represents "B" and 1 represents "M." We have determined the frequency of these two circumstances. 212 individuals were at the malignant stage, leaving 357 patients in the initial stage in dataset A. Another dataset B had 239 patients in the malignant stage and 444 individuals in the initial stage. Figures 3.1 for dataset A and 3.2 for dataset B below display the ratio. The dataset was split into two pieces. They go through testing and training. We've chosen 20% for the exam portion and another 80% for the learning portion.

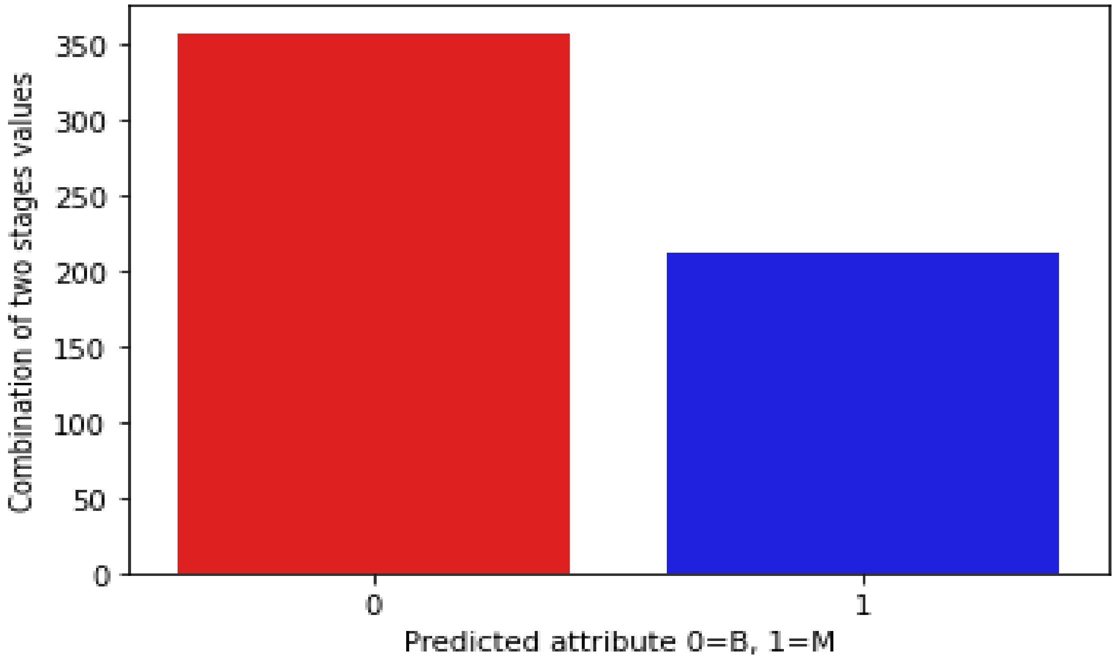


Figure 3.1: Number of target values dataset A

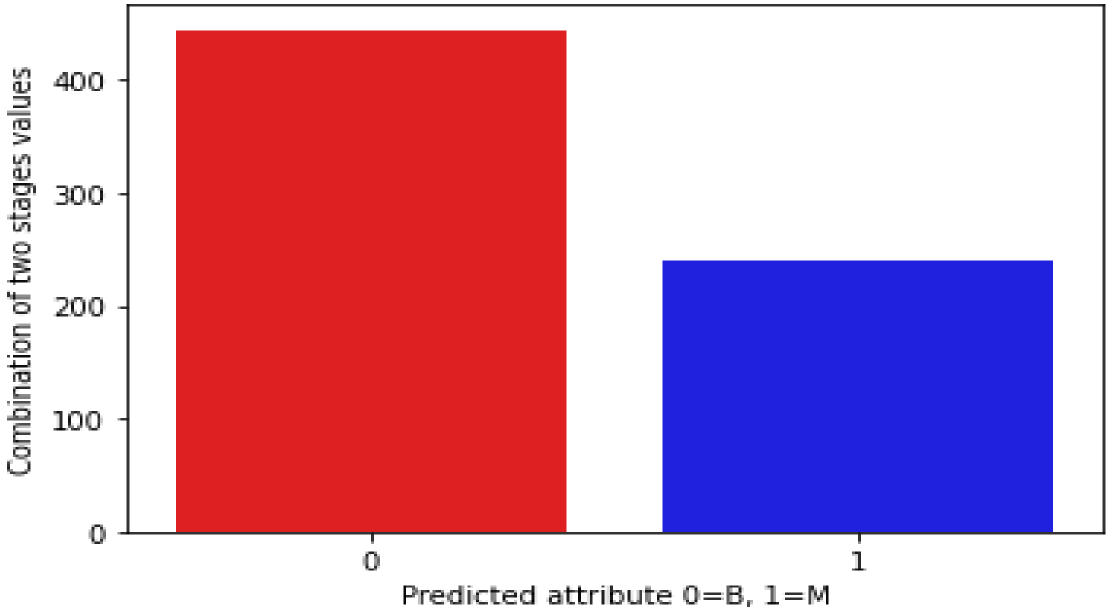


Figure 3.2: Number of target values dataset B

The dataset contains nominal values and there were no missing or incorrect values. A comprehensive explanation of the dataset with its range is displayed in table 3.1 and 3.2.

Table 3.1: Details of the dataset A

Attributes	**Description**	**Value Range**	**Types of values**
Diagnosis	Malignant or Begin	0 and 1	Integer
Radius_mean	Radius of Lobes	6.98 to 28.1	Float
Texture_mean	Mean of Surface Texture	9.71 to 39.28	Float
Perimeter_mean	Outer Perimeter of Lobes	43.8 to 188.5	Float
Area_mean	Mean Area of Lobes	143.5 to 2501	Float
Smoothness_mean	Mean of Smoothness Levels	0.05 to 0.163	Float
Compactness_mean	Mean of Compactness	0.02 to 0.345	Float
Concavity_mean	Mean of Concavity	0 to 0.426	Float
Concave points_mean	Mean of Concave Points	0 to 0.201	Float
Symmetry_mean	Mean of Symmetry	0.11 to 0.304	Float
Fractal_dimension_mean	Mean of Fractal Dimension	0.05 to 0.1	Float
Radius_se	SE of Radius	0.11 to 2.87	Float
Texture_mean	SE of Texture	0.36 to 4.88	Float
Perimeter_se	Perimeter of SE	0.76 to 22	Float
Area_se	Area of SE	6.8 to 542	Float
Smoothness_se	SE of Smoothness	0 to 0.03	Float
Compactness_se	SE of Compactness	0 to 0.14	Float
Concavity_se	SE of Concavity	0 to 0.4	Float
Concave points_se	SE of Concave Points	0 to 0.05	Float

Symmetry_se	SE of Symmetry	0.01 to 0.08	Float
Fractal_dimension_se	SE of Fractal Dimension	0 to 0.03	Float
Radius_worst	Worst Radius	7.93 to 36	Float
Texture_worst	Worst Texture	12 to 49.54	Float
Perimeter_worst	Worst Perimeter	50.4 to 251	Float
Area_worst	Worst Area	185 to 4254	Float
Smoothness_worst	Worst Smoothness	0.07 to 0.22	Float
Compactness_worst	Worst Compactness	0.03 to 1.06	Float
Concavity_worst	Worst Concavity	0 to 1.25	Float
Concave points_worst	Worst Concave Points	0 to 0.29	Float
Symmetry_worst	Worst Symmetry	0.16 to 0.66	Float
Fractal_dimension_worst	Worst Fractal Dimension	0.06 to 0.21	Float

Table 3.2: Details of the dataset B

Attributes	**Description**	**Value Range**	**Types of values**
Clump Thickness	Thickness of Clump	1 to 10	Integer
Uniformity of Cell Size	Cell size	1 to 10	Integer
Uniformity of Cell Shape	Cell shape	1 to 10	Integer
Marginal Adhesion	Adhesion Marginal value	1 to 10	Integer
Single Epithelial Cell Size	Cell size	1 to 10	Integer
Bare Nuclei	Number of Nuclei	1 to 10	Integer
Bland Chromatin	Number of Bland Chromatin	1 to 10	Integer

Normal Nucleoli	Number of Normal Nucleoli	1 to 10	Integer
Mitoses	Number of Mitoses	1 to 10	Integer
Class	Malignant or Begin	0 and 1	Integer

3.2.1 Categorical Data Encoding

The process of converting categorical variables into a numerical value is known as the numerical and categorical encoding method. The categorical encoding strategy was crucial to our investigation since machine learning only accepts and outputs numeric data. To use the categorical encrypting data approach, we had such a gender column.

3.2.2 Missing Value Imputation

It involves filling in the blanks or missing information with imputed values that were determined by research with other dataset data. However, it is gratifying that our dataset contained no missing values.

3.2.3 Handling Imbalanced Data

It refers to the process of changing a dataset's class distribution. It manages the data by systematically adding more examples to the dataset. While using the entire dataset as input, the data for minorities is increased.

3.2.4 Feature Scaling

It is a procedure for normalizing the variety of independent data variables. When there are no negative values or [-1,1] otherwise, the MinMax scaler is employed to scale all of the data features.

