

**A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS
FOR BREAST CANCER**

BY

**ARNAB SAHA
ID: 201-51-3084**

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

Supervised By
Mr. Amir Shoel
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By
Mr. Rahmatul Kabir Rasel Sarker
Lecturer
Department of CSE
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
DECEMBER 2024**

APPROVAL

This Project titled “A Comparative Analysis of Machine Learning Algorithms for Breast Cancer”, submitted by Arnab Saha, ID:201-15-3084 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 22 January 2024.

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori (SRH)
Professor & Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

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

Internal Examiner

Mohammad Monirul Islam (MMI)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University


External Examiner

Dr. Md. Arshad Ali (DAA)
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology
University

DECLARATION

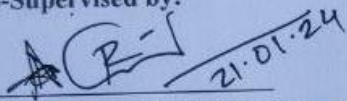
I hereby declare that, this project has been done by us under the supervision of **Mr. Amir Shoel, Lecturer, Department of Computer Science and Engineering** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

Supervised by:



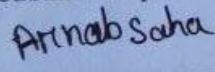
Mr. Amir Shoel
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Rahmatul Kabir Rasel Sarker
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Arnab Saha
ID: 201-15-3084
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

I really grateful and wish our profound our indebtedness to of **Mr. Amir Shoel**, Lecturer, 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 ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express our heartiest gratitude to Mr. Amir Shoel, **Dr. Sheak Rashed Haider Noori** Professor and Head, Department of Computer Science and Engineering, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

With the rapid expansion of medical research in recent years, early diagnosis is now more critical than ever. A growing global population increases the risk of death from breast cancer, making it the second most severe cancer reported. For this reason, automated diagnostic systems are becoming a valuable adjunct for clinicians. This system helps in accurate diagnosis and makes it reliable, effective, and fast. By doing so, it plays a vital role in reducing the mortality associated with breast cancer. Integrating such technological advances humanizes the approach to healthcare, ensuring timely interventions that can significantly impact patient outcomes and well-being. As we navigate the challenges of the evolving clinical landscape, emphasizing early detection through the Action Plan underscores our commitment to the growing challenges posed by breast cancer and improving cancer healthcare delivery. Although these characteristics naturally vary from person to person, thorough testing and a wealth of clinical data combine to determine normal levels for a healthy individual. Age at survey Attempts were made to evaluate stratification strategies for quantifying the risk of individuals according to gender and specific characteristics associated with breast cancer risk. When employing machine learning with intense object pressure, various classification methods are used in the analysis, such as Support Vector Machine (SVM), Decision Tree Algorithm (DT), K-Nearest Neighbour (KNN), Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), etc. Forecasting cancer rates accurately shows progression and is notable for considering the wide range of factors that influence breast cancer risk—study methods, such as confusion matrix coefficient-based selection of features to improve model predictions further. A thorough analysis of the data, comparisons, and evaluations are made, and key performance indicators, including accuracy, precision, F-1 score, recall, sensitivity, and specificity, are reviewed, providing information.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 INTRODUCTION	1-2
1.2 MOTIVATION	2
1.3 RATIONALE OF THE STUDY	2-3
1.4 EXPECTED OUTPUT	3-4
1.5 REPORT LAYOUT	4
CHAPTER 2: BACKGROUND	5-8
2.1 INTRODUCTION	5
2.2 RELATED WORK	5-7
2.3 SCOPE OF THE PROBLEM	7-8
2.4 CHALLENGES	8
CHAPTER 3: RESEARCH METHODOLOGY	9-21
3.1 WORKING MODEL	9
3.2 DATASET COLLECTION	9-10
3.3 DESCRIPTION OF BREAST CANCER	11
3.4 DATA PRE-PROCESSING	12
3.5 LABEL ENCODING	12-13
©Daffodil International University	vi

3.6 SMOTE	13-14
3.7 CORRELATION	14-15
3.8 FETURES SELECTIONS	15-17
3.9 TRAINING & CLASSIFICATIONS	17
3.10 PROPOSED MODEL	17-21
CHAPTER 4: Experimental Result & Discussion	22-34
4.1 RESULT & DISCUSSION	22-23
4.2 RESULT ANALYSIS	23-30
4.3 COMPRATIVE ANALYSIS	30-33
4.4 DISCUSSION	34
CHAPTER 5: Impact on Society, Environment & Sustainability	35-36
5.1 IMPACT ON SOCIETY	35
5.2 IMPACT ON ENVIRONMENT	35-36
5.3 SUSTAINABILITY PLAN	36
CHAPTER 6: Summary, Conclusion, Recommendation & Implementation for the Future Study	37-38
6.1 SUMMERY OF THE STUDY	37
6.2 CONCLUSION	37
6.3 IMPEMANTION OF THE FUTURE STUDY	38
REFERENCES	39-40
PLAGIARISM REPORT	41

LIST OF FIGURES

FIGURES	PAGE NO
Fig 3.1.1: Working Process of the Breast Cancer Prediction	9
Fig 3.2.1: Diagnosis (M = malignant, B = benign)	10
Fig3.5.1 Diagnosis (1 = malignant, 0 = benign)	13
Fig 3.6.1 Balanced Dataset for Diagnosis Class	14
Fig 3.7.1: Correlation matrix for the breast cancer dataset.	15
Fig 3.10.4 SVM generated hyper-planes	20
Fig 4.2.1: Different Machine Learning Algorithms	24
Fig 4.2.2: Metrics of Different Machine Learning Algorithms Confusion Matrix	25
Fig 4.2.3: ROC Curve	26
Fig 4.2.4: AUC Score	27
Fig 4.2.5: Metrics of Different Machine Learning Algorithms Jaccard Score	28
Fig 4.2.6: Metrics of Different Machine Learning Algorithms Mean Cross-Validation Score	29
Fig 4.2.7: Metrics of Different Machine Learning Algorithms Mean Classification Error Score	30
Fig 4.3.1: Performance using all Algorithm	31
Fig 4.3.2: Algorithms (Boasting) Performance all Features	31
Fig 4.3.3: Performance all Algorithm using PCC-FS	33
Fig 4.3.4: Algorithms (Boasting) Performance all Features using PCC-FS	33

LIST OF TABLES

TABLES	PAGE NO
Table 3.2.1: Ten real-valued features are computed for each cell nucleus	10
Table4.2.1: Applied Different Machine Learning Algorithms	23
Table4.3.1: Applied Different Machine Learning Algorithms with Boasting by AdaBoast	31
Table 4.3.2: Comparison of Accuracy with Earlier Studies	32
Table4.3.3: Applied Different Machine Learning Algorithms with Boasting by AdaBoast using PCC-FS	33

