

**A Machine Learning-Based Traditional and Ensemble Technique for Predicting
Breast Cancer**

BY

Aunik Hasan Mridul
ID: 191-15-2732
AND

Md. Jahidul Islam
ID: 191-15-2753

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

Supervised By

Mushfiqur Rahman
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 “A Machine Learning-Based Technique for Predicting Breast Cancer”, submitted by **Aunik Hasan Mridul**, ID No: 191-15-2732 and **Md. Jahidul Islam**, ID No: 191-15-2753 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 05-02-2023.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Nazmun Nessa Moon

Associate Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner




Raja Tariqul Hasan Tusher

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Ahmed Wasif Reza

Professor

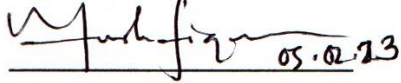
Department of Computer Science and Engineering
East West University

External Examiner

DECLARATION

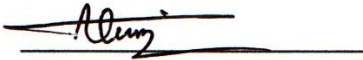
We hereby declare that this project has been done by us under the supervision of **Mushfiqur Rahman, Lecturer (Senior Scale) Department of CSE Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

 05.02.23

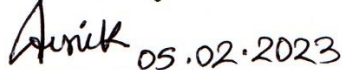
Mushfiqur Rahman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised by:



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

Submitted by:

 05.02.2023

Aunik Hasan Mridul
ID: -191-15-2732
Department of CSE
Daffodil International University



Md. Jahidul Islam
ID: -191-15-2753
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 **Mushfiqur Rahman, 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

Breast cancer is a physical disease and increasing in recent years. The topic is known widely in the recent world. Most women are suffering from problem of breast cancer. The disease is measured by the differences between normal and affected area ratio and the rate of uncontrolled increase of the tissue. Many studies have been conducted in the past to predict and recognize breast cancer. We have found some good opportunities to improve the technique. We propose predicting the risks and making early awareness using effective algorithm models. The dataset was collected from Kaggle [9]. We have used Random Forest, Logistic Regression, Gradient Boosting, and K-Nearest Classifier Algorithms. Logistic Regression and Random Forest Classifier were performed well with 98.245% testing accuracy. Other algorithms like Gradient Boosting 91.228%, and K-Nearest 92.105% testing accuracy. We also used some different ensemble models to justify the performances. We have used Bagging, Boosting, and Voting algorithms. To assign the optimal parameters to each classifier, we employed hyper-parameter tweaking. The experimental investigation demonstrated more precise breast cancer forecasts and assessed the outcomes of previous recent studies, with the greatest performance being 99.122% accuracy.

Keywords: Breast cancer, Prediction, Machine Learning, Algorithms.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval Page	ii
Declaration	iii
Acknowledgments	iv
Abstract	v
CHAPTER	
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 1.2 Motivation	1
1.3 Rationale of the Study	2
1.4 Research Questions	2
1.5 Expected Output	2
1.6 Project Management and Finance	3
1.7 Report Layout	3
CHAPTER 2: BACKGROUND	4-5
2.1 Preliminaries	4
2.2 Related Works	4
2.3 Comparative Analysis and Summary	5
2.4 Scope of the Problem	5
2.5 Challenges	5
CHAPTER 3: RESEARCH METHODOLOGY	6-13
3.1 Research Subject and Instrumentation	6

3.2 Data Collection Procedure	6
3.3 Statistical Analysis	9
3.4 Proposed Methodology	9
3.5 Implementation Requirements	13
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	14-30
4.1 Experimental Setup	14
4.2 Experimental Results & Analysis	21
4.3 Discussion	29
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	31-32
5.1 Impact on Society	31
5.2 Impact on Environment	31
5.3 Ethical Aspects	32
5.4 Sustainability Plan	32
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH	33-34
6.1 Summary of the Study	33
6.2 Conclusions	33
6.3 Implication for Further Study	33
6.4 Limitations	34
REFERENCES	35-36

LIST OF TABLES

TABLES	PAGE NO
Table 3.1: Details of the dataset	7

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Number of target values	7
Figure 3.2: Methodology	10
Figure 3.3: Correlated Features of Breast Cancer Dataset	12
Figure 4.1: Logistic Regression	15
Figure 4.2: Random Forest	16
Figure 4.3: Gradient Boosting	17
Figure 4.4: K-Nearest	18
Figure 4.5: Bagging	19
Figure 4.6: Boosting	20
Figure 4.7: Voting	21
Figure 4.8: Experimental Results of Classifiers	22
Figure 4.9: AUC-ROC Curve Analysis of Classifiers	23
Figure 4.10: Experimental Results of Bagging	24
Figure 4.11: AUC-ROC Curve Analysis of Bagging	25
Figure 4.12: Experimental Results of Boosting	26
Figure 4.13: AUC-ROC Curve Analysis of Boosting	27
Figure 4.14: AUC-ROC Curve Analysis of Voting	28
Figure 4.15: Compilation Time (Millisecond)	29

CHAPTER 1

INTRODUCTION

1.1 Introduction

As the disease is the best part of our day-to-day life, several tissues are being damaged or growing uncontrolled, known as cancer. When uncontrolled tissue or damaged tissue creates cancer in a woman's breast it is known as breast cancer. The rate of this patient is increasing at a significant rate. But the main problem is to identify or recognize the damaged area at the time of diagnosis. Machine learning can be the best part of predicting the presence of breast cancer from collected health datasets by exploring patient diagnosis records. In our work, we have explored the patient's diagnosis reports and found some important parameters to determine the disease. The dataset was about the shape and size of tissues in a woman's body and identifying the presence of cancer in her breast or not. Many other researchers have collaborated to use machine learning algorithms to identify the cancer tissue in the body. But their accuracy and technique were not suitable or smooth to predict breast cancer. To improve the prediction of breast cancer in a woman's body, we propose our technique to improve the accuracy rate. Two types of machine learning approaches are present. One of them is supervised and another is unsupervised. Supervised learning works with the data which is labeled and gives an output from input based on the example input-output pairs. The working data is training data from the dataset. Unsupervised learning starts with unlabeled data and builds a model to deal with the patterns and information that were previously undetected.

1.2 Motivation

As most women are affected by breast cancer, it is getting more common and the increasing rate is growing day by day. The reason for breast cancer, food habit, cosmetic cream, etc. is responsible. In 2020, the World Cancer Research Fund International survey says that the number of affected people was 2261419. The death number was 684996 [1]. They also said that the evidence of breast cancer was alcohol, greater birth weight, and adult-attained

@Daffodil International University

eight. There are a few research happened to predict cancer. We do a lot of studies about the prediction of breast cancer. The majority of them do not reach better accuracy. As a result, we get more driven, and we eventually discover the best accuracy of our procedure. We have made a proposed method to predict breast cancer among suspected or regular patients.

1.3 The rationale of the study

In our research, we suggested a model for predicting breast cancer in humans. Recently we noticed that our society is getting affected by this cancer. But we also noticed that we are facing a lack of awareness and diagnosis technologies. In our developing country, it is costly to detect cancer and analysis the symptoms of a patient. As a researcher, we are trying to solve the problem in machine learning approach.

1.4 Research Questions

How are the algorithms working in this proposed model?

How can we predict breast cancer?

What are the benefits of our proposed model?

What are the real-life implementation possibilities with this work?

What is the future plan for this work?

What are the precautions for this work?

How can we evaluate our model to predict breast cancer?

What is the complexity of this work?

What are the requirements for this work?

1.5 Expected Output

People are getting affected by breast cancer. People also do not know if she is affected or not. To predict or identify the disease by exploring the diagnosis report we are proposing the best method. Our method can accurately determine the effect, improve decision-making, and find breast cancer patients. It can analyze the correlated problems and also

measure the problems of daily life. It can make awareness of breast cancer disease. The proposed model can evaluate the disease within the shortest period of time.

1.6 Project Management and Finance

Our proposed model is based on low cost and is effective in our daily life. It can be a valuable asset for our nation to evaluate breast cancer. The prediction process needs common tools to implement in real life. If we use high-configuration tools, our model will give the best outputs and will run smoothly. But it can be performed while we will use lowconfigured tools.

1.7 Report Layout

Chapter 1 is covering introduction. Chapter 2 is covering the related work of the previous researchers. We need to study the introduction and motivation before starting research. As a result, we discuss Introduction which can give details knowledge about the proposed method and the motivation part can give the prediction. After completing the Introduction part, we focused on related work related to our topic and collected internal information needed for our work. After selecting the topic, In our methodology section, we picked algorithms for machine learning and apply them to our data to discover the best one. After pre-processing part, the data we test, and finally, we have evaluated the expected outcome. That was explained in our last part which is known as the conclusion.

CHAPTER 2

BACKGROUND

2.1. Preliminaries

Machine Learning ideas are used for getting the precise arrangement of breast cancer. In this segment, we attempt to explore the investigations, related to the assessment examination of the diagnosis report of the patient. These models incorporate some calculations like Random Forest, Logistic Regression, and Gradient Boosting. This segment executes profound learning models for playing out the exploration. A few researchers have utilized more than one model in their research, and there are referenced under the segment.