3.3 Statistical Analysis

Every type of research project needs an analysis section. This section depends on creating and assessing the algorithms I've employed. We must take a few procedures to prepare the dataset to make it useable because we have decided to use a comma-separated value (CSV) file. We have taken a number of measures, including data collecting and pre-processing.

We employed four distinct kinds of algorithms in this work, including Random Forest (RF), Logistic Regression (LR), K-Neighbors Classifier (KN), and Decision Tree Classifier (DT). RFBO and LRGD had the highest accuracy, which was 98.24% for dataset A. DTB, RFBO and LRGD had the highest accuracy, which was 98.24% for dataset B. Following the use of bagging, boosting, and voting algorithms, we obtained the best LRGD accuracy of 99.27%. Hyperparameter tweaking and 10-fold cross-validation have both been employed.

3.4 Proposed Methodology

Flow chart:

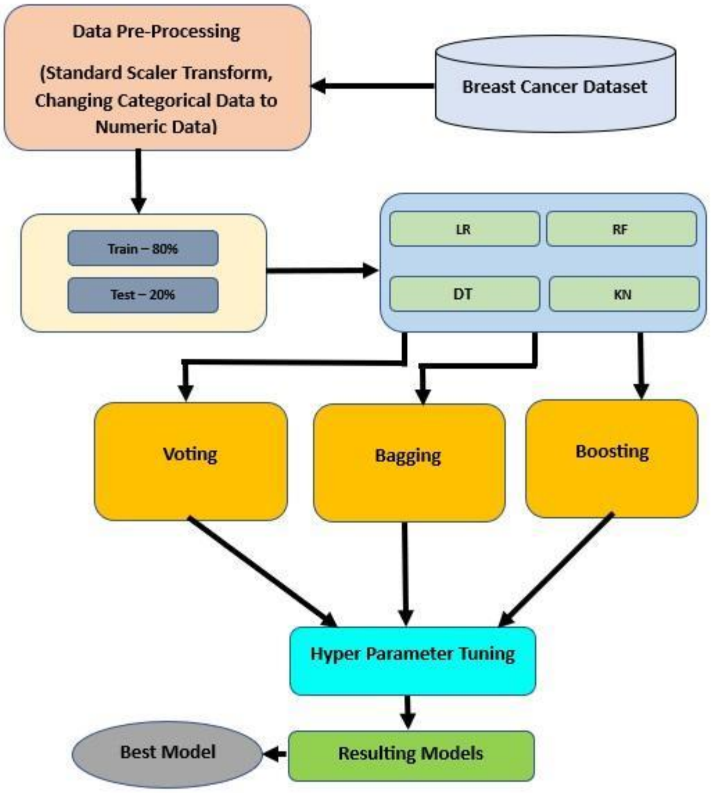


Figure 3.3: Methodology of Breast Cancer

In this section, we have predicted breast cancer using a process diagram. The dataset for the system's training and testing was initially introduced. Next, we used data preparation techniques such as the Standard Scaler Transform. Categorical data conversion to numeric data. We utilized 80% for the training portion and 20% for the testing portion. After that, we implemented algorithms and assessed the outcomes. Then, in order to get the highest forecast accuracy, we employed ensemble methods. Bagging, Boosting, and Voting are the ensemble algorithms. The outcomes of the ensemble algorithms that were used were then assessed. Then we used Hyper Parameter Tuning to verify the outcome. Then, using outcome analysis, we assessed the models that had been put into practice. Figure. 3.3 displays the recommended model technique. The identification of internal dependencies between two variables, or how one variable changes as a result of the change in another, is referred to as a correlation subplot. The more interdependence between variables suggests that it will be successful to predict one variable from another. It alludes to a deeper comprehension of the dataset and aids in our ability to identify the crucial factors [10].

3.5 Implementation Requirements

We require data sources in order to examine or train our suggested model. For things to go well, we must clean the dataset. A number of filtering techniques will be used to clean the dataset. Then, data pre-processing techniques like Standard Scaler Transform were used. Categorical data conversion to numeric data. We utilized 80% for the training portion and 20% for the testing portion. After that, we implemented algorithms and assessed the outcomes. Then, in order to get the highest forecast accuracy, we employed ensemble methods. Bagging, Boosting, and Voting are the ensemble algorithms. The outcomes of the ensemble algorithms that were used were then assessed. Then we used Hyper Parameter Tuning to verify the outcome. Then, using outcome analysis, we assessed the models that had been put into practice. Then we need to execute the data analysis part to start the learning process. Then we need to execute model learning and fit the method of predictions. Then we need to bagging, boosting, and voting the models to get the best accuracy. Then we can decide the best model to implement considering the best accuracy, precision, recall, and F-1 score. The learning process must then be initiated by carrying out the data analysis step. Next, we must put model learning into practice and fit the predictions approach. To acquire the best accuracy, we must then vote, boost, and bag the models. The best model may then be chosen for implementation based on accuracy, precision, recall, and F-1 score.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

A supervised learning method, which functions based on training and testing, was employed in this paper. The classification model is built using the training dataset. To obtain the outcome, the generated model is applied to the testing dataset. The machine-learning algorithm will be swiftly illustrated in the following sections.

4.1.1 Classifier Algorithms

In our study, we used Machine Learning (ML) based classifiers like Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KN).

Logistic Regression

A classifier approach based on machine learning called logistic regression (LR) contains two categories for the class label: yes or no, like a binary (0/1) scale. Although it permits the combined value of continuous data and discrete predictors, logistic regression is appropriate for discrete variables [11]. The idea is depicted in Figure. 4.1 below. Logistic regression adopts the supervised machine learning approach. The fundamental equation 1 is displayed below [12].

$$h(x) = \frac{1}{1} + e - (\beta o + \beta 1X) \text{ ………..(1)}$$

'$h(x)$' is the output of the logistic function, where 0 ≤ $h\Theta(x)$ ≥1

'$\beta 1$' is the slope

‘βo’ is the y-intercept

‘X’ is the independent variable

($\beta o + \beta 1X$) *– derived from the equation of a line Y (predicted) =* ($\beta o + \beta 1X$) *+ Error.*

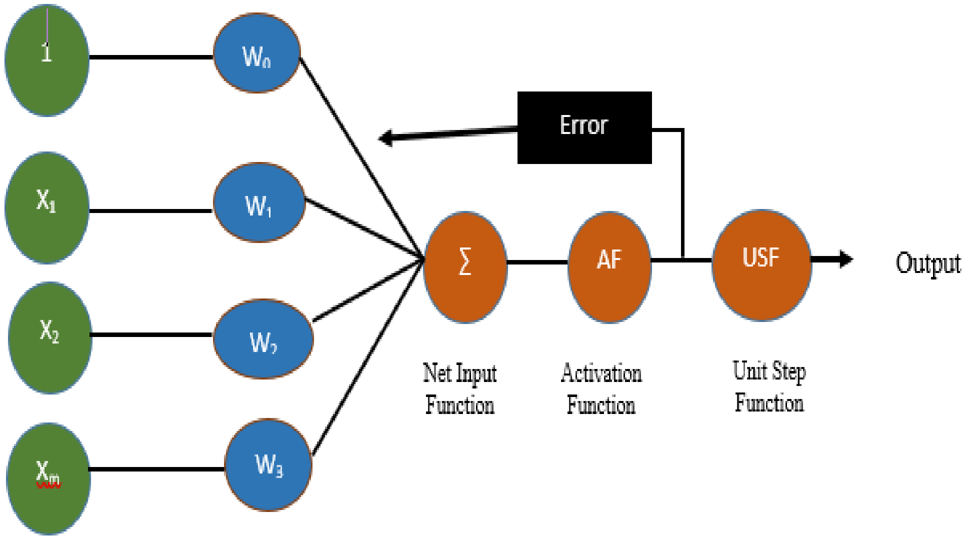


Figure 4.1: Logistic Regression

Random Forest

Different Decision Tree algorithms make up the Machine Learning (ML) driven classifier ensemble approach known as Random Forest (RF) [13]. In order to provide an ideal decision model with more accuracy than that of the single decision tree model, RF builds several decision trees while the algorithm is being trained. The notion is depicted in Figure. 4.2 below.

However, it may be used with big datasets. The mean of all decision tree methods is calculated using the Random Forest algorithm [14] [15]. The Random Forest method estimated the average of two decision tree algorithms.

$$_{j} = \frac{1}{B} + \sum^{B} fb(X') \text{ ....(2)}$$

$$b=1$$

Concerning X = $\{x1,2,x3,\dots\dots\dots\dots\dots\dots\ \ xn\}$ with respect to Y = $\{y1,y2,y3,\dots\dots\dots\dots\dots\dots\ yn\}$ with the lower to upper limit is 1 to B. Sample x' = mean of the sum of the prediction $\sum^{B}{}_{b=1} f(X^{'})$ for every summation.