CHAPTER 1

Introduction

1.1 Introduction

According to the American Cancer Society (ACS), breast cancer makes up 25% of cancer diagnoses in women globally [1]. Breast cancer tumours can be categorized into two main groups: (i) benign (noncancerous), whereby the immune system segregates benign tumours from normal cells; (ii) malignant (cancerous), which initiates abnormal cell growth and invades nearby tissue. Identifying normal, benign, and malignant tissues is crucial for treatment [12, 13]. A type of cancer called breast cancer starts from breast cells. Both men and women can get it, but women are more likely to get it. The disease usually begins in the milk-producing lobules or tubes that carry milk to the nipples. Breast lumps or changes in the appearance of the size and shape of the breast skin can be a sign of breast cancer. Doctors use screening methods like mammography to detect breast cancer in its early stages so that it can be seen. Age, family background, or hormonal changes can also cause specific DNA mutations that can put you at risk. Treatment options for this disease include surgery, radiation therapy, hormone therapy, and some appropriate medications. Awareness and regular checkups are two essential things that can help people with cancer. Breast carcinoma can be diagnosed by detecting tumours. Malignant and benign are two different kinds of tumours. Doctors require an active determination technique to recognize these tumours. But mostly, it is tough to acknowledge tumours, even by specialists [14]. As per the World Health Organization (WHO), the most influential problem in the field of medical research is breast cancer diagnosis [8]. Many researchers are working on machine learning to discover the severity of breast cancer, i.e., whether tumours are cancerous or noncancerous. Now, we have to work on the two most essential questions. Among the two critical questions are: what is the role of the machine, and how does the machine learn by combining medical data to predict the severity of the disease? Machine learning is one way to make decisions based on data with minimal human intervention. Machine learning is being used more and more in medicine because it has applications for predicting the future and making diagnoses, particularly cancer prognoses. It is also being used a lot in biological research at present. The data analysis journey commenced with meticulous preprocessing, entailing data scrutiny for missing and null values.

Subsequently, linear regression coefficients were judiciously deployed to discern and eliminate redundant features. Label encoding techniques were adeptly employed to combat the challenge of an imbalanced class distribution, complemented by feature scaling to address outliers. These diligent preparatory measures paved the way for applying various classifiers, including logistic regression, random forest, decision trees, naive Bayes, support vector machines, and K-nearest neighbors. In this regard, we will be comparing two things: one is the standard algorithm, and the other is to find out the accuracy through the Pearson correlation coefficient. Besides, the boosting model will be used in both algorithms. And we will build the boosting model with the help of the Ada boost algorithm.

1.2 Motivation

Better identification and even diagnostic accuracy, precision, and efficiency are the driving forces behind using machine learning in breast cancer studies. Recently published studies have highlighted the possibilities of machine learning approaches for improving the detection rate of breast cancer. Regarding swift recognition & prompt action, such increased accuracy has become crucial. Creating sophisticated prediction models for breast cancer using various sources of information is another source of inspiration. Synthesis of several characteristics is made possible by machine learning, which produces tools for forecasting that are more durable and trustworthy. Previous studies have shown limits in the future outlook of breast cancer, as evidenced by current articles. To overcome these restrictions, machine learning acts as a tactical instrument that offers ways to improve the durability and reliability of diagnostic methodologies. The ultimate objective is to use machine learning algorithms to improve the detection of breast cancer. This improvement aims to provide precise and effective findings while guaranteeing a high likelihood of finding possible instances throughout periodic checks. To put things briefly, the specific aim is to use machine learning to transform the diagnosis, prognosis, and early detection of breast cancer, hence improving patient outcomes: results and enhanced medical procedures.

1.3 Rationale of the Study

The explanation for utilizing machine learning in the detection of breast cancer is rooted in its capacity to transform precision, effectiveness, and customized treatment. Recent

research has shown that machine learning algorithms are remarkably accurate for identifying breast cancer. The capacity to analyze intricate data sets facilitates quick recognition of minute trends, which in turn permits prompt action and enhanced patient outcomes. To maximize the likelihood of breast cancer, attention is paid to applying machine learning frameworks such as random forests and ADA boosting. By improving medical diagnostic precision, these mathematical models give doctors the trustworthy information they need to make wise judgments. This study aims to create machine learning-based algorithms that can accurately identify breast cancer in digital photos. This ability is crucial for streamlining the diagnostic procedure and reducing the time needed to receive findings. Systematic examinations have shown that machine learning approaches help provide a thorough overview of research on the outlook of breast cancer. This concept methodical technique aids in assessing the efficacy and dependability of different machine learning techniques. The paper highlights machine learning applications for precise cancer type categorization, outcomes for patients forecasting, and recognizing viable options for treatment. This individualized strategy will enhance patients' responsiveness and medical care approaches. To put it briefly, the idea is to use machine learning to improve diagnostic models, diagnose breast cancer with greater accuracy, expedite the procedure, and eventually provide patients with safer and more individualized therapy.

1.4 Expected Output

The supposed outcome of machine learning-based breast cancer prognosis will be a particular and precise model that helps with many facets of breast cancer care. By identifying minute patterns and anomalies within healthcare information, machine learning algorithms have enabled early identification of breast cancer. Beneficial patient results & the potential of having a successful course of therapy are significantly increased by a quick diagnosis. The creation of prognosis designs, which predict results and the likelihood of survival for individuals with breast cancer, is facilitated by sophisticated machine learning methods. Using this knowledge, doctors may create treatments that consider the illness's expected progression. The application of machine learning, especially its deep-learning methods, aids in the creation of targeted treatment strategies for those suffering from breast cancer. Optimizing treatment results entails studying omics data to adapt measures of treatment. The ultimate aim is to increase the

rates of breast cancer survival, which may be attributed to machine learning-enabled prompt diagnosis, precise outlook, and individualized therapy management.

1.5 Report Layout

This chapter included Machine Learning Algorithms for Breast Cancer. The justification of the research study argument and thesis findings reporting format will come later.

In Chapter 2, we will discuss the background of our study topic.

In Chapter 3, we will talk about the research methodologies

In Chapter 4, we will talk about the experimental result and discuss comparative analysis.

In Chapter 5, we will discuss the impact on Society, environment & sustainability for our study.

In Chapter 6, we will discuss the summary of our study and the conclusion, recommendations, and implementation for future studies.

CHAPTER 2

Background

2.1 Introduction

Using supervised learning with datasets, cancer, several research have examined the use of machine learning techniques for breast cancer diagnosis, offering insightful information and valuable strategies that can be used to identify, train, evaluate, and adjust the system. There is evidence that the data has previously been honored with excellence; we are confident that the larger the data set, the better the machine learning outcomes are once our classifier has been trained using the input data.