2.2. Related works

Some Machine Learning classifiers we have implemented for our breast cancer classification and they are suitable for our proposed work. The meaning of tree structure is Machine Learning algorithms that are based on decision tree models to run decision models [1] [2]. Researchers Rani and Dhenakaran have proposed models based on Modified Neural Network to make predictions of cancer tissue growth rate. The proposed model resulted in 97.80% accuracy [3]. Li et al. also developed an SVM classifier to predict the cancer tissue. The proposed model performed with 84.12% accuracy, 78.80% specificity, and 92.86% of sensitivity [4]. Gomez-Flores and Hernandez-Lopez proposed a model to detect cancer tissue with an 82.0% AUC score [5]. Liu et al. developed an SVC model to acquire the classification of breast cancer tissue with 67.31% accuracy, 47.62% sensitivity, and 80.65% specificity [6]. Irfan et al. also proposed CNN and SVM models to classify breast cancer with a precision rate of about 98.9% [7]. SVM, AdaBoost, Naive Bayesian, K-NN, Perceptron, and Extreme Learning Machine models were proposed by Lahoura et al. with 98.68% accuracy, 91.30% recall, 90.54% precision, and 81.29% F1-score [8].

2.3. Comparative Analysis and Summary

Nowadays there are very commonly using the model named Machine Learning. We had to do the difficult task to find our relative work. All related works had low accuracy and all other low resulting models. We had to use various types of Machine Learning models to find the best accuracy in predicting the dataset. We had faced using high-configuration devices to run the models. We utilized some solo calculations to resulting the classification rates. The complex models can create long-time runtime by adding costly GPUs.

2.4. Scope of the Problem

The problem was about easing and familiarizing the diagnosis system of breast cancer among women. As we have found huge related works with machine learning, we tried to achieve the best accuracy with our proposed model. We had limited scope to improve the mechanism but we could implement the idea with common tools to minimize the diagnosis of breast cancer.

2.5 Challenges

The dataset was collected from kaggle[9]. The data was fully usable and easy to implement. After completing the data collection, we need to manually check it so there is any missing data in the dataset. We have dropped two unnamed columns, and they were not useful for us. No one has achieved our accuracy with this dataset.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrument

We have implemented some different algorithms and hybrid models to find the best accuracy from the dataset. We needed some instruments like good configuration devices with the best GPUs. We have used python programming language, and tools like the anaconda, Jupiter Notebook, and Google Collaboratory. It permits the writing and performing of arbitrary python code through the browser. Our experiments were run with AMD Ryzen 5 3600 6-core Processor 3.59 GHz speed 8 GB RAM Windows 10 Pro the 64-bit operating system.

3.2 Data Collection Procedure

As the dataset was taken from Kaggle, dataset was almost ready for implementation. The column and the row size are 32 and 569 respectively. The diagnosis column classifies the rate of breast cancer. All the attributes were important to predict breast cancer. Patients are separated into 2 conditions Malignant and Begin. Here Malignant was used as M and Begin was used as B. We have converted these values with nominal values. There 0 denotes 'B' and 1 denotes 'M'. We have calculated the rate of these two conditions. There were 357 patients in Begin stage and the rest 212 patients were in the Malignant stage. The ratio is shown below in figure 3.1. We have separated the dataset into two parts. They are trained and tested. We have selected 80% for the training part and another 20% for the test part.

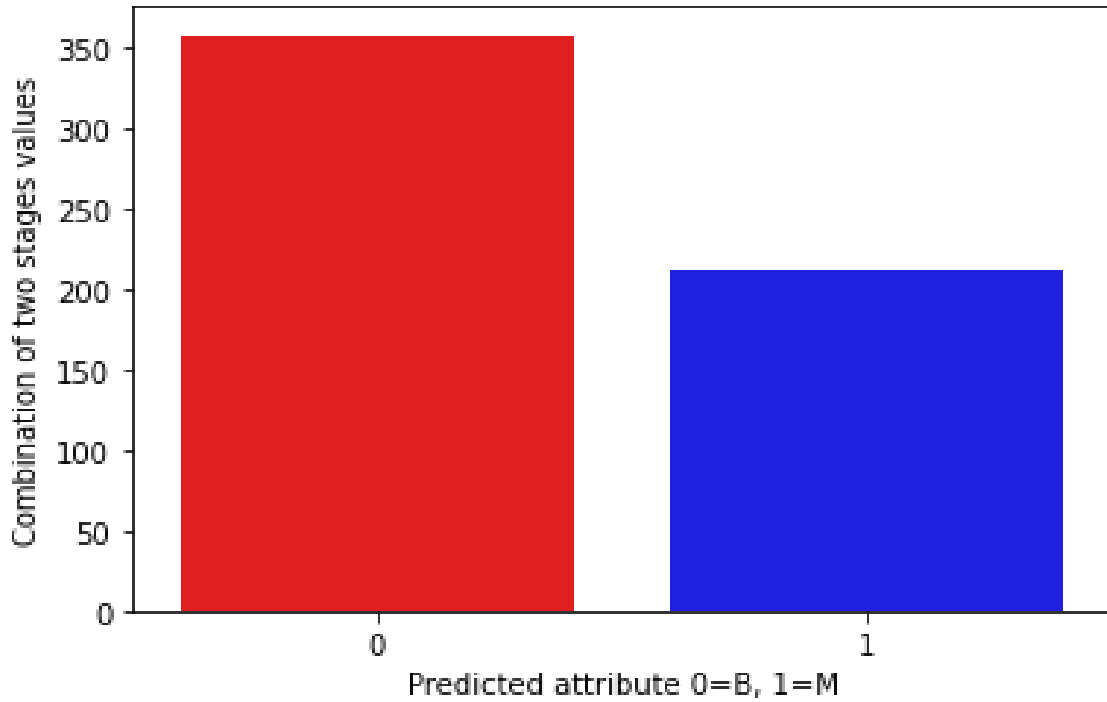


Fig 3.1: Number of target value

The dataset has 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.

Table 3.1: Details of the dataset

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

3.2.1 Categorical Data Encoding

The term data encoding system of categorical method means the technique of transformation of data of categorical into a value of nominal. As we have to input and output only numeric data in machine learning, the categorical encoding method played important role in our study. We had a gender column to run the categorical data encryption method.

3.2.2 Missing Value Imputation

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

3.2.3 Handling Imbalanced Data

It is the process of adjusting the category distribution of a source data. It manages the dataset 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 method for normalizing a large number of independent different data. When there are no local variables, the MinMax measure scales all relevant data in the region $[0, 1]$, otherwise it scales all relevant data in the range $[-1, 1]$.

3.3 Statistical Analysis

The analysis part is an important part of any kind of research work. This segment depends on developing and evaluating the algorithms I have used. As we have chosen comma separated valued (CSV) file to implement, we have to follow some steps to clean the dataset and make it usable. We have used several steps like data gathering, pre-process, etc.

In this study, we have used different four types of algorithms like Random Forest (RF), Logistic Regression (LR), K-Neighbors Classifier (KN), and Gradient Boosting Classifier (GB). The best accuracy was LR and RF about 98.25%. Then Bagging, Boosting and Voting algorithms were used and we got the best score in RF was 99.122%. We employed 10-fold cross-validation as well as hyperparameter tweaking.

3.4 Proposed Methodology

Flow chart:

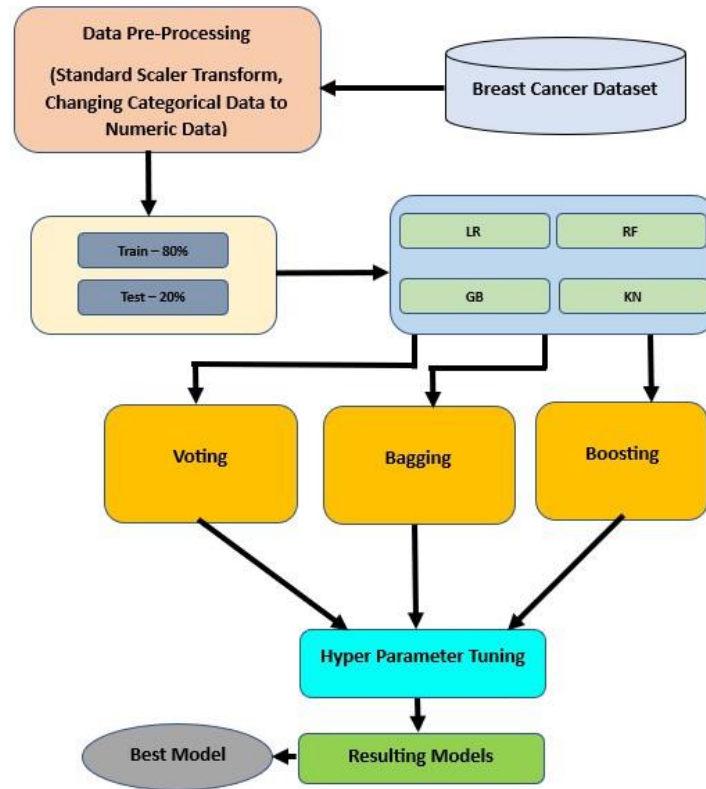


Fig 3.2: Methodology

In this part, we have used a process diagram to predict breast cancer. At first, we inserted the dataset into the system to be trained and tested. Then we implemented data preprocessing like Standard Scaler Transform. Changing Categorical Data to Numeric Data. We have used 80% in training and 20% in the testing part. Then we implanted Algorithms and evaluated the results. Then we implemented ensemble algorithms to give the best accuracy of predictions. The ensemble algorithms are Bagging, Boosting, and Voting. Then we evaluated the results of the implemented ensemble algorithms. Then we checked the output through Hyper Parameter Tuning. Then we evaluated the implemented models and took decisions with result analysis. The full process of our method is showed in Fig 3.2.