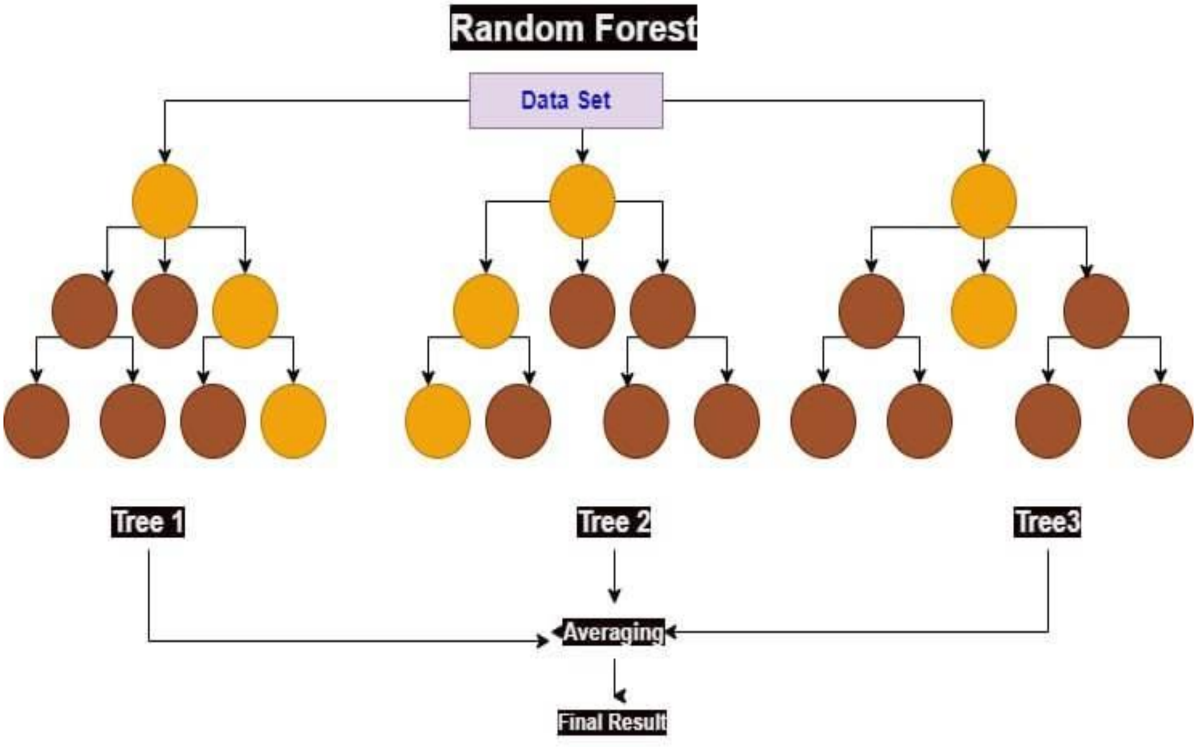


Figure 4.2: Random Forest

Gradient Boosting

The loss function is the main component of the boosting method known as Gradient Boosting (GB), which is based on Machine Learning (ML). The notion is depicted in Figure. 4.3 below. It works by combining and optimizing weak learners to reduce a model's loss function. To improve an algorithm's performance, overfitting is eliminated . Here (x) = loss function with correlated negative gradients $(-\rho i\ x\ (X))$, m = number of iterations. Feature increment (i) = 1,2,3, $\dots\ .\ .\ .$, m. Therefore, the optimal function (X) after m^{-th} iteration is shown below [16].

$$m$$

$$F\ (X) = \sum_{i=0} (x)\ \ldots\ldots\ldots.(3)$$

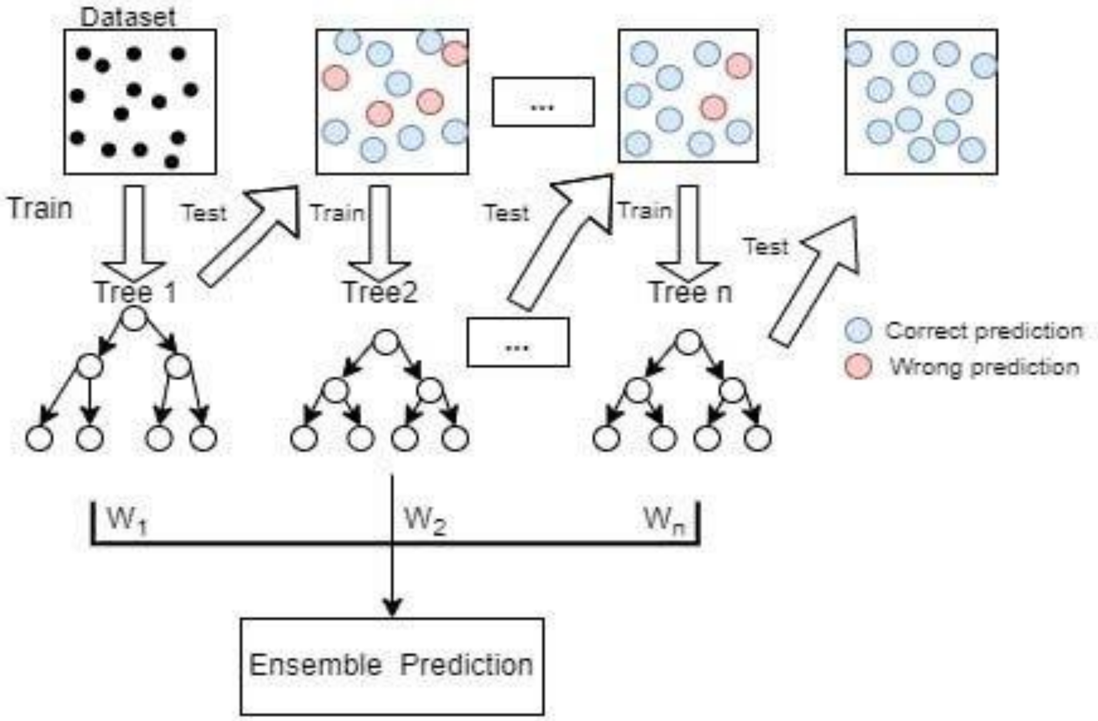


Figure 4.3: Gradient Boosting

K-Nearest

Due to its ability to treat new and old data equally, the Machine Learning (ML) algorithm K-Nearest Neighbors (KN) is frequently employed in non-parametric categorization techniques. The notion is depicted in Figure. 4.4 below. It calculates the Euclidean distance between new (x_1, x_2) and existing (y_1, y_2) data [19][20].

$$Euclidean\ Distance = \surd(x_2 - x_1)^2 + (y_2 - y_1)^2 \ldots\ldots\ldots.(4)$$

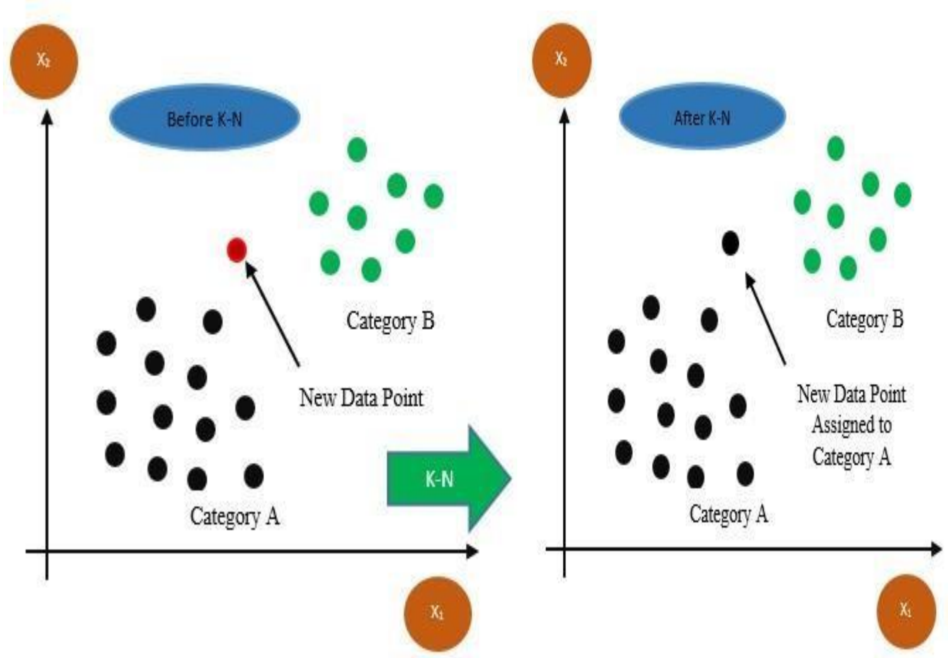


Figure 4.4: K-Nearest

4.1.2 Ensemble Methods of Machine Learning

The term "ensemble approach" refers to the use of several classifiers to turn weak classifiers into strong classifiers by producing the greatest accuracy and effectiveness. It was used in our investigation due to variable handling, bias, and uncertainty since it lowers variances, merges predictions from several models, and narrows the prediction spread [21]. In our investigation, three ensemble approaches were employed. We employed ensemble models for bagging, boosting, and voting.

Bagging

Bagging describes missing variables, decreased handling, and a decrease in variance. It improves stability for a variety of algorithms, but decision tree methods benefit the most. The notion is depicted in Figure. 4.5 below. The Bagging model's classification formula is shown below .