2.2 Related Work

In their 2020 study, Fatima et al. thoroughly examine and compare various machine-learning approaches to breast cancer prediction. The authors discuss data mining, deep learning, and machine learning to determine the efficacy of these breast cancer prediction approaches. The study provides valuable information on the veracity of these methods' performance predictions and examines them thoroughly. If researchers and doctors want to know what factors affect breast cancer prognosis, this study would be an excellent place to start.[1] Islam et al. conducted extensive comparative analyses utilizing machine learning approaches to investigate breast cancer prognosis in these articles. This work tackles the urgent need for accurate models to forecast breast cancer occurrence. The authors contribute to the growing body of literature on machine learning by methodically comparing and analyzing different approaches. Findings highlight the merits and shortcomings of the approaches by highlighting their nuanced distinctions. This comprehensive review is an excellent tool for researchers looking to learn more about applying machine learning for breast cancer prediction. Everyone interested in keeping up with the latest developments in breast cancer research can benefit from Islam and his team's work, especially those seeking better prediction models.[2] In their article, Yasin et al. describe computer-assisted breast cancer screening with an emphasis on imaging methods. The authors examine machine learning technologies to increase diagnostic abilities and evaluate their efficacy. The study gives a comprehensive understanding of the pros and cons of the methods used

by Yasin and colleagues to highlight the importance of these methods in advancing breast cancer treatments; accuracy is valued more than ineffectiveness. It does this by combining the findings of multiple methodologies. If researchers and doctors are interested in using machine learning for breast cancer diagnosis, their study—which includes a thorough literature review—is required reading. Excellent methodology underpins this all-encompassing study, which sheds light on the present state of affairs and the innovative approaches to this crucial task.[3]. In this work, a notable study was carried out by Bayrak et al., which compared several machine-learning approaches for breast cancer diagnosis. They have presented their findings at the Scientific Assembly of Computer Science, Biomedical Engineering, and Electronics, and their research is a systematic investigation of breast cancer detection methods. The research sheds light on the pros and cons of machine learning and gives valuable information. This study provides a wealth of information for medical professionals and academics curious about the accuracy of breast cancer detection using machine learning algorithms.[4] At the 2018 CTMS conference, Sharma et al. published a paper that delves into the topic of breast cancer diagnosis through the use of machine learning algorithms. The writers used the Wisconsin breast cancer search database to examine the efficacy of several machine learning algorithms. Three popular MLs for breast cancer detection were compared in this study. The profession is constantly striving to improve breast cancer detection procedures, and these contributions provide valuable knowledge to that end.[5] Noor et al. Wisconsin investigated the use of machine learning for breast cancer classification in a landmark study. Their study, published in the International Journal of Engineering and Technology, provides valuable insight into the performance of machine learning in breast cancer diagnosis. The study's substantial field information could help academics and physicians better understand machine learning applications in breast cancer diagnosis.[6] The efficacy of ML algorithms in detecting and diagnosing breast cancer is investigated in the article by Bajajeh et al. Three popular ML methods are examined in the study: random forest, support vector machine (SVM), and an extra method that remains mysterious. The authors analyze the algorithms' performance through extensive research, paying particular attention to accuracy. Model accuracy and hyperparameter tuning are both examined in a comparative study. This study stands out because it highlights the importance of high-resolution images in improving the accuracy of breast cancer diagnoses. The results

provide valuable information on the pros and cons of machine learning approaches and may be used to make educated decisions in clinical practice. Clinical work.[7] In their research papers, Gupta et al. compare supervised machine learning methods for studying breast cancer diagnosis. This study examines several algorithms and their potential applications in improving accuracy; it was presented at ICCMC 2018. The authors shed fresh light on the efficacy of breast cancer detection methods by analyzing the current state of supervised machine learning. Essentially exploratory, this work sheds light on the growing importance of machine learning in healthcare research.[8] In their 2019 study, Osmanovich et al. cover every angle of using machine learning to categorize breast cancer. Methods for distinguishing between benign and malignant cases are discussed in the study presented at the World Congress of Medical Biotechnology and Biotechnology. Medical research is advanced thanks to the authors' illuminating commentary on the usefulness of machine learning for precise breast cancer classification. Professionals looking into state-of-the-art cancer diagnostic tools will find this article an essential reference.[9] A thorough evaluation of the uses of machine learning is provided by Chaurasia and colleagues in their examination of breast cancer prediction. Their research on breast cancer diagnostic outcome prediction was published in SN Computer Science. This study includes an ever-growing body of literature on machine learning in healthcare decision-making. The writers add to our understanding of how crucial high-tech equipment may be for detecting and diagnosing breast cancer by analyzing the effectiveness of different approaches.[10]

2.3 Scope of the Problem

Conceivable benefits of using machine learning for breast cancer prediction include enhanced precision, early identification, and tailored care. It analyzes complicated information, such as records of patients and imaging data, using sophisticated algorithms. Mention certain facets of the issue. Putting algorithms for machine learning into practice to find minute trends in medical data to diagnose breast cancers early and accurately. To identify the best breast cancer prediction technique, compare machine learning models, including random forests, decision trees, and logistic regression. Resolving the issue affirmation, outlining the parameters, and providing evidence for machine learning studies relating to breast cancer diagnosis. Creating models that can identify breast cancer that already exists and forecast how it will spread, therefore

assisting with all-encompassing patient treatment. Recommending specific machine learning techniques, such as decision trees, to create an accurate and efficient predictive algorithm for breast cancer. The main objective is to use machine learning techniques in breast cancer studies to maximize the pinpointing of diagnosis, support timely treatment, and eventually enhance the quality of life for patients.

2.4 Challenges

Because breast cancer is a complicated illness with many subgroups and varied features, it might be difficult for machine learning algorithms to recognize and categorize different trends. The durability and generalization of algorithms that use machine learning can be impacted by skewed forecasts caused by the scarcity of large, well-balanced datasets for algorithm training. Achieving the respect of medical experts about predictive machine learning models related to breast cancer depends on their accessibility. Ensuring algorithms are intelligible and offer therapeutically significant discoveries can be challenging. It is difficult to integrate machine learning techniques within current healthcare procedures in a seamless manner. Acceptance of an intuitive interface, analysis in continuous time, and interoperability with medical systems. Progressive machine learning methods, intense learning ones, can have high computing requirements. Providing practical model training and forecasting while controlling computing expenses is an ongoing task. Another significant difficulty is verifying the machine learning algorithms' medical efficacy and dependability for breast cancer diagnosis in various patient groups. Thorough testing is required before extensive adoption in medical environment

CHAPTER 3

Research Methodology

3.1 Working Model

The proposed system is shown as a block diagram below in figure 3.1.1.