The identification of internal dependencies between variables that alter their connection to one 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]. All the linked characteristics with the anticipated attribute "Diagnosis" were displayed in Fig. 3.3.

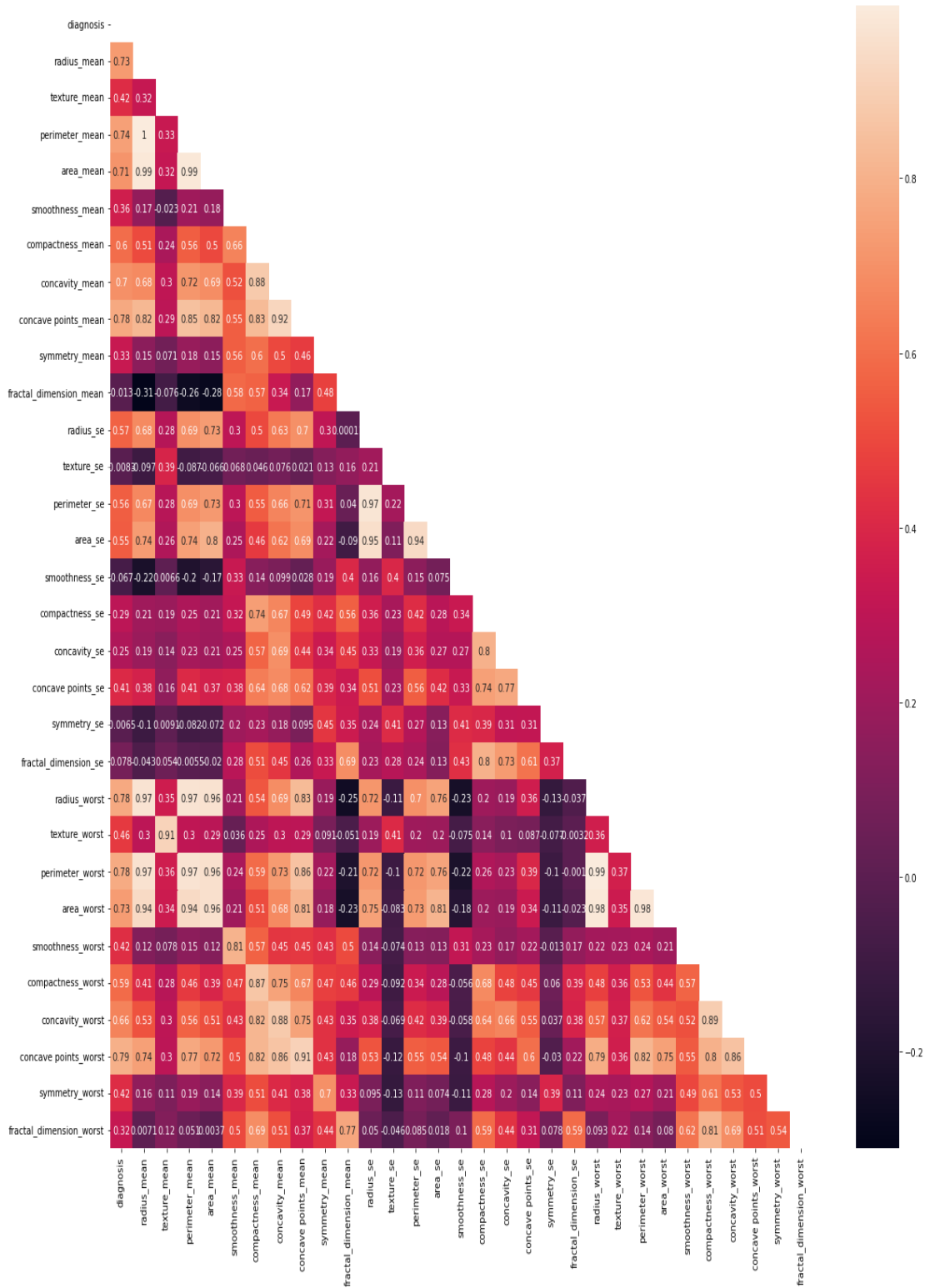


Fig 3.3: Correlated Features of Breast Cancer Dataset

3.5 Implementation Requirements

To implement our proposed model, we need data sources to study or train the model. We have to clear the dataset to run smoothly. The dataset will be cleaned with several filtering processes. Then we implemented data pre-processing like Standard Scaler Transform. Changing Categorical Data to Numeric Data. We have used 80% in training and 20% in the testing part. Then we implemented Algorithms and evaluated the results. Then we implemented ensemble algorithms to give the best accuracy of predictions. The ensemble algorithms are Bagging, Boosting, and Voting. Then we evaluated the results of the implemented ensemble algorithms. Then we checked the output through Hyper Parameter Tuning. Then we evaluated the implemented models and took decisions with result analysis. 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.

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), Gradient Boosting (GB), and K-Nearest Neighbors (KN).

Logistic Regression

Logistic Regression (LR) is a Machine Learning (ML) based classifier algorithm where the class label has two categories, there are yes or no like a binary (0/1). Logistic regression is useful in discrete variables but it allows the mixed value of continuous variables and discrete predictors [11]. The concept is shown in below Fig 4.1. Logistic Regression accepts the method of supervised machine learning. The basic equation is shown below [12].

$$h\theta(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

' $h\theta(x)$ ' is the output of the logistic function, where $0 \leq h\theta(x) \leq 1$

' β_1 ' is the slope

' β_0 ' is the y-intercept

'X' is the independent variable

$(\beta_0 + \beta_1 X)$ – derived from the equation of a line Y (predicted) = $(\beta_0 + \beta_1 X) + \text{Error}$.

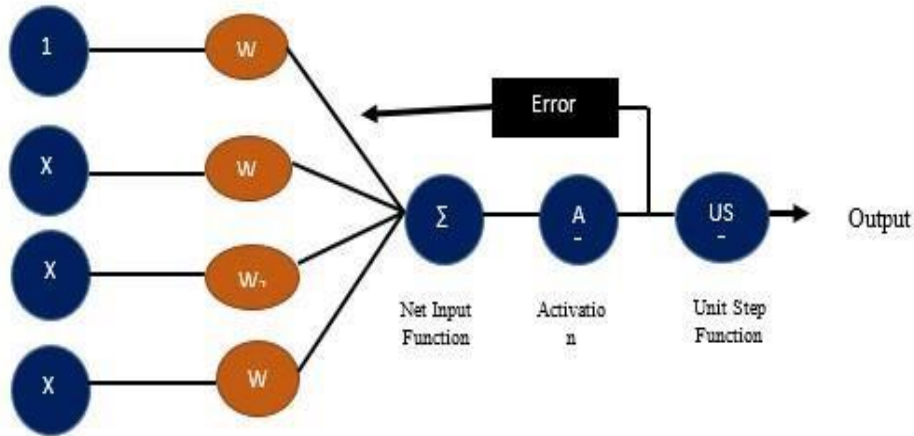


Fig 4.1: Logistic Regression

Random Forest

Random Forest (RF) is a Machine Learning (ML) based classifier ensemble method that consists of different Decision Tree algorithms [13]. RF creates several multiple decision trees during the time of algorithm training to result in an optimal decision model which can result in the best accuracy than the single decision tree model. The concept is shown in below Fig 4.2.

But it is applicable in large datasets. The Random Forest algorithm calculates the mean of total decision tree algorithms [14] [15].

$$j = \frac{1}{B} + \sum_{b=1}^B fb(X')$$

Concerning $X = \{x_1, x_2, x_3, \dots, x_n\}$ with respect to $Y = \{y_1, y_2, y_3, \dots, y_n\}$.

Sample x' = mean of the sum of the prediction $\sum_{b=1}^B fb(X')$ for every summation.