Here$f^{'}(x)$ is the average of $fi(x)$ for i = 1,2,3,….T.

$$f^{'}(x) = sign(\sum_{i=1}^{T} fi(x))\text{……….(5)}$$

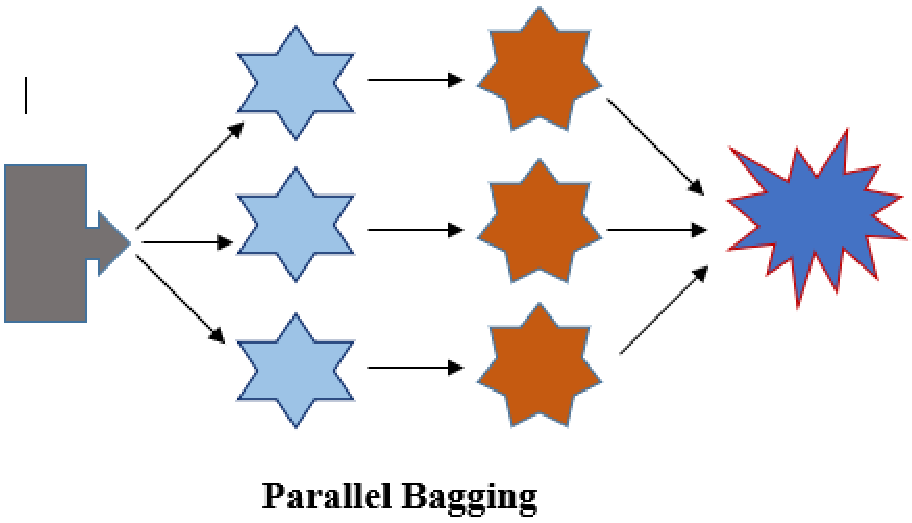


Figure 4.5: Bagging

Boosting

The term "boosting" refers to a method that converts weak learners become strong learners by using a weighted average to operate with many algorithms and create the loss functions . The notion is depicted in Figure. 4.6 below. In our work, the training and testing phase of the hybrid model construction uses the boosting approach. The formula is shown below .

Here, Υ_t = ½-ϵ_t (how much f_t is on the weighted sample).

$$\frac{1}{n}\sum_{i=1}^{n} I(y_j g(x_i) < 0) \ \le \ \prod_{t=1}^{T}\sqrt{1-4{\Upsilon_t}^2}\text{..........(6)}$$

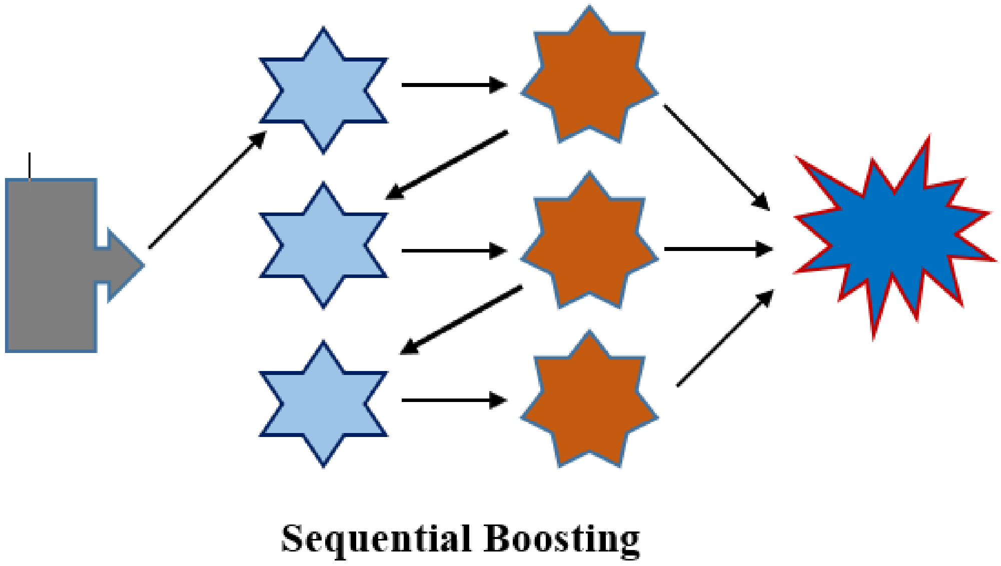


Figure 4.6: Boosting

Voting

Voting classifiers are a group of classifiers that are used to forecast the class with the best majority of votes. It implies that the model trains using many models to anticipate outcomes by aggregating the results of voting [17] [18].

The notion is depicted in Figure. 4.7 below. The formula we employed is shown below [24]..

Here, w_j = weight that can be assigned to the j^{th} classifier.

$$y' = argmax \sum_{j=1}^{m} w_j p_{ij} \text{ ....(7)}$$

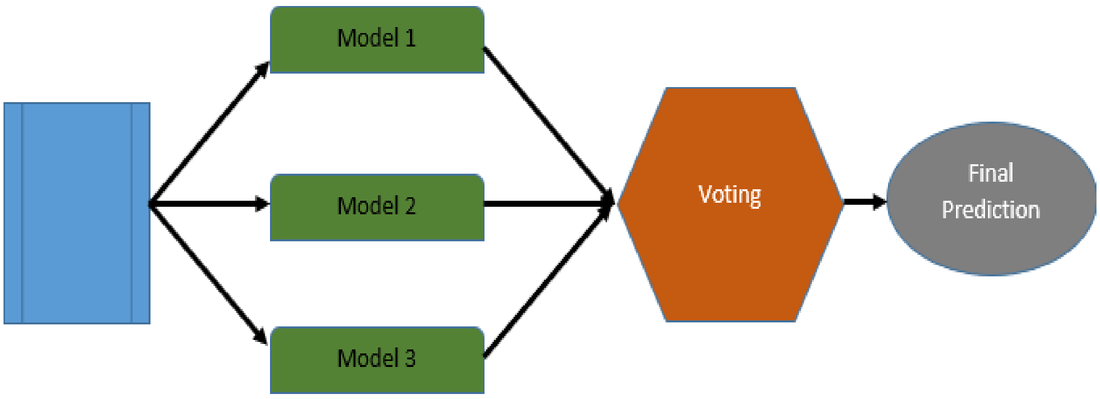
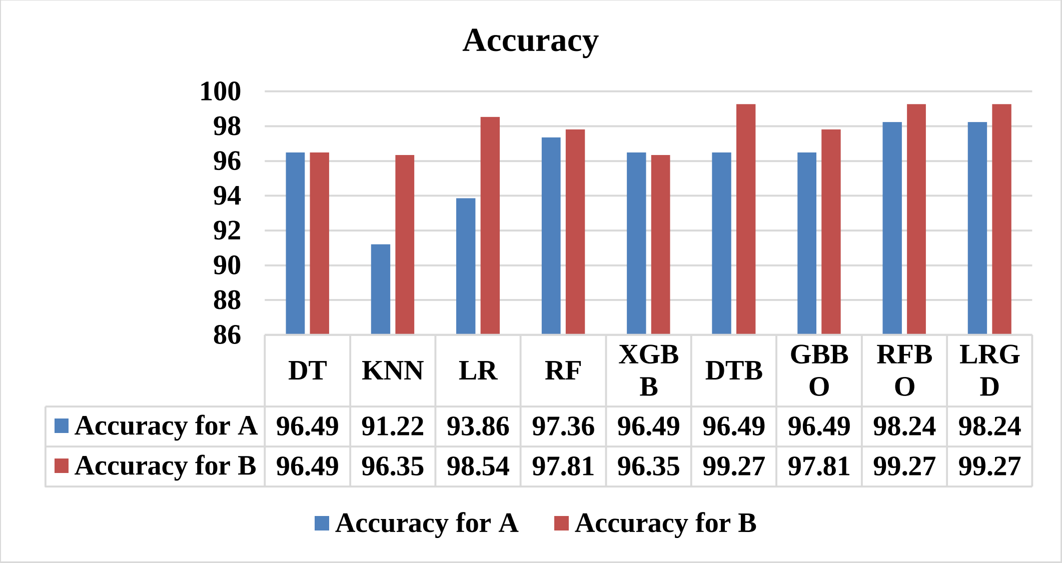