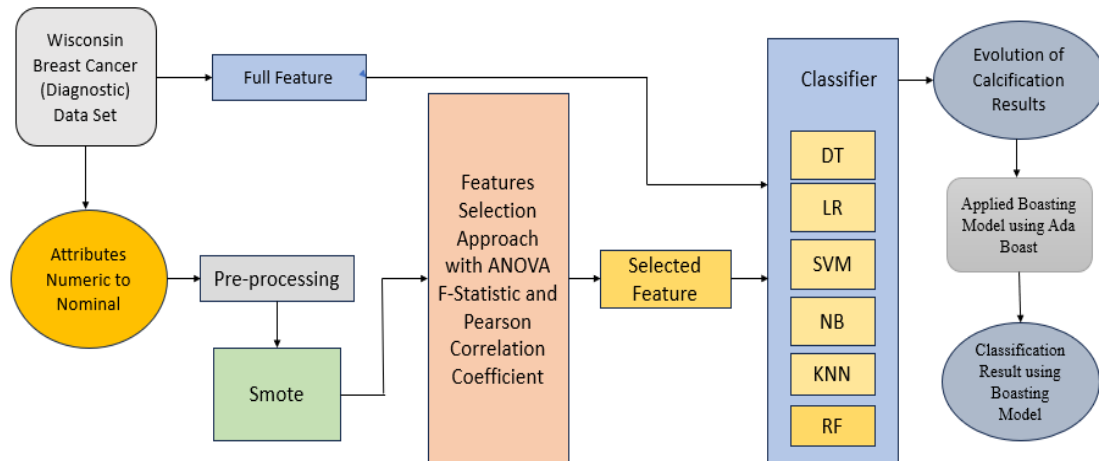


Fig 3.1.1: Working Process of the Breast Cancer Prediction

The following method's machine learning models suggest completion techniques for collecting information, including benign and malignant data, which are used to prepare data. Decision trees, logistic regression, Naive Bayes, KNN, SVM, and random forests are the models.

3.2 Dataset Collection

For this study, the dataset used the "Wisconsin Breast Cancer (Diagnostic) Data Set" with 569 instances and 32 attributes. This data set was created by Dr. William H. Wolberg of the University of Wisconsin to diagnose breast cancer, i.e., (M = malignant, B = benign) [10]. The dataset is located at archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. The breast cancer medical data collection comprises 569 instances, with 357 classified as soft and 212 as malignant. The data was reported on November 1, 1995, and includes the patient identity number and diagnosis (malignant or benign) for each scenario. Attributes of Dataset [18]

(1) ID number

(2) Diagnosis M=malignant, B=benign

(3–32) Ten real-valued features are computed for each cell nucleus:

(a) radius (mean of distances from center to points on the perimeter)
(b) texture (standard deviation of gray-scale values)
(c) perimeter
(d) area
(e) smoothness (local variation in radius lengths)
(f) compactness ($\text{perimeter}^2/\text{area} - 1.0$)
(g) concavity (severity of concave portions of the contour)
(h) concave points (number of concave portions of the contour)
(i) symmetry
(j) fractal dimension (“coastline approximation” – 1)

Table 3.2.1: Ten real-valued features are computed for each cell nucleus

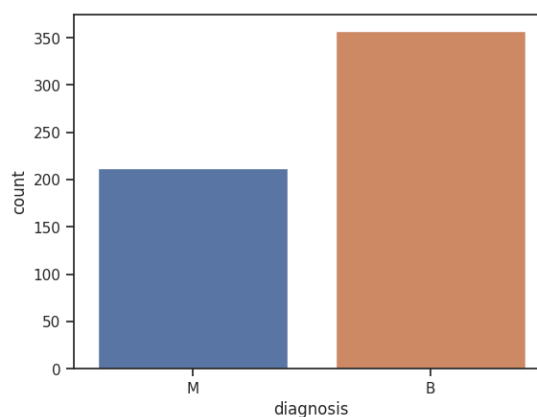


Fig 3.2.1: Diagnosis (M = malignant, B = benign)

3.3 Description of Breast Cancer

One kind of cancer which starts in the breast cells is called breast cancer. It has been defined as aberrant cells growing out of control to develop tumors. It is additionally the second-most prevalent cancer in women discovered globally, behind cancer of the skin. The appearance of a lump, modifications to the breast's form, skin dimpling, and nipple fluid flow are possible symptoms. For those with breast cancer, early identification and advancements in treatment techniques, including radiation, chemotherapy, surgery, and specialized treatment, have greatly enhanced prognosis. To Regular exams and public education campaigns are crucial in diagnosing and treating this frequent type of cancer.

3.3.1 Malignant

The previously unchecked proliferation of aberrant cells in a woman's breast tissue is the hallmark of breast cancer, a malignant disease. This is among the most frequent cancers in women identified globally, and it may spread to develop malignancies that infect surrounding tissue. The development of a lump, alterations in the size of the breasts, skin dimpling, and breast discharge are examples of early signs. Malignant breast cancer cells pose a serious danger to one's welfare since they can grow and spread to various other regions of the human organism if left treatable. For those with malignant breast cancer, early diagnosis via monitoring and advancements in treatment modalities, including surgery, chemotherapy, and specific treatment, are crucial to treating the disease and enhancing prognoses.

3.3.2 Benign

Many different kinds of benign tumors, cysts, and fibroadenomas may develop in the breasts. These abnormalities are not malignant. These disorders do not provide the same hazard of metastasis as malignant breast cancer. Benign breast diseases may manifest as lumps, changes in breast tissue, or breast discharge; often, mammography or clinical exam is used for diagnosis. Although these illnesses are usually harmless, they need medical attention for accurate assessment and treatment. Differentiating benign problems from malignant ones and ensuring optimum breast health, early identification, and adequate medical attention are of the utmost importance.

3.4 Data Pre-processing

One of the most critical aspects of classification that is necessary to improve prediction accuracy is Data preprocessing. It is the quality of data on which a prediction algorithm's performance is heavily dependent. Predictability is lowered if a dataset contains too many missing values, outliers, and irrelevant features. The primary aim of our research paper is to enhance the precision and efficacy of heart patient classification through improved prediction methods. Here, data preprocessing includes encoding categorical feature values, handling outliers, removing irrelevant features, and oversampling.

3.4.1 Feature Scaling

The vital process of feature scaling is instrumental in normalizing feature values to a predefined range, constituting a crucial and indispensable step in constructing a robust machine-learning model. Feature scaling reduces training time and helps achieve faster convergence for many machine learning and neural network algorithms. Mean and Standard deviation-based scaling may suffer if a dataset contains too many outliers. In our case, we have used the 'Z-score outlier detection technique' to detect the outliers and take care of those outliers using 'Robust Scaling.' 'Robust Scaling' uses the 'Interquartile Range' to handle the outliers and scale the data.

3.4.2 Data Cleaning

Data cleansing, a fundamental data preparation process, involves meticulously removing or modifying erroneous, incomplete, irrelevant, duplicated, or improperly formatted data. The utmost importance of data quality in training machine learning models necessitates rigorous data cleaning to ensure optimal performance and accuracy in the subsequent analysis.

3.5 Label Encoding

Label coding is an essential step in the preparation of machine learning for finding breast cancer. In this case, it means using math to turn category factors, like the type of tumor or the patient's traits, into a machine learning system. Offering a unique mathematics name to each class in this way makes understanding the method easier. There could be several valuable implementations and pieces of code for label encoding.

A one-hot encoding method that doesn't use numbers but instead uses transfer learning was also suggested as an alternative way to classify breast cancer. The name is shown with vectors. Label encoding makes it easier for machine learning systems to match category data, which helps them better organize breast cancer. The target value of my data set was diagnosis, so there were two groups: (i) benign and (ii) malignant. I changed these two values to 0 and 1. In the diagnostic data set for breast cancer, the attribute "diagnosis" is replaced by B's 0 and M's 1. In the data set, when the units of measurement are different, we need to standardize the data [10].