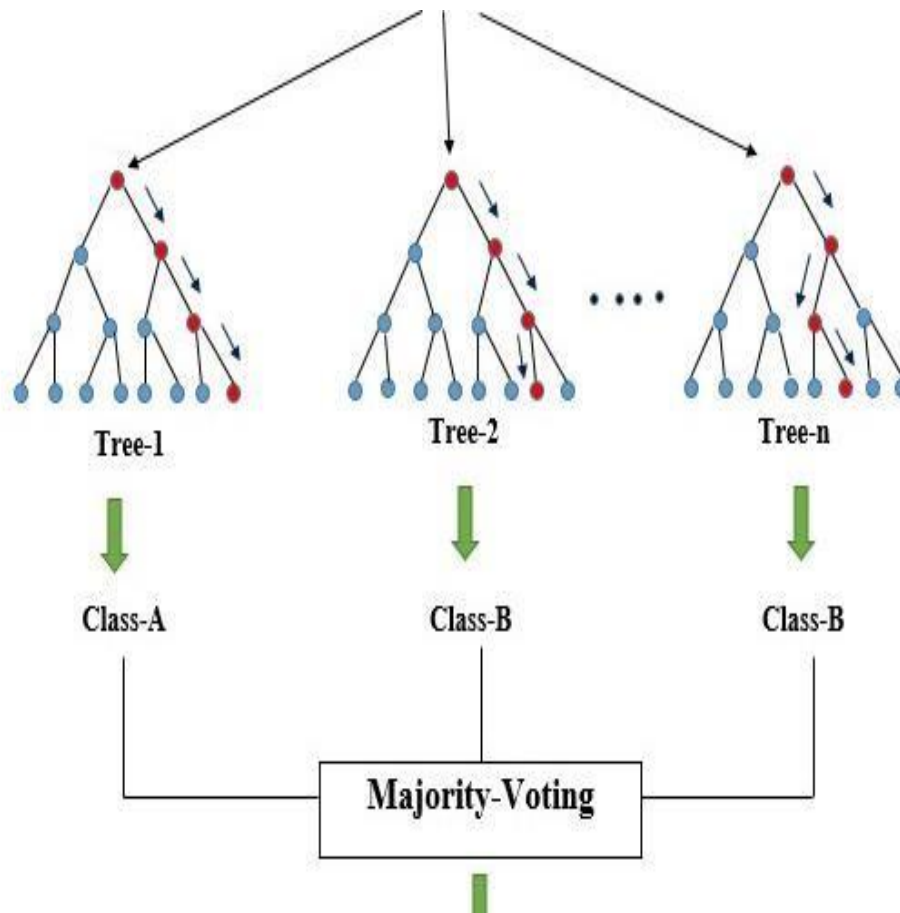


Fig 4.2: Random Forest

Gradient Boosting

Gradient Boosting (GB) is a Machine Learning (ML) based boosting algorithm that is composed of the loss function. The concept is shown in below Fig 4.3. It works with the combination and optimization of weak learners to decrease the loss function of a model. It removes overfitting to increase the performance of an algorithm. Here $f_i(x)$ = loss function with correlated negative gradients ($-\rho_i x g_m(X)$), m = number of iterations.

Feature increment (i) is equal to 1, 2, 3,..., m. As a result, the best function $F(X)$ after m iterations is displayed below [16].

$$F(X) = \sum_{i=0}^m f_i(x)$$

Here, g_m = the path of loss function's fast decreasing $F(X) = F_{n-1}(X)$ the decision tree's target is to solve the mistakes by previous learners [17][18]. The negative gradient for the m^{th} iteration is shown below.

$$F(X) = \sum_{i=0}^m f_i(x)$$

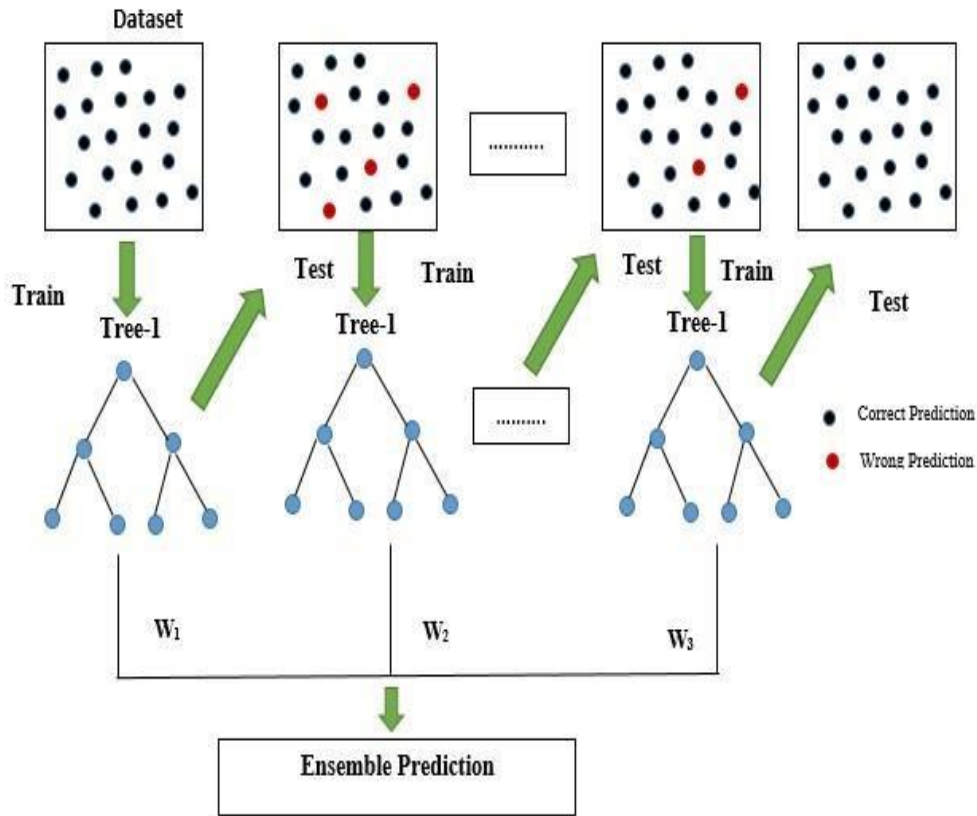


Fig 4.3: Gradient Boosting

K-Nearest

K-Nearest Neighbors (KN) is a Machine Learning (ML) algorithm that is mostly used in non-parametric classification methods as it allows the equivalence of new and existing data. The concept is shown in below Fig 4.4. It calculates the Euclidean distance between new (x_1, x_2) and existing (y_1, y_2) data [19][20].

$$Euclidean\ Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

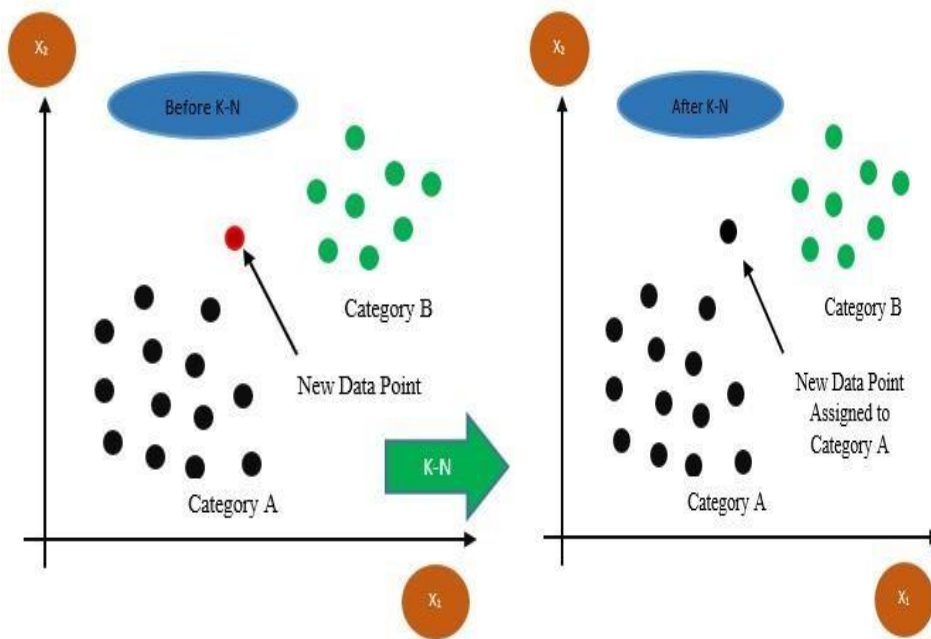


Fig 4.4: K-Nearest

4.1.2 Ensemble Methods of Machine Learning

The ensemble method refers to the multiple classifiers that result in the best accuracy and effectiveness for the weak classifiers to create them as a strong classifier. It was applied in our study because of variable handling, uncertainty, and bias, reduces variances, combines prediction of multiple models, and reduces the spread of predictions [21] [22]. Three ensemble methods were used in our study. We used Bagging, Boosting, and Voting ensemble, models.

Bagging

Bagging refers to the decrease of variance, diminishing handling, and missing variables. It enhances stability for different algorithms but is mainly applicable to decision tree algorithms. The concept is shown in below Fig 4.5. The formula of the Bagging model for classification is shown below [23].