Figure 4.7: Voting

4.2 Experimental Result & Analysis

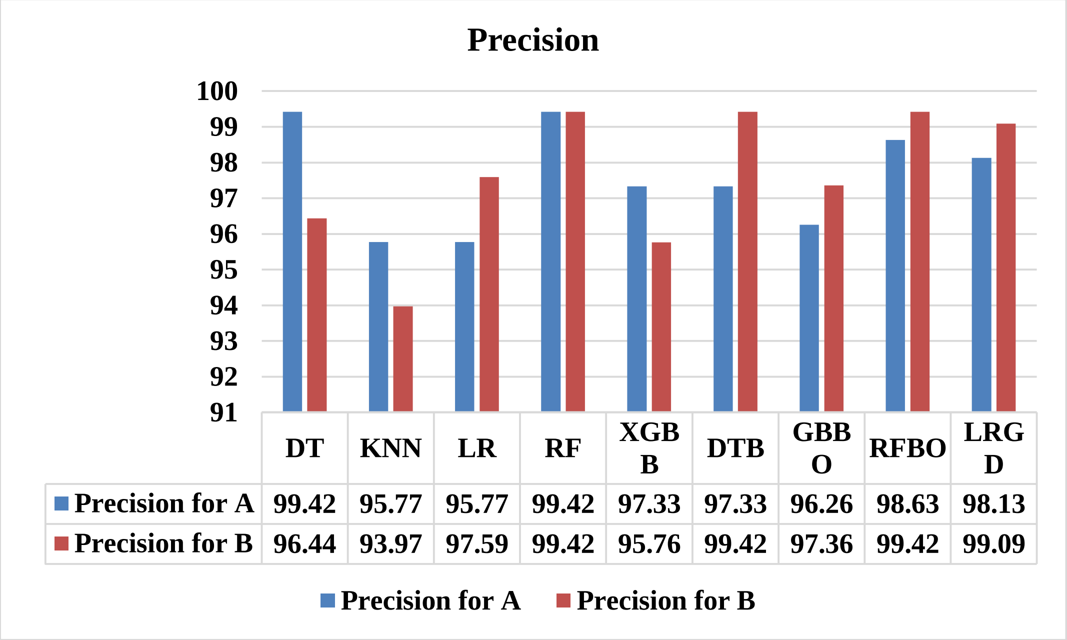
At this point, we had to assess how well the current models performed. To verify the effective performance of our suggested model, we may utilize several performance assessment measurements and approaches. These techniques calculate the total performance based on hypothetical data. In this section, we must present an analysis report based on the results of our machine learning experiments on the targeted dataset for breast cancer. We initially put our chosen dataset into practice. Our dataset has been filtered to remove any missing or erroneous values. We put a variety of algorithms into practice and evaluated how well they worked. With two separate datasets of breast cancer, we evaluated the Accuracy, Precision, Recall, and F-1 Score of our suggested algorithms. These confusion matrices for conventional methods have been measured. We tested K-Nearest, Decision Tree, Random Forest, and Logistic Regression (LR, RF, DT) (KN). With confusion matrices, we have seen many ensemble approaches in action. We assessed ensemble approaches for bagging, boosting, and voting.



	DT	KNN	LR	RF	XGB B	DTB	GBB O	RFB O	LRG D
■ Accuracy for A	96.49	91.22	93.86	97.36	96.49	96.49	96.49	98.24	98.24
■ Accuracy for B	96.49	96.35	98.54	97.81	96.35	99.27	97.81	99.27	99.27

Figure 4.8: Accuracy comparison of dataset A and B

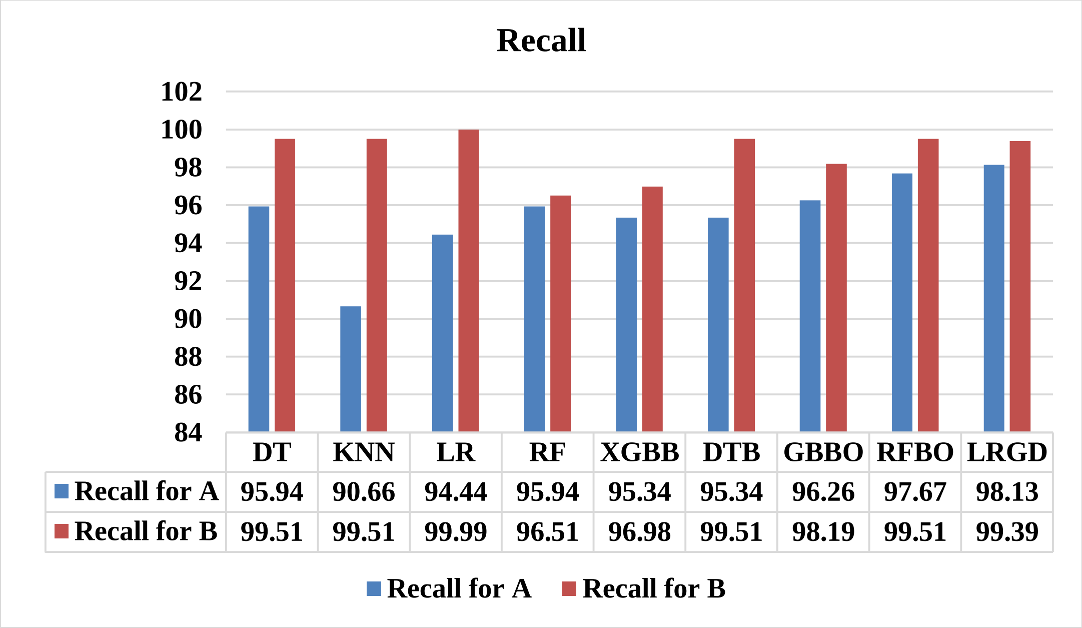
Firstly, we have measured the accuracy of both dataset A and B. The best accuracy was achieved for dataset A about RFBO and LRGD of 98.24%. For dataset B the accuracy was DTB, RFBO and LRGD of 99.27%. The output is shown in Figure 4.8.



	DT	KNN	LR	RF	XGB B	DTB	GBB O	RFBO	LRG D
■ Precision for A	99.42	95.77	95.77	99.42	97.33	97.33	96.26	98.63	98.13
■ Precision for B	96.44	93.97	97.59	99.42	95.76	99.42	97.36	99.42	99.09

Figure 4.9: Precision comparison of dataset A and B

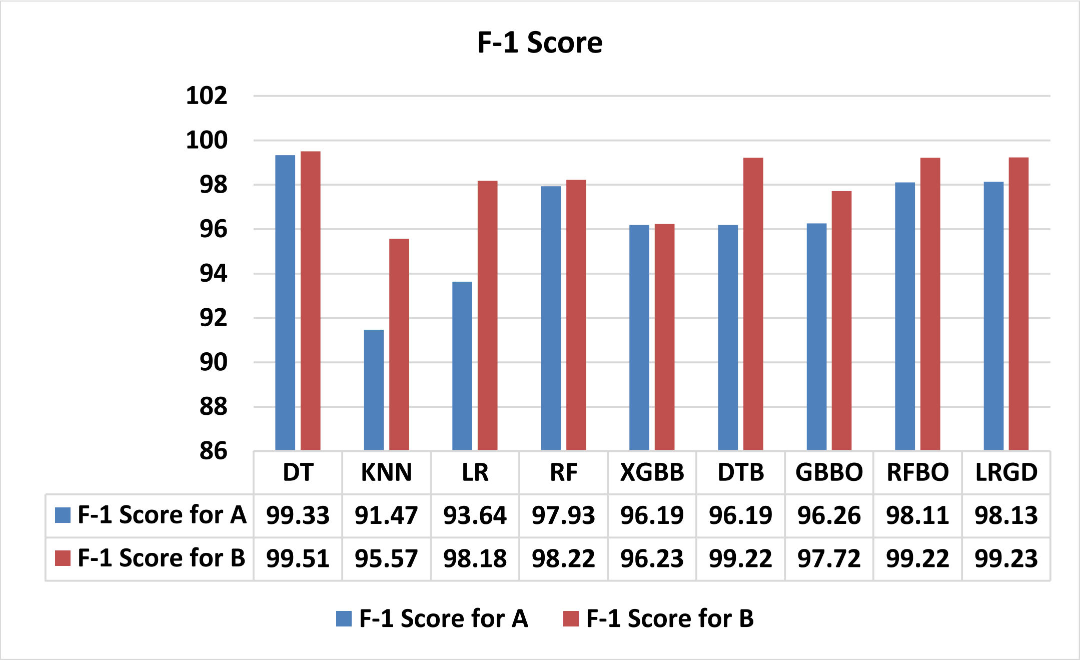
Then we have measured the Precision of both dataset A and B. The best Precision was achieved for dataset A about RF and DT of 99.42%. For dataset B the accuracy was RF and DTB of 99.42%. The output is shown in Figure 4.9.



	DT	KNN	LR	RF	XGBB	DTB	GBBO	RFBO	LRGD
🟦 Recall for A	95.94	90.66	94.44	95.94	95.34	95.34	96.26	97.67	98.13
🟥 Recall for B	99.51	99.51	99.99	96.51	96.98	99.51	98.19	99.51	99.39

Figure 4.10: Recall comparison of dataset A and B

Then we have measured the Recall of both dataset A and B. The best Recall was achieved for dataset A about LRGD of 98.13%. For dataset B the accuracy was LR of 99.99%. The output is shown in Figure 4.10.



	DT	KNN	LR	RF	XGBB	DTB	GBBO	RFBO	LRGD
🟦 F-1 Score for A	99.33	91.47	93.64	97.93	96.19	96.19	96.26	98.11	98.13
🟥 F-1 Score for B	99.51	95.57	98.18	98.22	96.23	99.22	97.72	99.22	99.23

Figure 4.11: F-1 Score comparison of dataset A and B

Then we have measured the F-1 Score of both dataset A and B. The best F-1 Score was achieved for dataset A about LRGD of 98.13%. For dataset B the accuracy was LR of 99.99%. The output is shown in Figure 4.11.