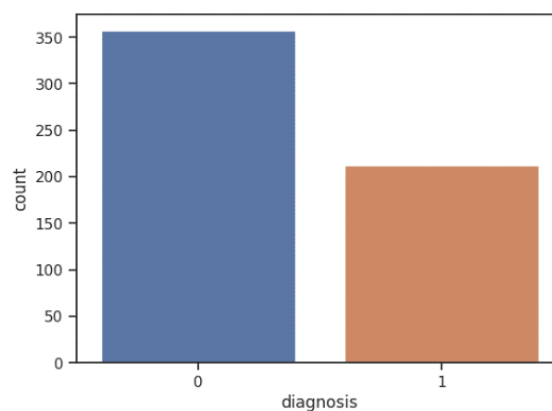


Fig3.5.1 Diagnosis (1 = malignant, 0 = benign)

3.6 Smote

When dealing with unbalanced datasets, such as those used for breast cancer classification, a machine learning technique called SMOTE (Synthetic Minority Over-sampling Technique) might be helpful. To improve the model's minority class recognition, it creates synthetic members of the minority class to even out the distribution of classes. By using SMOTE to datasets like mammography data, breast cancer detection models may be trained to be more sensitive to cancer-related microcalcifications. Breast cancer prediction models have improved their classification ability using oversampling, under-sampling, and SMOTE techniques. According to the research, when combined with other methods, such as under-sampling, SMOTE improves the overall performance of breast cancer prediction models.

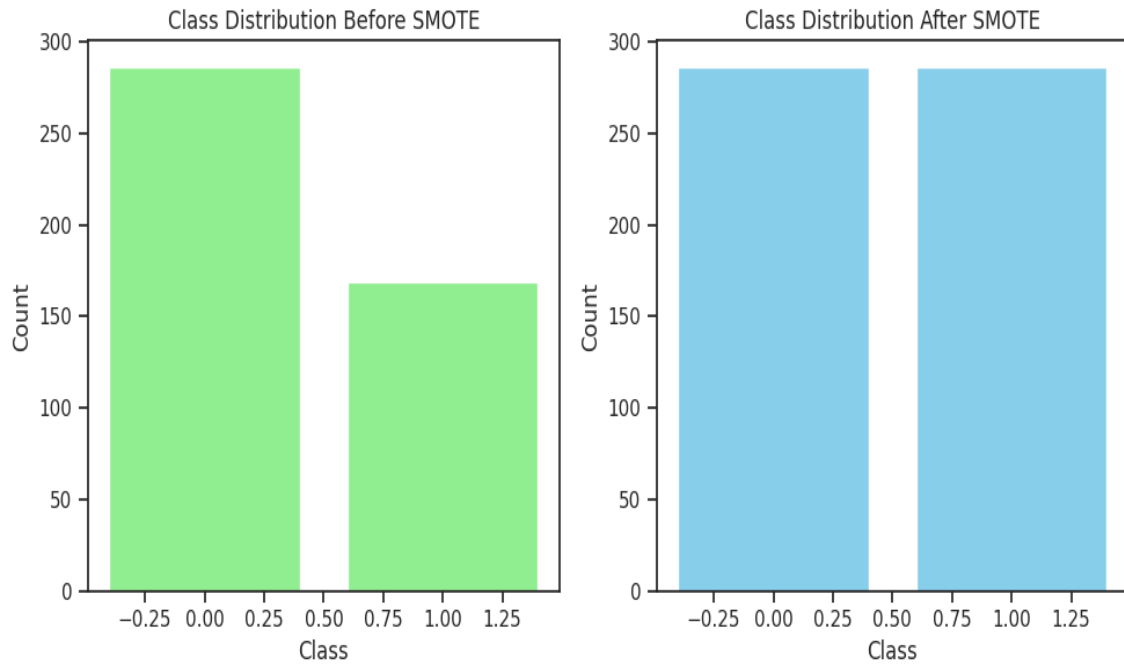


Fig 3.6.1 Balanced Dataset for Diagnosis Class

3.7 Correlation

Machine learning correlation measures the statistical link between breast cancer dataset characteristics, showing how changes in one variable affect changes in another. Correlation analysis helps researchers comprehend dataset structure by revealing patterns and connections between characteristics. Researchers may visually identify favorably, negative, and uncorrelated information using the correlation matrix. Machine learning algorithms use this knowledge to choose important characteristics or comprehend how factors affect breast cancer prediction models. Using correlation analysis and machine learning algorithms, breast cancer forecasts are more interpretable and accurate.

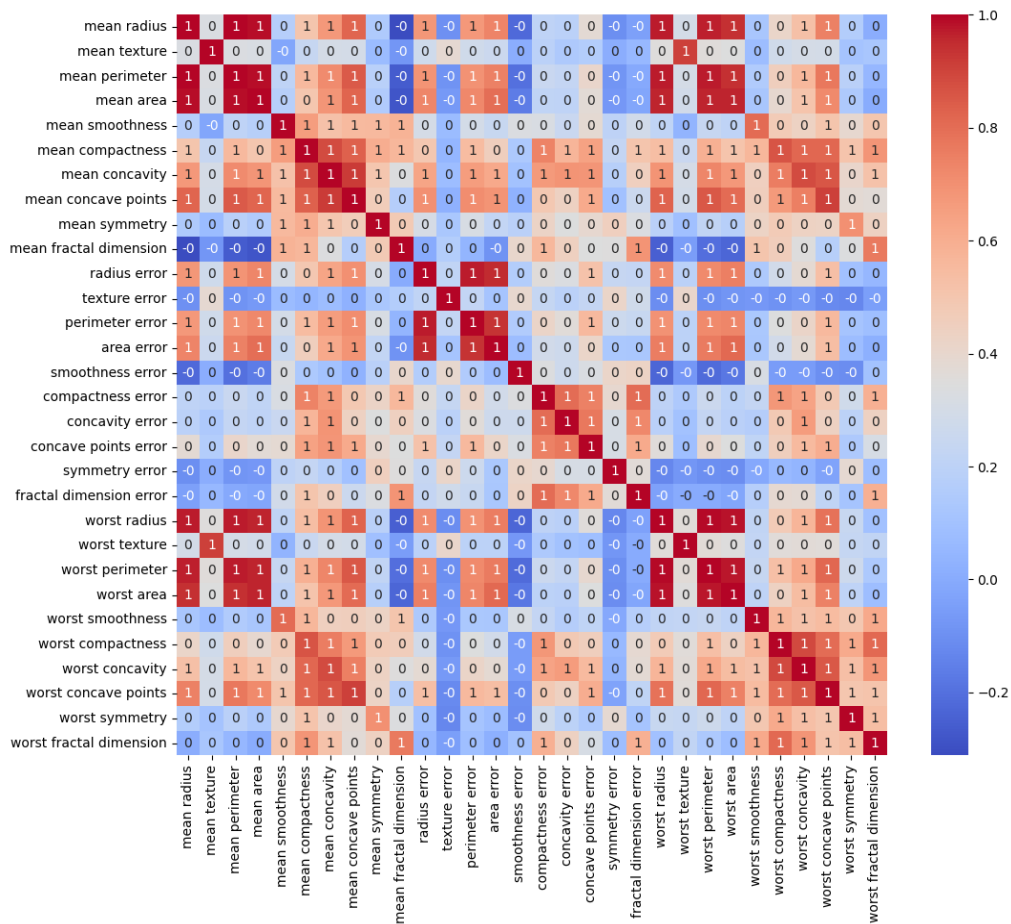


Fig 3.7.1: Correlation matrix for the breast cancer dataset.