Here $f'(x)$ is the average of $f_i(x)$ for $i = 1, 2, 3, \dots, T$. $f'(x) = \text{sign}(\sum_{i=1}^T f_i(x))$

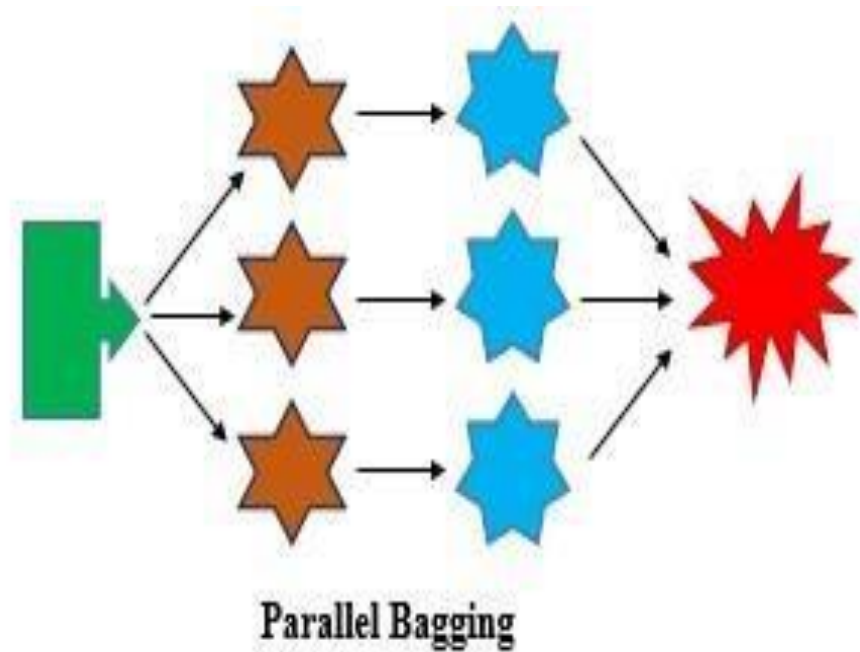


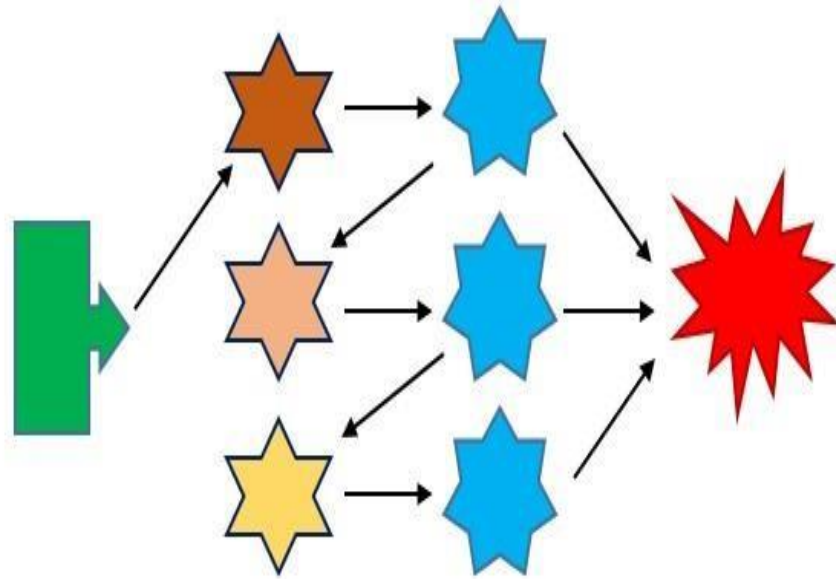
Fig 4.5: Bagging

Boosting

Boosting refers to the technique that uses a weighted average to work with several algorithms and makes the weak learners strong learners boost the accuracy of independent models creating the loss functions [24]. The concept is shown in below Fig 4.6. In our study, the boosting method is applied in the training and testing portion to construct the hybrid model. The equation is shown below [23].

Here, $\gamma_t = \frac{1}{2} - \epsilon_t$ (how much f_t is on the weighted sample).

$$\frac{1}{n} \sum_{i=1}^n I(y_j g(x_i) < 0) \leq \prod_{t=1}^T \sqrt{1 - 4\gamma_t^2}$$



Sequential Boosting

Fig 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.

The concept is shown in below Fig 4.7. The equation we have used is shown below [24] [25] [26].

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

$$y' = \operatorname{argmax} \sum_{j=1}^m w_j p_{ij}$$

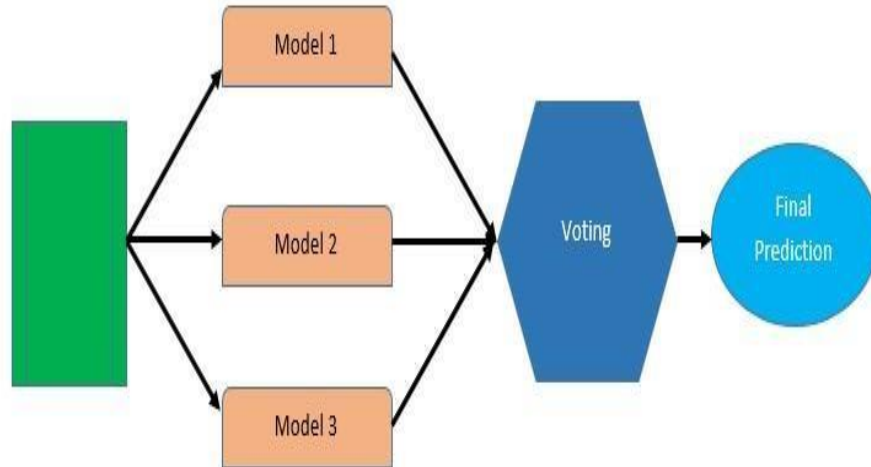


Fig 4.7: Voting

4.2 Experimental Result & Analysis

At this stage, we had to evaluate the performance of existing models. We can use some performance evaluation measures methods to check our proposed model's efficient performance. These methods estimate the overall performance which is performed on the unseen data. In this segment, we need to show an analysis report based on our experimental outputs of the machine learning models of our targeted Breast cancer dataset. At first, we implemented our chosen dataset. We have calculated the missing or incorrect values and filtered these from our dataset. We have implemented some different algorithms and analyzed their performances. We have measured Confusion matrices Accuracy, Precision, Recall and F-1 Score for our proposed algorithms. We have measured these confusion matrices for traditional algorithms. We have evaluated for Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB) and K-Nearest (KN). We have observed different ensemble techniques with confusion matrices. We have evaluated for Bagging, Boosting and Voting ensemble techniques.

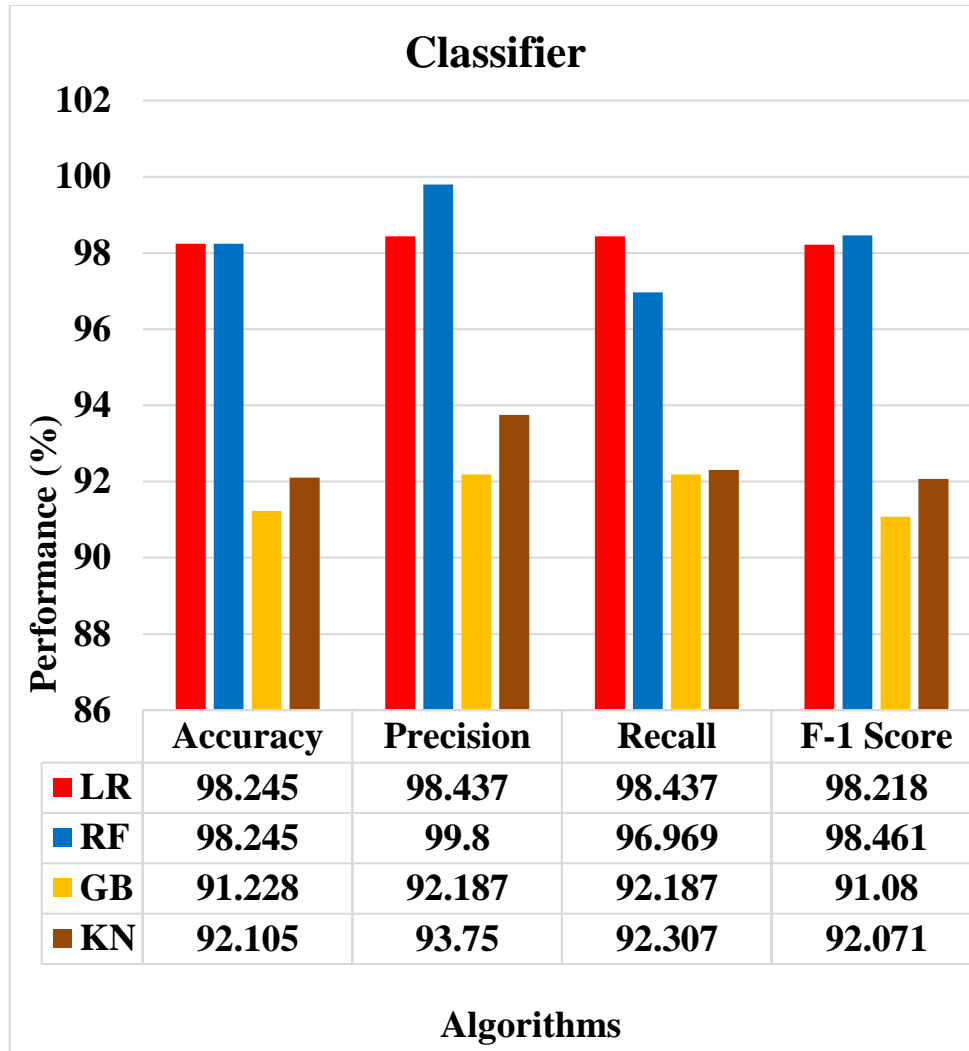


Fig 4.8: Experimental Results of Classifiers

Firstly, we considered the performances of algorithmic classifiers, the best accuracy had obtained at 98.245% using two different algorithms named Random Forest (RF) and Logistic Regression (LR). Gradient Boosting (GB) and K-Nearest (KN) algorithms achieved the accuracy respectively 91.228% and 92.105%. The output is shown in below Fig 4.8. The best precision score of 99.8% had obtained by applying Random Forest (RF). Logistic Regression (LR), Gradient Boosting (GB) and K-Nearest (KN) algorithms achieved the Precision respectively 98.437%, 92.187% and 93.75%. The best recall score of 98.437% had obtained by applying Logistic Regression (LR). Random Forest (RF) Gradient Boosting (GB) and K-Nearest (KN) algorithms achieved the Recall respectively

96.696%, 92.187% and 92.307%. The best F-1 score of 98.461% had obtained by applying Random Forest (RF). Logistic Regression (LR), Gradient Boosting (GB) and K-Nearest (KN) algorithms achieved the F-1 Score respectively 98.218%, 91.08% and 92.071%.