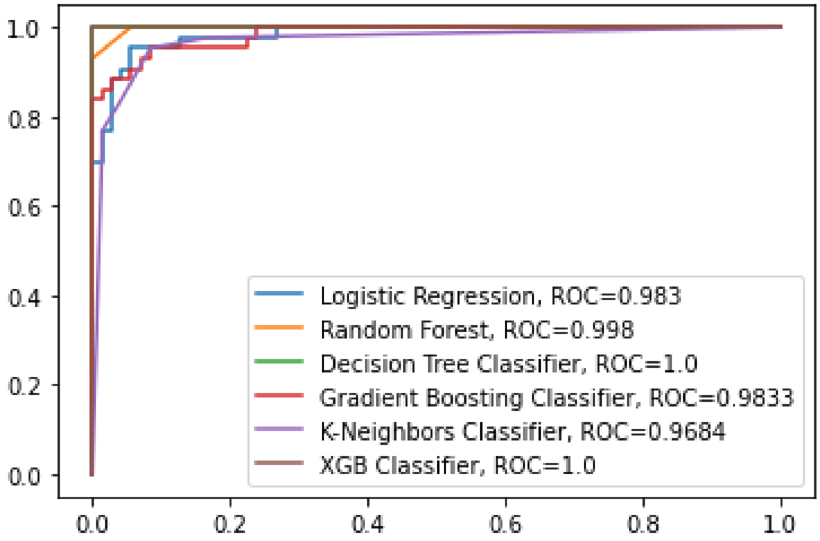


Figure 4.12: Traditional Classifier AUC Curve with dataset A

Then we have measured the AUC Curve of dataset A. The best accuracy was achieved with DT and XGBB. The output is shown in Figure 4.12.

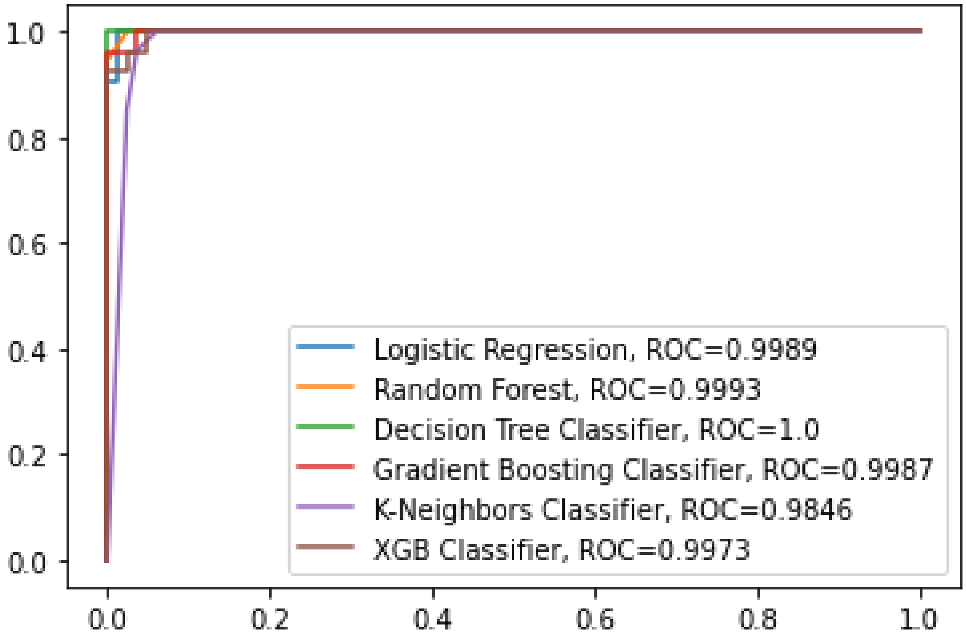


Figure 4.13: Traditional Classifier AUC Curve with dataset B

Then we have measured the AUC Curve of dataset B. The best accuracy was achieved with DT. The output is shown in Figure 4.13.

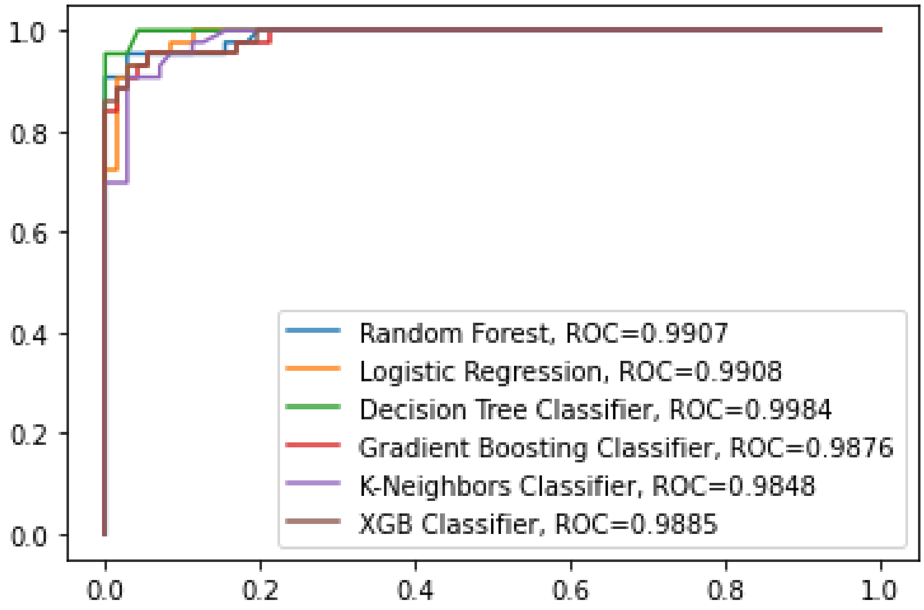


Figure 4.14: Bagging Classifiers AUC Curve with dataset A

Then we have measured the Bagging AUC Curve score of dataset A. The best accuracy was achieved with DTB. The output is shown in Figure 4.14.

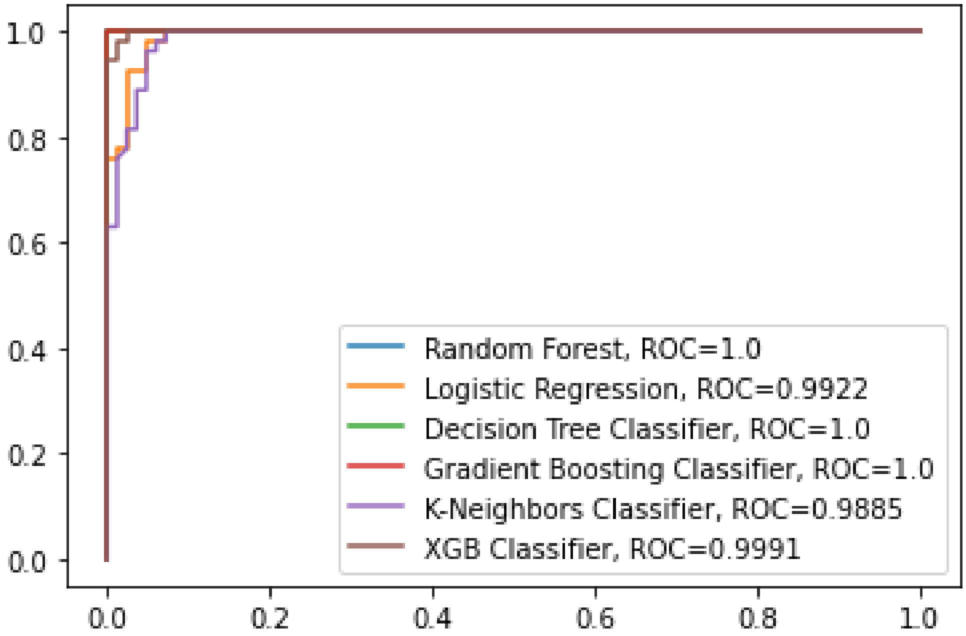


Figure 4.15: Bagging Classifiers AUC Curve with dataset B

Then we have measured the Bagging AUC Curve score of dataset B. The best accuracy was achieved with RFBO, GBBO and DTBB. The output is shown in Figure 4.15.