3.8 Features Selection

Finding those most essential characteristics that help with precise forecasting is the goal of feature selection in machine learning applied to breast cancer. Model achievement, interpretability, and computing cost may be improved by decreasing noise and concentrating on essential variables such as tumor size, shape, and morphology of cells. The characteristics demonstrating the most vital link with the result may be ranked and chosen using methods such as Recursive Feature Elimination (RFE) and SelectKBest. This method increases the models' prediction ability and helps doctors recognize what aspects are essential for prognosis and how to treat patients.

3.8.1 ANOVA F-statistic

The ANOVA F-statistic is a type of F-test used in parametric statistical hypotheses testing. It helps determine if there is a significant difference between numerical input characteristics. It assesses the difference in means across various attributes rather than

©Daffodil International University 15

within each attribute. Regarding each trait, ANOVA calculates an F-statistic that shows the variability ratio between groups to within-group variance. Features are ordered according to their F-statistic value. Higher F-statistic values indicate greater significance in the differences between trait means. The ANOVA F-statistic helps identify features that show significant differences in mean values between benign and malignant cases. Understanding the various factors that contribute to breast cancer is crucial. The model prioritizes a subset of the most informative features by choosing features with high F-statistic values. This improves the accuracy of predictions and provides valuable insights into the critical variables related to breast cancer by bringing them to the forefront. Feature selection is crucial in mitigating the risk of overfitting, mainly when dealing with smaller datasets. Choosing the most appropriate features helps simplify the model, enhancing its ability to make accurate predictions on new data. Choosing features based on their F-statistic relevance ensures that computational resources are directed toward the most impactful variables. This capability is convenient when there are constraints on computational power or data volume. Certain features have been chosen to create a brief list of factors that impact the outcome of breast cancer. Understanding interpretability is essential for medical professionals and researchers who want to grasp the vital variables and possible biomarkers linked to disease. The F-statistic for a feature i is calculated as:

$$F_i = \frac{\text{Variance Between Classes (group } i)}{\text{Mean Squared Within Classes (error)}} \quad (1)$$

3.8.2 Pearson's correlation coefficient

One way to assess the degree of linear relationship between two variables, X and Y, is using Pearson's correlation coefficient (r), which may take on values between one and one and one negative, with one being the most strongly correlated variable and zero the least. R may be used in breast cancer machine learning to evaluate the direction and strength of the link between parameters such as tumor size (X) and malignancy probability (Y). Here is the formula:

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}} \quad (2)$$

Using this coefficient, we may determine which clinical characteristics are most strongly correlated with cancer outcomes, narrowing down the feature space to only include the most useful variables for predictive models.

3.9 Training & Classification

To find out the extent to which the algorithm worked, we used the subsequent phases in our study's learning and categorization process: We discuss how objective data is trained and evolved, including data categorization techniques that divide data into several categories. The two primary components of my data set upon diagnosis are malignant and benign. I use a multi-class categorization technique for systems. Nevertheless, I compared the sixth classification model against Decision Tree Logistic Regression, Naïve Bayes, SVM, KNN, and Random Forest for improved outcomes. The model that is most helpful in diagnosing breast cancer is random forest. Because it contains more temperature characteristics than previous models, it is pretty effective and capable of indicating breast cancer with the maximum degree of accuracy. It's a more challenging option since it can be learned and used rapidly. Applications in the real world where accuracy and resilience are crucial. We look in those metrics, including accuracy, precise memory, and the F1 score, to assess the effectiveness of our approach.

3.10 Proposed Model

Deep learning and collaborative methodologies are two cutting-edge strategies in the suggested machine-learning model for breast cancer prediction. This model has a high capacity to identify breast cancer by effectively analyzing digital mammography with a variety of problems. Incorporating characteristics found via multi-model analysis improves dependability and accuracy. The importance of prognostic characteristics in distinguishing between benign and malignant patients is emphasized in the research. The model uses ensemble machine learning methods to forecast breast cancer automatically in an organized manner. We have primarily used six algorithms: Decision Tree Logistic Regression, Naïve Bayes, SVM, KNN, and Random Forest. This creative method supports continuing efforts to enhance breast cancer preventive and early detection methods.

3.10.1 Logistic Regression (LR)

Logistic Regression (LR) is an elegant and powerful predictive analysis machine learning algorithm that draws inspiration from probability. Primarily designed to tackle two-class (binary) classification challenges, where outcomes are either 0 or 1, LR exhibits its finesse by gracefully compressing its output within the confines of a charming range—0 to 1—courtesy of its signature Logistic function. This intrinsic characteristic of LR imbues it with a subtle yet captivating charm, making it a favored choice in classification tasks. Statistical forecasting using logistic regression involves associating the probability of an outcome with a collection of explanatory factors. When several distinct factors determine what happens in the information set, it is utilized for analysis. The result is evaluated using a binary variable with a maximum of two potential outcomes. With this method, you may collect random variables to make a yes or no, true or false, or 1/0 prognosis. The LR model can be represented by the following equations [2]:

$$x = C_0 + \sum_{i=1}^n c_i x_i \quad (3)$$

$$P(x) = \frac{e^x}{1+e^x} \quad (4)$$

3.10.2 Decision Tree (DT)

The ingenious Decision Tree algorithm tackles classification and regression problems through its remarkable tree-like structure. At the breast of this structure lies the optimal dataset feature, thoughtfully chosen based on attribute selection measures like the Gini index or the Information gain. Each inner node elegantly represents an attribute or feature, while the graceful leaf nodes hold the key to the outcome, class, or dependent variable. With its versatile nature, this algorithm seamlessly handles both categorical and numerical data, making it an invaluable tool for various data analysis endeavors. Both absolute and numerical data can be resolved using this algorithm.