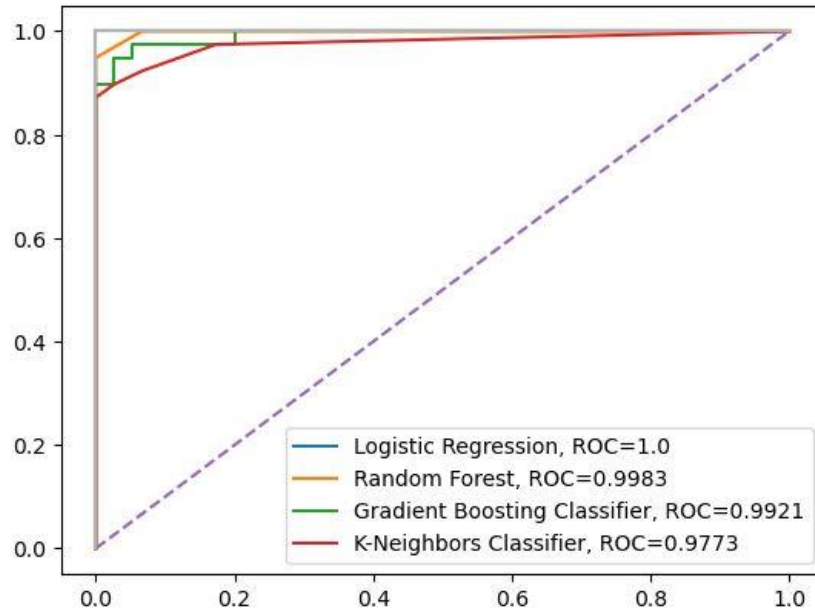


Fig 4.9: AUC-ROC Curve Analysis of Classifiers

The AUC-ROC Curve depicted in Fig 4.9, also showed that Logistic Regression (LR) had acquired the best score of 99.99%. But Random Forest (RF) had acquired the 99.83% which is so much closer to the LR. Hence according to the above analysis as well as the detailed results with graphical representation, Random Forest Classifier can be stamped as the best algorithmic classifier. Gradient Boosting (GB) and K-Nearest (KN) had acquired 99.21% and 97.73% respectively.

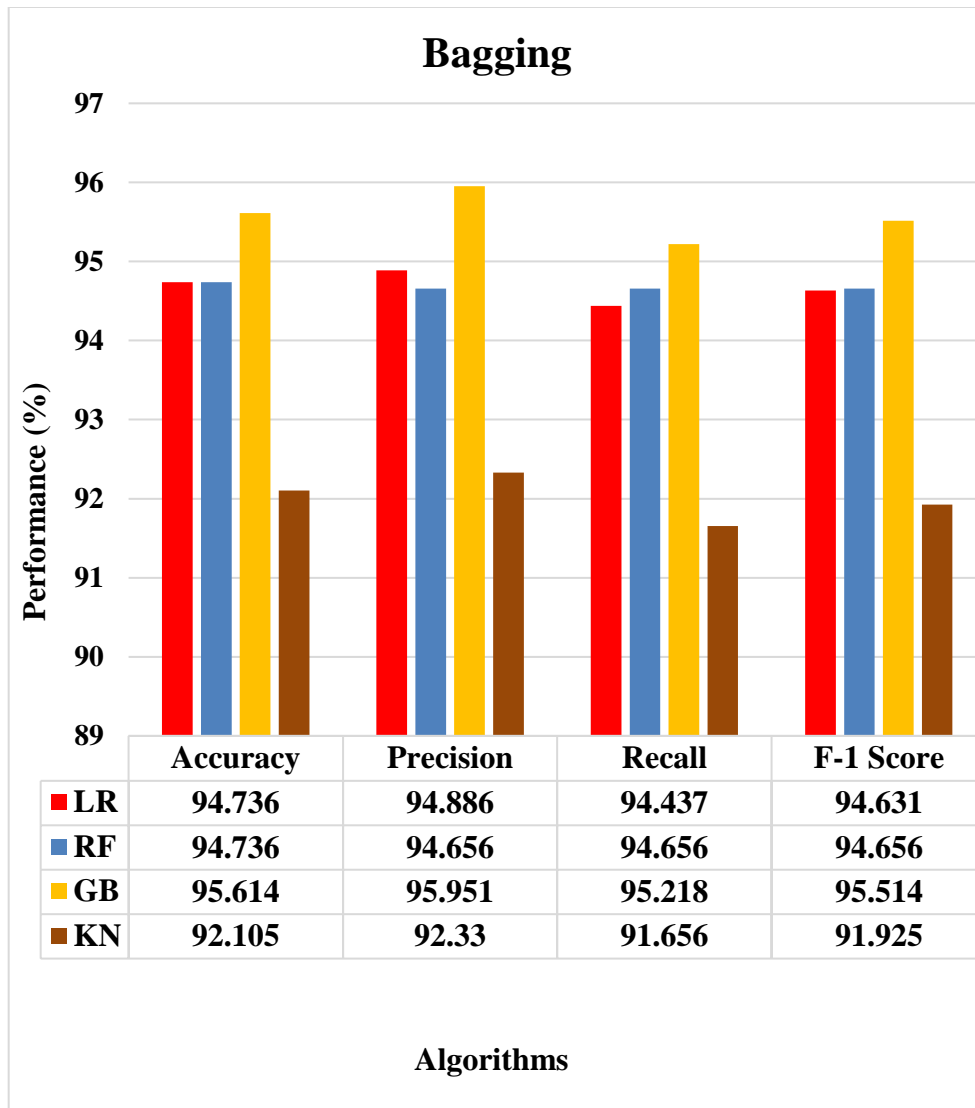


Fig 4.10: Experimental Results of Bagging

Secondly, the performance obtained by bagging classifiers had shown in Fig 4.10. The best accuracy of 95.614% had obtained by Gradient Boosting (GB) but the second largest accuracy 94.736% had obtained by Logistic regression (LR) and Random Forest (RF). KNearest (KN) algorithm had acquired 92.105% accuracy. Logistic Regression had 94.886% precision, 94.437% recall and 94.631% F-1 score. Between LR and RF the best outcome had come from Random Forest (RF) with about 94.656% precision, 94.656% recall, and 94.656% F-1 score. Hence, we can assume that Random Forest gained the second highest results.

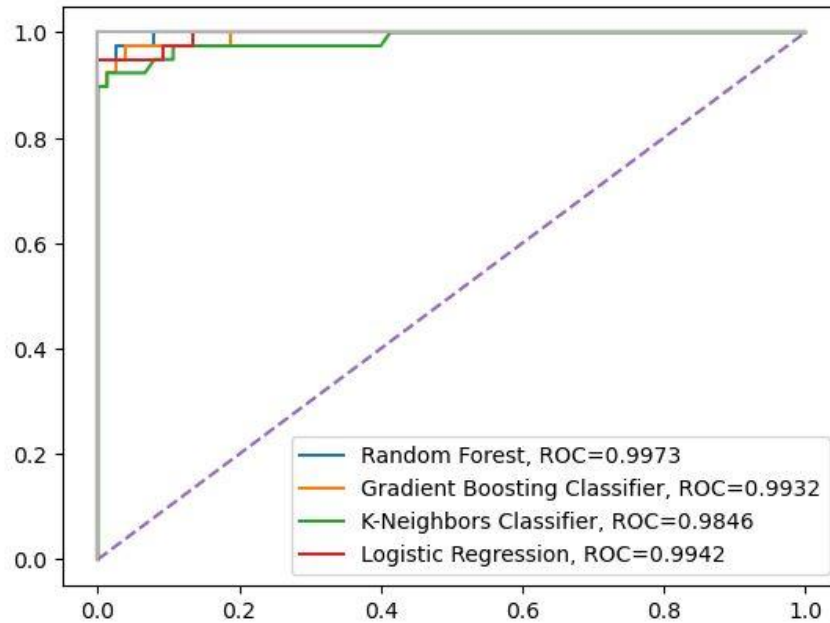


Fig 4.11: AUC-ROC Curve Analysis of Bagging

The AUC-ROC Curve depicted in Fig 4.11, also showed that Random Forest (RF) had acquired the best score of 99.83%. But Logistic Regression (LR) had acquired the 99.42% which is so much closer to the RF. Hence according to the above analysis as well as the detailed results with graphical representation, Random Forest Classifier resulted in the best algorithmic classifier. Gradient Boosting (GB) and K-Nearest (KN) had acquired the score respectively 99.32% and 98.46%.