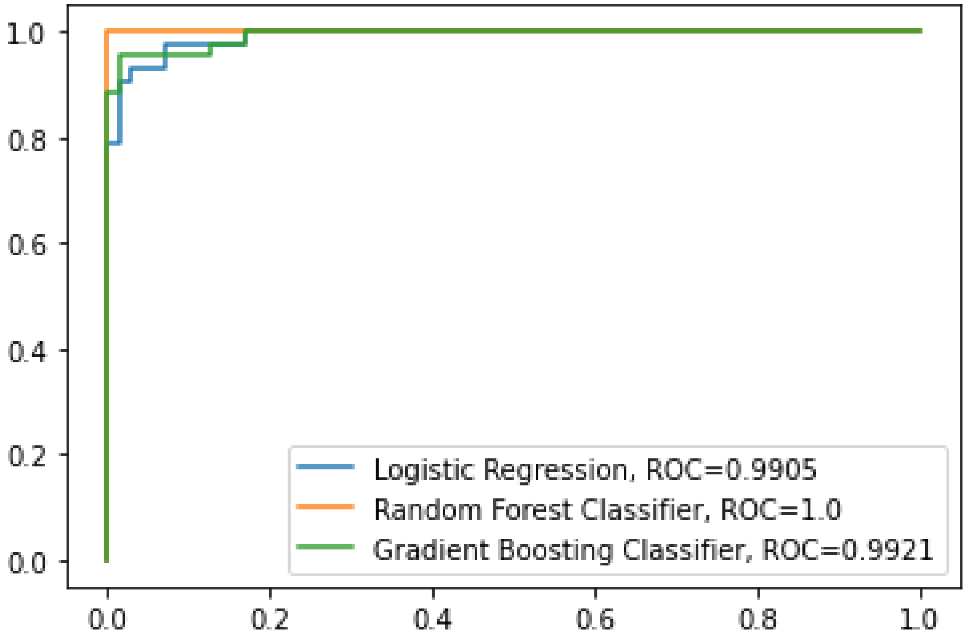


Figure 4.16: Boosting Classifiers AUC Curve with dataset A

Then we have measured the Boosting AUC Curve score of dataset A. The best accuracy was achieved with RFBO. The output is shown in Figure 4.16.

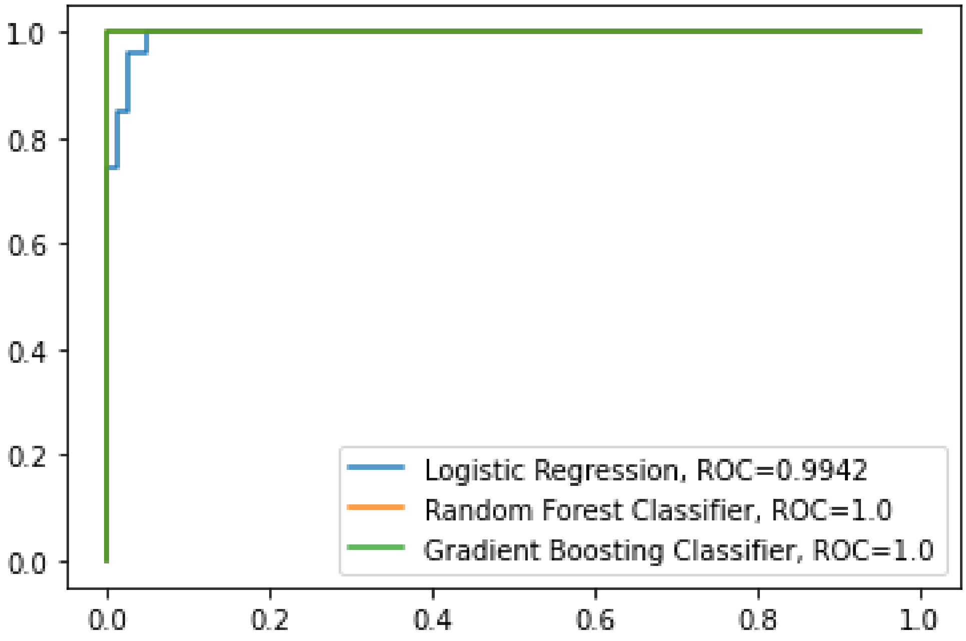


Figure 4.17: Boosting Classifiers AUC Curve with dataset B

Then we have measured the Boosting AUC Curve score of dataset B. The best accuracy was achieved with RFBO, GBBO and DTB. The output is shown in Figure 4.17.

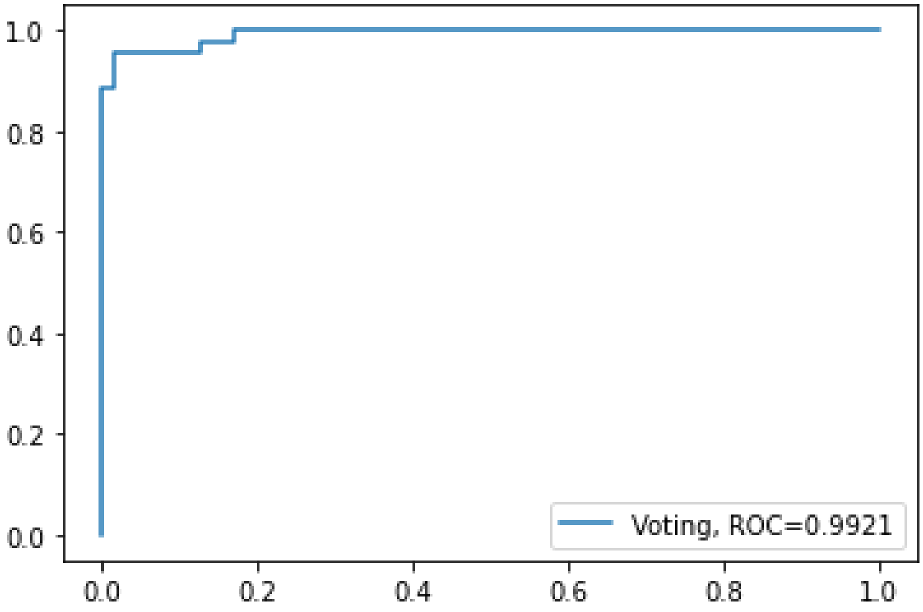


Figure 4.18: Voting Classifier AUC Curve with dataset A

Then we have measured the Voting AUC Curve score of dataset A. The best accuracy was achieved with LRGD. The output is shown in Figure 4.18.

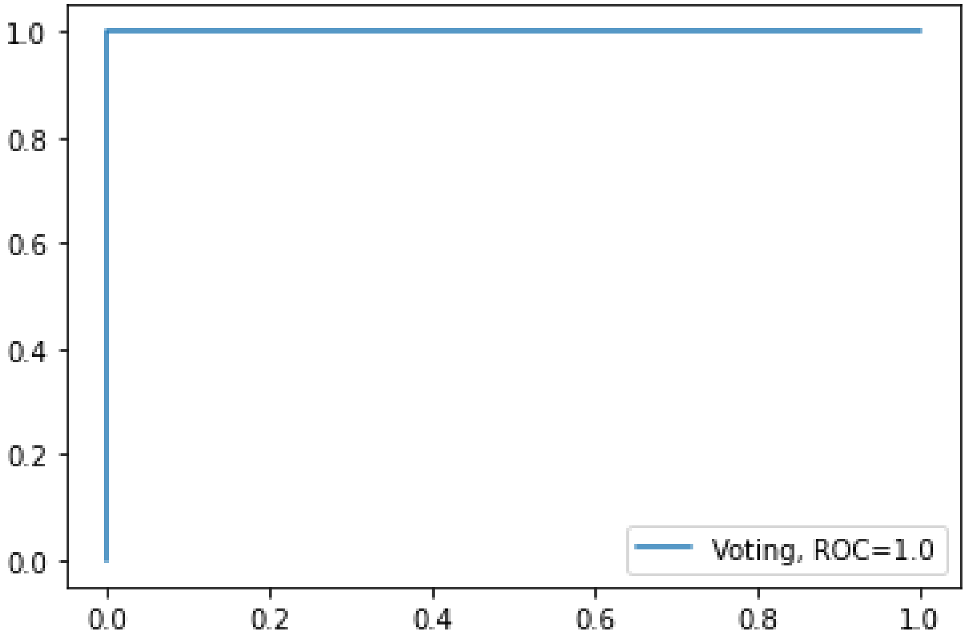


Figure 4.19: Voting Classifier AUC Curve with dataset B

Then we have measured the Voting AUC Curve score of dataset B. The best accuracy was achieved with LRGD. The output is shown in Figure 4.19.

4.3 Discussion

We shall now define the judicial system of our suggested paradigm. We have considered the F-1 score, recall, accuracy, and precision.

4.3.1 Accuracy

It speaks about the proportion of testing data predictions that were correct. Whereas accessibility of the measures with actual measurements is performed by accuracy. It is founded on a solitary variable. Accuracy only addresses deliberate mistakes. It is one of the most straightforward measurement methods for any model. For our models, we must strive for maximum accuracy.

$Accuracy$ = $(TruePositive$ + $TrueNegative)$ / $(TruePositive$ + $FalsePositive$ + $TrueNegative$ + $FalseNegative)$ ….(8)

4.3.2 Precision

It speaks about the percentage of positively expected observations that really occurred. The genuine true portion of all the cases where they correctly predicted true are identified by precision. For any type of model, a high recall might also be highly deceptive.

$Precision$ = $(TruePositive)$ / $(TruePositive$ + $FalsePositive)$ …..(9)

4.3.3 Recall

It speaks about the percentage of positively anticipated observations from a model. High accuracy, though, might occasionally be deceptive. The ratio of projected positives to all positive labels is determined by normally recall.

$Recall$ = $TruePositive$ / $(TruePositive$ + $FalseNegative)$…. (10)

4.3.4 F-1 Score

It speaks of the precision and recall harmonic means. Both the recall and precision ratios are relevant. We assume the model is quite terrible if the harmonic mean is lower.