$$E = -\sum_{i=0}^k P_i \log_2 P_i \quad (5)$$

3.10.3 Random Forest (RF)

Random Forest (RF) is an ensemble Machine Learning technique capable of addressing regression and classification tasks. The potent supervised categorization technique is the random forest classifier. The RF classification is a combined technique that may be analyzed as a variation of the closest-neighbor prediction. The process of systematically developing and integrating techniques from statistics, such as classifiers or credible sources, to address a particular intelligence computing challenge is known as collective learning. Instead of producing a single classification tree from a given dataset, Random Forest creates an array of them. The following trees all include classifications with a particular combination of qualities. The method builds multiple decision trees and combines them to produce more robust and accurate predictions. By doing so, RF effectively mitigates the overfitting issues encountered with individual Decision Trees (DTs) when constructing models from training data. DT generally picks the best feature from the samples, leading to overfitting. But when several DTs are being considered, the overfitting problem can be easily solved. Each of these trees is a classification for a given set of attributes.[22]

$$\hat{Y} = \frac{1}{N} \sum_{i=1}^N f_i (X) \quad (6)$$

3.10.4 Support Vector Machine (SVM)

SVM is a supervised ML classification technique widely applied in the field of cancer diagnosis and prognosis. SVM functions by selecting critical samples from all classes, known as support vectors, and separating the classes by generating a linear function that divides them as broadly as possible using these support vectors.[7] It can be used particularly in healthcare-related issues and regression problems, but the main strength of SVM lies in classification. Support Vector Machines can effectively handle both linear and non-linear problems. Moreover, it demonstrates the accurate predictions of the target classes (labels) for new instances. Fig shows a scatter plot of two courses with two properties. A linear hyperplane is defined as $ax_1 + bx_2$, and the aim is to find a , b , and c such that $ax_1 + bx_2 \leq c$ for class 1 and that $ax_1 + bx_2 > c$ for class 2 [15].

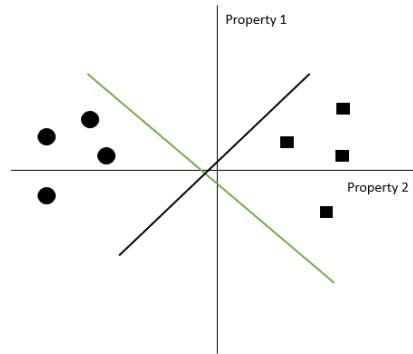


Fig 3.10.4 SVM generated hyper-planes

3.10.5 Naive Bayes (NB)

Naive Bayes is a fundamental and widely used AI technique, serving as a probabilistic classifier that makes classifications based on conditional independence assumptions with pre-trained datasets. It is commonly employed in various applications, including spam detection and medical diagnostics. The essence of Naive Bayes classifiers lies in discovering the probability distribution of classification problems. Bayes' Theorem, at the core of Naive Bayes, mathematically expresses the probability of an event occurring based on the probability of another event that has already taken place. The equation for Bayes' Theorem is as follows:

$$P(A|B) = P(B|A) * P(A)/P(B) \quad (7)$$

3.10.6 K-nearest Neighbor (KNN) The arrangement of and identification of patterns uses the K-nearest neighbor technique. Its prognostic analytic applications are many. When additional information arrives, the K-NN algorithm identifies the most nearby current information values. Any attributes that can differ on a large scale may sufficiently influence the interval between data points [19]. During instruction, we preserve the characteristics matrices and the group names. The measurement spatial representation of the information's values is assumed by K-NNs. During the classification process, the value is initially defined through the K retraining sampling neighbors, who are more constant. When that happens, the computation will find K neighbors within the fresh collection immediately nearby. The distance calculation method becomes a significant concern because the information's locations fall within the measurement area. If the number of neighbors is denoted by N in K-NNs, then N samples are considered using the following distance metric value [2]:

$$\text{Minkowski Distance: } \text{Dist.}(x, y) = (\sum_{i=1}^n |x_i - y_i|^p)^{\frac{1}{p}} \quad (8)$$

The distance between two points is the Manhattan distance when $p=1$, the Euclidean distance when $p=2$, or this distance when $p=\infty$.

On a worldwide basis, the most popular option is the Euclidean distance. The mathematical procedure will next look at the sum of data focused on each class amongst this K neighbors, and finally, the signal will assign the latest information to a classification that covers the most crucial portion.

3.11.7 Ada Boost (AB)

Because of machine learning jobs like breast cancer diagnosis, the AdaBoost (Adaptive Boosting) method merges several inadequate classifiers into one powerful one. The fundamental concept is to apply an ineffective learning technique iteratively, usually decision stumps, to updated data sets, each time providing greater importance to examples that were incorrectly categorized. The power of the method lies in its ability to edit the parameter values of each classification according to its mistake ϵ_t :

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \quad (9)$$

Afterwards, all future classifiers are compelled to focus on the examples that were incorrectly categorized earlier. The end model is a weighted combination of these ineffective classifiers $h_t(x)$:

$$H(x) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right) \quad (10)$$

By concentrating on the most difficult situations, AdaBoost may increase the prediction model's sensitivity and specificity in breast cancer diagnosis via iterative improvement of malignant tumor identification.

CHAPTER 4

Experimental & Result

4.1 Result and Discussions

In our study, we use various classification algorithms and assess their performance using evaluation metrics like accuracy, sensitivity, specificity, precision, and the F1 measure. The confusion matrix provides essential values such as True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Precision, a significant metric, measures the model's ability to distinguish between healthy and patient cases based on predictions. Analyzing these metrics gives us valuable insights into the algorithms' performance and makes informed decisions in our specific context.

$$Accuracy = \frac{(True\ positive + True\ negative)}{(True\ positive + False\ positive + True\ negative + False\ negative)} \quad (11)$$

The sensitivity test determines the accuracy in identifying patients with liver disease, focusing on the actual positive instances of the test. It is also called Recall or True Positive Rate (TPR).

$$Sensitivity = \frac{True\ positive}{(True\ positive + False\ positive)} \quad (12)$$

In particular, it shows the negative consequences of the disease. It gives the extent of the missing disease of the patients. It is otherwise called the True Negative Rate (TNR).

$$Specificity = \frac{True\ negative}{(False\ positive + True\ negative)} \quad (13)$$

Precision is otherwise called positive predictive value. It gives the proportion of an accurately predicted positive outcome by classifier algorithms.

$$Precision = \frac{True\ positive}{(True\ positive + False\ positive)} \quad (14)$$

F1 measures the precision of the model by a blend of accuracy and recall. It gives the proportion of both FP and FN of a model.

$$F1\text{-score} = \frac{2 \times (\text{Recall} + \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (15)$$

4.2 Result Analysis

4.2.1 Discussion Result

In this experiment, we considered different analyses to examine the six-machine learning classifier for the classification of breast cancer dataset. Regarding accuracy, NB achieved the highest accuracy of 0.96, and DT achieved the worst performance of 0.93. Concerning precision, RF, LR, SVM, KNN, and NB achieved the highest score of 0.96, and DT performed the worst at 0.94. When considering the specificity, NB & KNN reached the highest value, 0.99, and RF obtained the worst, 0.93. LR, NB & RF were also the best performers in the f1 measure at 0.96, and DT received the worst performance at 0.93, almost the same. According to these measurement criteria, the Naïve Bayes classification technique is more effective than the other classifiers for predicting breast cancer.