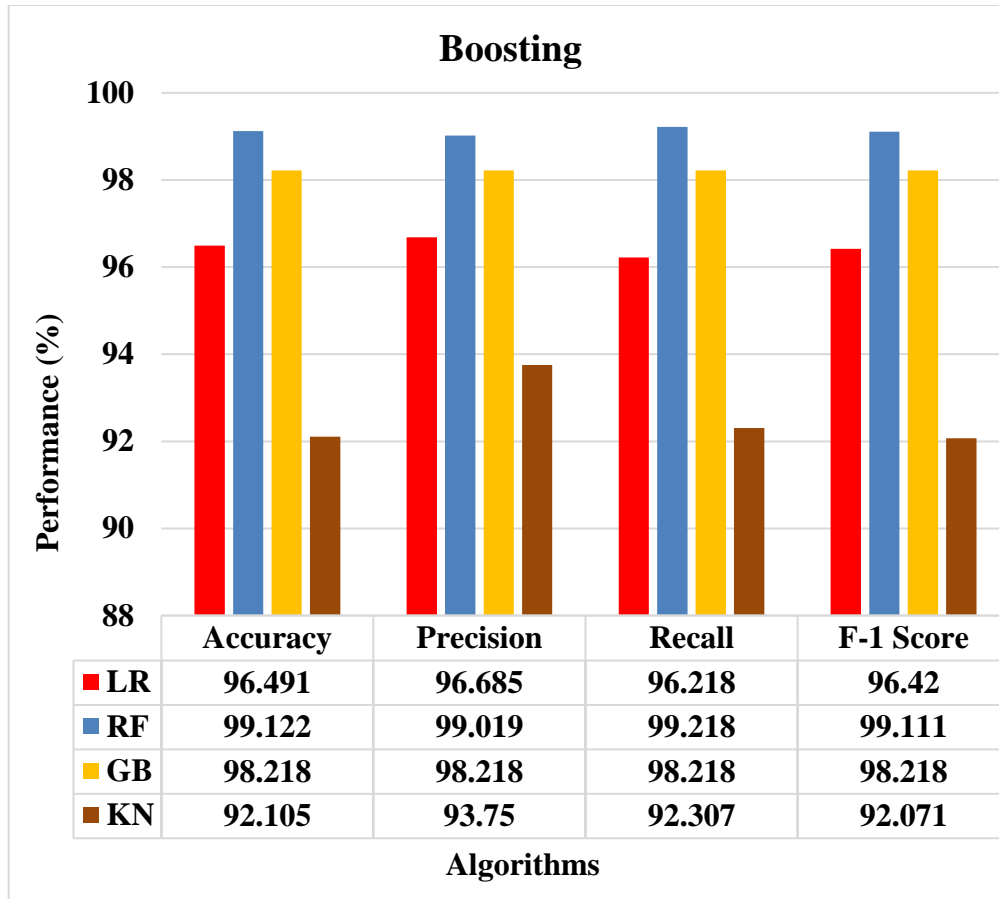


Fig 4.12: Experimental Results of Boosting

The final consideration should be the performance obtained using boosting algorithms. After applying boosting algorithms, the best accuracy had obtained from Random Forest (RF) at about 99.122%. The precision had 99.019%, the recall had 99.218% and the F-1 score had 99.111%. The accuracy had been improved. But Logistic Regression had 96.491% accuracy, 96.685% precision, 96.218% recall, and 96.42% F-1 score. The score of Logistic Regression (LR) had decreased. Another algorithm Gradient Boosting (GB) and K-Nearest had improved with 98.218%, and 98.456% accuracy respectively. The other performance evaluation measuring method precision had 98.218% and 98.455%, the recall had 98.218% and 98.456%, and the F-1 score had 98.218% and 98.457% respectively. The Random Forest (RF) achieved the best algorithm in our study which is shown in Fig 4.12.

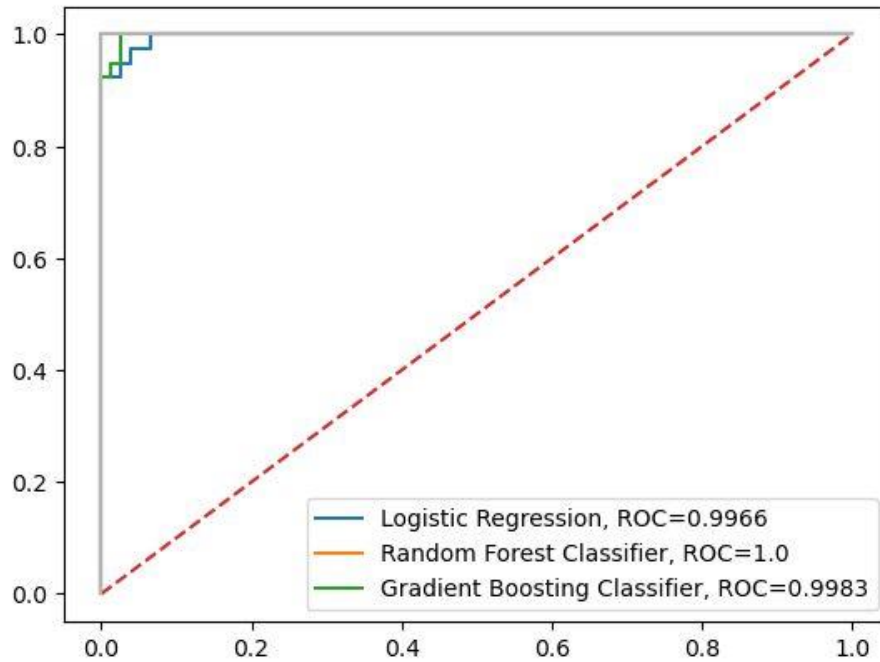


Fig 4.13: AUC-ROC Curve Analysis of Boosting

The AUC-ROC Curve depicted in Fig 4.13, also showed that Random Forest (RF) had acquired the best score of 99.99%. But Logistic Regression (LR) had acquired the 99.66% which is so much closer to the RF. Hence according to the above analysis as well as the detailed results with graphical representation, Random Forest Classifier resulted in the best algorithmic classifier.

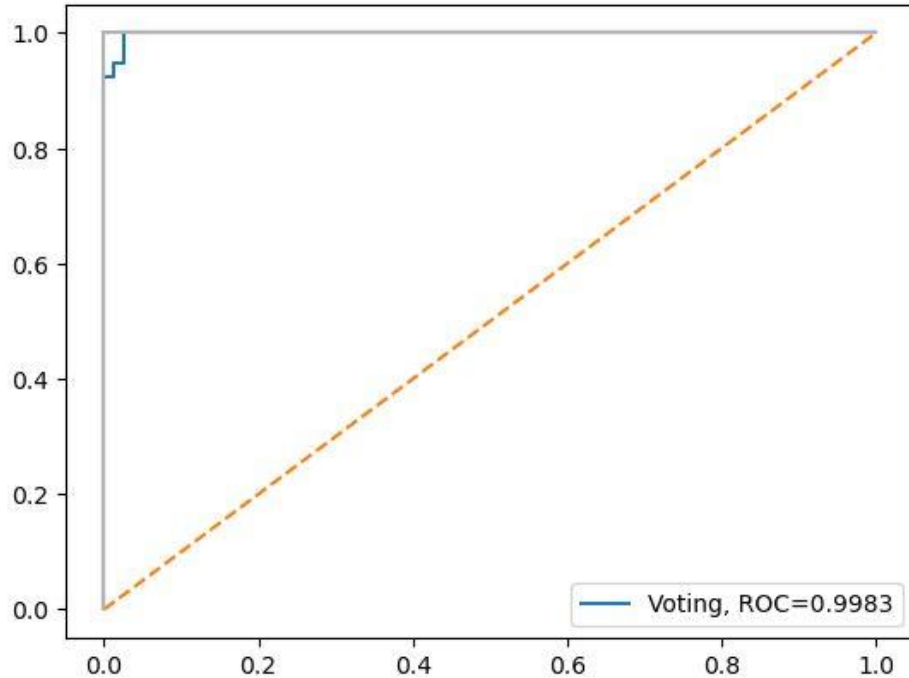


Fig 4.14: AUC-ROC Curve Analysis of Voting

We have applied a Voting algorithm for better calculation shown in Fig 4.14. We have used Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), and KNearest (KN) together with the ensemble method. The accuracy had 95.614%, the precision had 96%, the recall had 94.998% and the F-1 score had 95.427%. Which was not satisfactory but it touched the accuracy of our exciting models. We also showed the AUC Curve for the voting algorithm. The AUC score had achieved 99.83%.

We have evaluated the compilation time for our proposed method shown in Fig 4.15. We have measured particular compilation time for particular algorithms. The figure shows that Gradient Boosting (GB) took 492 Milliseconds, this is the highest compilation time among the algorithms. Then Logistic Regression Boosting (LRB) took 480 Milliseconds, this is the second highest compilation time among the algorithms. Then Logistic Regression Bagging (LRB) took 428 Milliseconds to compile. The Random Forest Bagging (RFB) took 248 Milliseconds to compile. The ensemble model (LRGK) took 168 Milliseconds to compile. The Logistic Regression (LR) took 91.5 Milliseconds to compile. The Boosting of Gradient Boosting (GBBO) took 61 Milliseconds to compile.