$F - 1\ Score = [\text{True Positive} / \{\text{True Positive} + (\text{False Positive} + \text{False Negative})/2\}]$….. (11)

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

Our suggested approach offers a number of advantages, both economically and socially. Our model explores and identifies the fundamental elements or characteristics of a breast cancer patient using data from real-world cases. The effort has societal significance in that it can inform young women about the incidence of breast cancer and breast cancer prevention measures. Through accurate diagnosis and frequent checkups, we can recommend early therapy. because it is simple for them to assess the likelihood of being impacted or not and since they are aware of breast cancer. Our approach requires fewer compilations and takes less time. As a result, illness prediction is simple and accurate. In order to better diagnose breast cancer, we have studied the information in our model to determine its underlying causes. We hope that our suggested approach will be adopted and put into practice on a societal level.

5.2 Impact on Environment

The streamlined diagnosis procedures in our suggested paradigm make it particularly useful in remote locations. Through the device model, we can cut down on both time and complexity. Our methodology has no adverse consequences and is simple, so we can guarantee that the environment will gain from it as well. The patients don't need to travel to metropolitan regions to find out if they have breast cancer or not. The patient's diagnosis report may be readily supported by the prediction model, which can also forecast potential outcomes. The patient won't need to worry about local therapies because it is not very expensive to diagnose breast cancer. Because it is less complicated, it can be used by individuals at any level. Our suggested model can make it clear if a patient has breast cancer or not. Our suggested model will improve the economic and social climate. If we are given the chance to put our suggested model into practice, we are confident that it will mark a significant advancement in current medical science technology.

5.3 Ethical Aspects

Before the system is put into operation, we must take some moral safeguards to prevent the disclosure of personal information, diagnostic reports, or humor. Our suggested approach may be used for real-world breast cancer detection and therapy as well as future research endeavors. We have determined that the issue affects not only a small area or region but also the entire planet. Anyone who has breast cancer or is aware of its risk can forecast it using the suggested model.

5.4 Sustainability Plan

We can guarantee that the technologies used in breast cancer diagnostics across the world will accept our suggested model. We are optimistic that the victim ladies who can anticipate their likelihood of developing breast cancer would find our suggested approach informative. We may be inspired and prepared to aid the rural regions if we are provided with the right tools and scope for implementation. We anticipate that our suggested paradigm will be advantageous and sustainable.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In our fascinating article, we assess the influenced rate of our individuals employing algorithms. With our model, we can successfully forecast the future. The prediction system may benefit from the diagnosing technology. People can gain from understanding if they will have an impact or not. They should presumably be aware about breast cancer. If individuals use our approach, they can quickly identify the various stages of breast cancer. Assuming our suggested model can also be beneficial to diagnosis authority. We have employed a variety of widely used algorithms that are quick to construct, simple to use, and accurate.

6.2 Conclusion

The world we live in today is a contemporary one. The globe is currently a technologically advanced and simple place. The new technology is accessible to anybody in the world. With the aid of technology, what we have suggested is really simple and quick. We have made an effort to simplify the process of predicting breast cancer in humans. Our innovative models can assist our people. We have to make sure the concept is workable, and we promise to add a lot more features and work on more well-liked topics in the future. We are starting this expectation.

6.3 Implication for Further Study

We have mortality because we are human. In our daily lives, we are impacted by several ailments. While most of us have malignancies, some of us have the necessities for healing. The therapy and diagnosis technologies are more advanced and precise since we live in a developing society. The time and difficulty involved in diagnosing breast cancer sickness have decreased because to new technology. We have made an effort to provide our folks something fresh. We hope that others will adopt our model. For better performance, we have worked on a few algorithms and want to add more in the future.

REFERENCE

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

[2] F. Khan, S. Kanwal, S. Alamri, and B. Mumtaz, “Hyper-parameter optimization of classifiers, using an artificial immune network and its application to software bug prediction,” *IEEE Access*, vol. 8, pp. 20954–20964, 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] 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).

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

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

[11] 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.

[12] “What is Correlation in Machine Learning?”, Accessed: November 8, 2021, Available: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47

[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] Nahar, Nazmun, and Ferdous Ara. "Liver disease prediction by using different decision tree techniques." International Journal of DataMining & Knowledge Management Process 8, no. 2 (2018): 01-09

[16] Aljahdali, Sultan, and Syed Naimatullah Hussain. "Comparative prediction performance with support vector machine and random forest classification techniques." International journal of computer applications 69, no. 11 (2013).

[17] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms." ArtificialIntelligence Review 54, no. 3 (2021): 1937-1967.

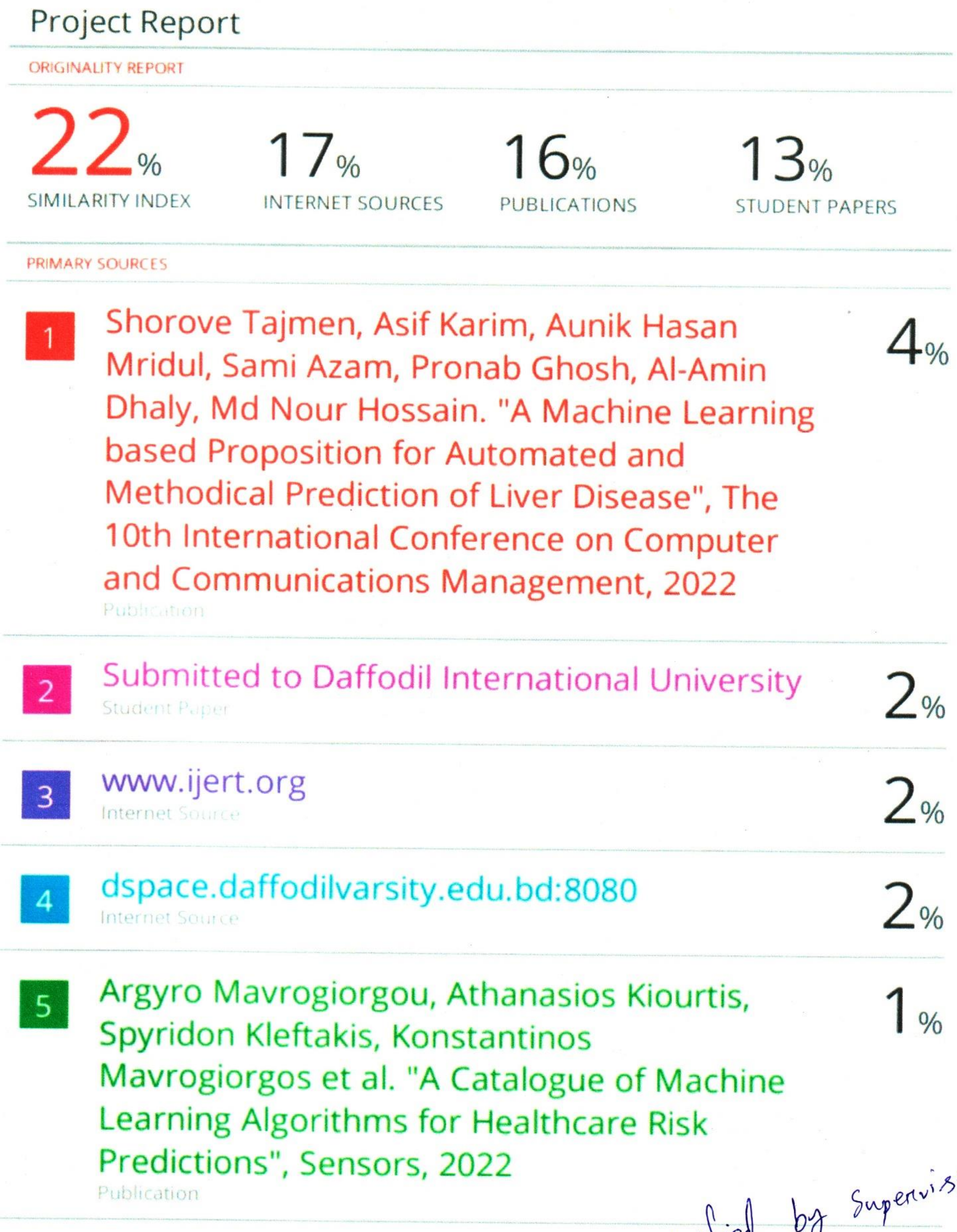
[18] Drucker, Harris, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. "Boosting and other ensemble methods." Neural Computation 6, no. 6 (1994): 1289-1301.

[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] 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.

[21] 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.

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



Project Report

ORIGINALITY REPORT

22%	17%	16%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Al-Amin Dhaly, Md Nour Hossain. "A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease", The 10th International Conference on Computer and Communications Management, 2022 Publication	4%
2	Submitted to Daffodil International University Student Paper	2%
3	www.ijert.org Internet Source	2%
4	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
5	Argyro Mavrogiorgou, Athanasios Kiourtis, Spyridon Kleftakis, Konstantinos Mavrogiorgos et al. "A Catalogue of Machine Learning Algorithms for Healthcare Risk Predictions", Sensors, 2022 Publication	1%

Verified by SuperviseR

02.01.2025