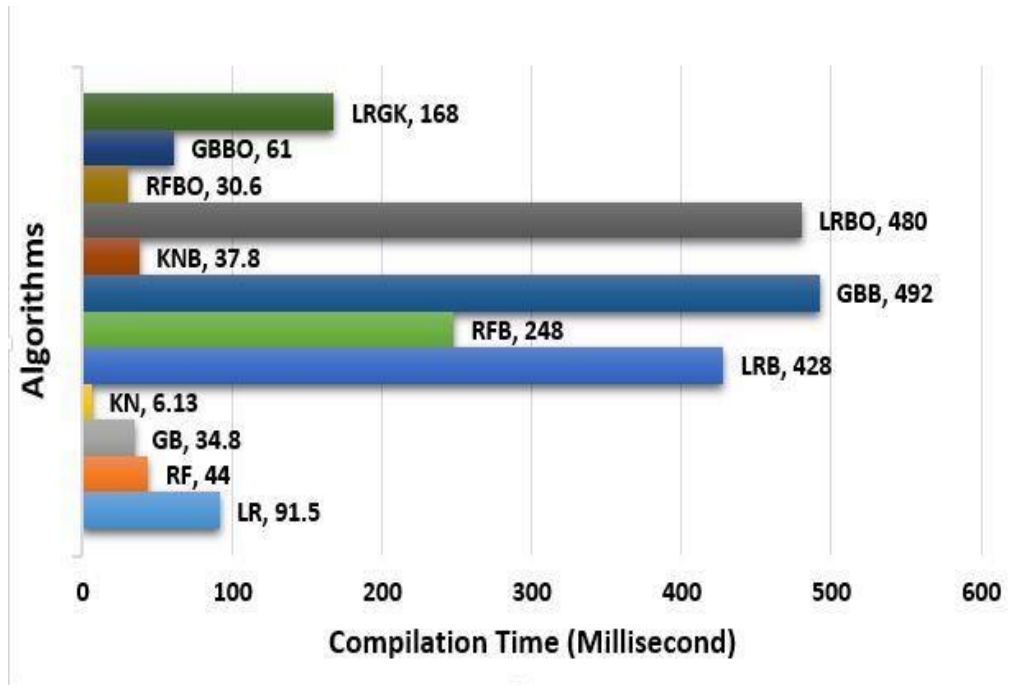


Fig 4.15: Compilation Time (Millisecond)

4.3 Discussion

In this stage we will clarify the judicial system of our proposed model. We have considered the accuracy, precision, recall and F-1 score.

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. We must strive to make our models as accurate as possible.

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

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.

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

4.3.3 Recall

It speaks about the percentage of positively anticipated observations from a model. High accuracy, though, might occasionally be deceptive. Normally Recall determines the proportion of expected positives to all positive labels.

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

4.3.4 F-1 Score

It speaks of the precision and recall harmonic means. Both the recall and precision ratios are relevant. If the harmonic mean is smaller, the model is probably not very good.

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

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

In our proposed model we have several benefits both socially and economically. Our model is based on a real-life dataset to explore and determine the basic points or parameters of a breast cancer patient. The work has social values like we can educate young women about the ratio of affected and precautions for breast cancer. We can suggest early treatment through diagnosis and regular check-ups. As they can be aware of breast cancer and easy to determine the possibilities to be affected or not. Our model is less time consuming and a few compilations time. So it is easy and accurate to predict the disease. In our model, we have analysed the dataset to identify the cause behind breast cancer with better diagnosis processes. We hope our proposed model will be accepted and will have real-life implementation socially.

5.2 Impact on Environment

Our proposed model is so beneficial in rural areas because of smooth diagnosis processes. We can reduce both time and complexity through the model of devices. We can assure that the environment will also be benefitted from our model as it has no side effects or complexity. The patients do not need to go to urban areas to check if they are the victim or not of breast cancer. The prediction model can easily justify the patient's diagnosis report and predict the possibilities. As it is not so much costly to determine breast cancer, patient will not be afraid of about it and no need to be worried of local treatments. Every stage of people can use it with its less complexity. Our proposed model can clarify the possible breast cancer affected patients or not. Our economic and social environment will be benefitted by our proposed model. If we can get the opportunity to implement our proposed model in real life, we can assure that, it will create milestone in recent medical science technologies.

5.3 Ethical Aspects

We have some ethical precautions like personal data or diagnosis report leakage or humour have to clarify the personal data exploration before the system runs. Our proposed model can be used for further research purposes and real-life implementation of the diagnosis and treatment of breast cancer. We have identified the topic as a problem not only for the limited area or region but also worldwide. Anyone the victim or aware woman can predict their breast cancer affect rate through the proposed model.

5.4 Sustainability Plan

We can assure that our proposed model can be accepted by worldwide research and breast cancer diagnosis technologies. We are confident our proposed model can be useful among the victim women who can easily predict their ratio of getting affected by breast cancer. If we get proper utilities and scope to implement, we can be motivated and we will be ready to implement in real life to help the rural areas. We hope our proposed model will be sustainable and beneficial.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

By using algorithms in our exciting paper, we can analyze the affected rate of our people. We can achieve an accurate prediction with our model. The diagnosis technology can enhance the prediction system. People can be benefited by knowing if they are going to affect or not. They can assume that they should be aware of breast cancer. If our model is received by the people, they can easily detect the stages of breast cancer disease. Diagnosis authorities also can be benefited by assuming our proposed model. We have used some different common algorithms which are low time consuming and easy to implement with high accuracy.

6.2 Conclusion

The present world is the modern world. Everything in the world is now technologically advanced and easy. Everyone in the world can familiarize themselves with the new technology. With the help of technology, we have proposed is so much easy and low time consuming. We have tried to reduce the complexity of breast cancer prediction among people. Our people can be benefited from our exciting models. We have to ensure the proposal is practical and we are promising to add many more features to our proposal in the future and ensure we will work on more popular things. We are expressing this hope.

6.3 Implication for Further Study

We are human beings, we have mortality. We are affecting several diseases in our daily life. Some of us have recovery essentials but most of us suffer from cancers. As we are living in a developing world, the treatment and diagnosis technologies are more dynamic and accurate. New technologies have shortened the time and complexity of breast cancer

disease identification. We have tried to do something new for our people. We hope our model will be accepted by the people. We have worked on some algorithms here and plan to add more in the future for better performance.

6.4 Limitations

We are human beings, we have mortality. We are affecting several diseases in our daily life. Some of us have recovery essentials but most of us suffer from cancers. As we are living in a developing world, the treatment and diagnosis technologies are more dynamic and accurate. If our proposed methodology is accepted by the people, we will work more deeply with this project. Our model is executed with a few data, as a result, the evaluation may differ from another research methodology.

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] V. Mary, K. Rani, and S. S. Dhenakaran, "Classification of ultrasound breast cancer tumor images using neural learning and predicting the tumor growth rate," *Multimedia Tools and Applications*, vol. 79, no. 23, pp. 16967–16985, 2020.
- [4] Y. Li, Y. Liu, M. Zhang, G. Zhang, Z. Wang, and J. Luo, "Radiomics with attribute bagging for breast tumor classification using multimodal ultrasound images," *Journal of Ultrasound in Medicine*, vol. 39, no. 2, pp. 361–371, 2020.
- [5] W. Gómez-Flores and J. Hernández-López, "Assessment of the invariance and discriminant power of morphological features under geometric transformations for breast tumor classification," *Computer Methods and Programs in Biomedicine*, vol. 185, article 105173, 2020.
- [6] Y. Liu, L. Ren, X. Cao, and Y. Tong, "Breast tumors recognition based on edge feature extraction using support vector machine," *Biomedical Signal Processing and Control*, vol. 58, no. 101825, pp. 1–8, 2020.
- [7] R. Irfan, A. A. Almazroi, H. T. Rauf, R. Damaševičius, E. A. Nasr, and A. E. Abdelgawad, "Dilated semantic segmentation for breast ultrasonic lesion detection using parallel feature fusion," *Diagnostics*, vol. 11, no. 7, p. 1212, 2021.
- [8] V. Lahoura, H. Singh, A. Aggarwal et al., "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," *Diagnostics*, vol. 11, no. 2, p. 241, 2021.
- [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" *ARNP 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. 5133-5142. 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] hou, Zhi-Hua. *Ensemble methods: foundations and algorithms*. CRC Press, 2012.
- [23] 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.
- [24] emmens, Aurélie, and Christophe Croux. "Bagging and boosting classification trees to predict churn." *Journal of Marketing Research* 43, no. 2 (2006): 276-286.
- [25] 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.
- [26] 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*.

Team_AtoZ

ORIGINALITY REPORT

27% SIMILARITY INDEX	22% INTERNET SOURCES	16% PUBLICATIONS	11% STUDENT PAPERS
--------------------------------	--------------------------------	----------------------------	------------------------------

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	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	7%
3	Submitted to Daffodil International University Student Paper	3%
4	downloads.hindawi.com Internet Source	3%
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%