Table4.2.1: Applied Different Machine Learning Algorithms

Algorithm	Accuracy	precision	Recall	F1-Score	Specificity
DT	0.92	0.93	0.93	0.93	0.94
LR	0.95	0.96	0.96	0.96	0.98
NB	0.96	0.96	0.95	0.96	0.99
SVM	0.95	0.96	0.95	0.95	0.90
KNN	0.94	0.96	0.95	0.95	0.99
RF	0.95	0.96	0.96	0.96	0.93

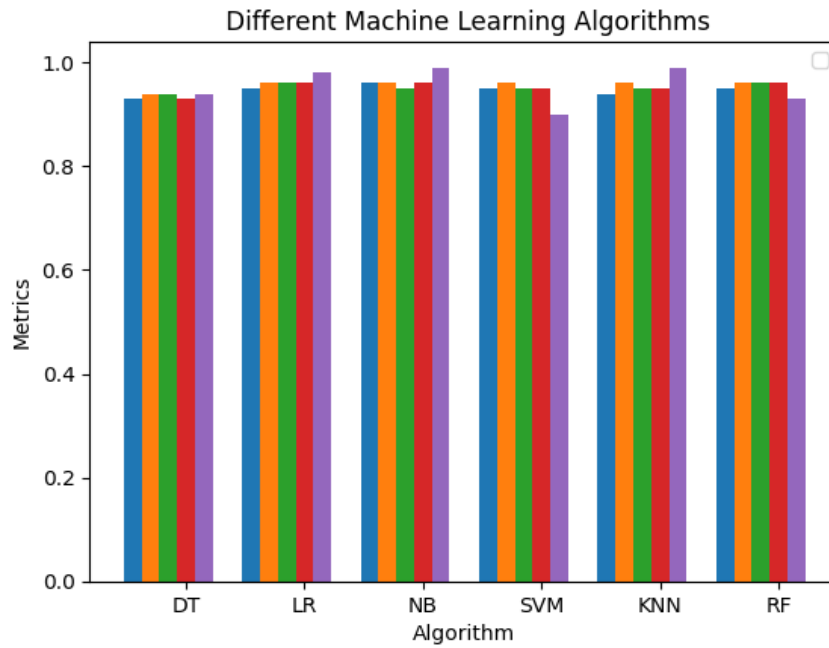


Fig 4.2.1: Different Machine Learning Algorithms

4.2.2 Confusion Matrix

When assessing the efficacy of categorization algorithms, a confusion matrix is an essential machine-learning tool. It clearly shows the model's predictions vs actual results, especially in identifying binary issues. Let's pretend we're doing a binary classification job with two classes: "malignant" (the positive class) and "benign" (the negative class), and use a set of data about breast cancer as an example of the confusion matrix. For a model to be considered "true positive," it must accurately forecast the occurrence of cancerous tumors. When the model accurately predicts cases of benign tumors, it is said to be true negative (TN). The model makes a Type I error—a false positive—when it wrongly predicts that an instance is cancerous when it is benign. A false negative (FN) occurs when an algorithm makes a type II mistake and wrongly predicts that a case is innocent while, in fact, it is cancerous.

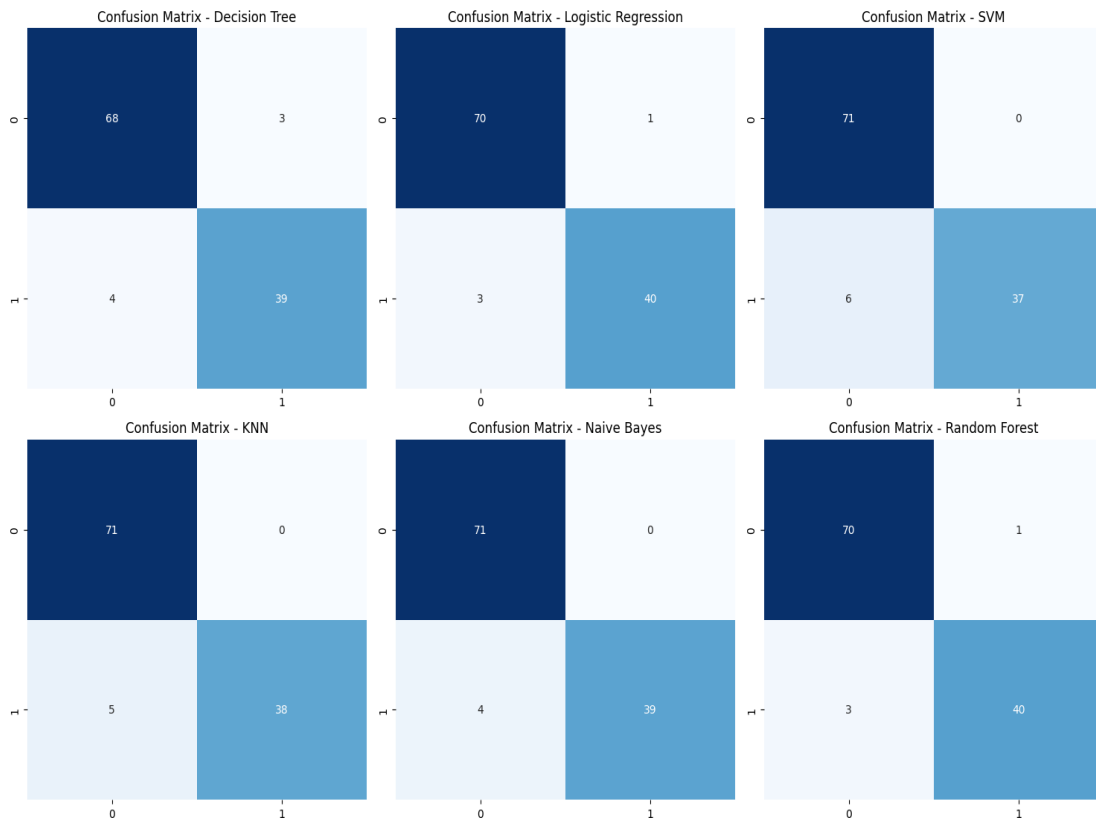


Fig 4.2.2: Metrics of Different Machine Learning Algorithms Confusion Matrix

4.2.3 Roc Curve

You may see the cost-benefit analysis for a classifier's effectiveness using a receiver operating characteristics (ROC) graph. Regarding ways to mine data, ROC is among the most popular and helpful efficiency gauges. This 2-D graph plots the TP rate (benefit) on the y-axis and the FP rate on the x-axis (cost) [20]. Around zero and one, the boundaries of the two axes extend. The TP and FP rates are obtained or shown on the graph for every conceivable classifier threshold. For the ROC to be considered more favorable, a point must be located in the top left quadrant, where the TP rate is significant and the FP rate is low. If the number of true positives and false negatives is equal, then the classifier is evil based on random guesswork. , A ROC graph's area under the curve represents the classifier's accuracy. You can cut the overall area of the graph by the area beneath the plot. The classifier performs better when the results are closer to 1.

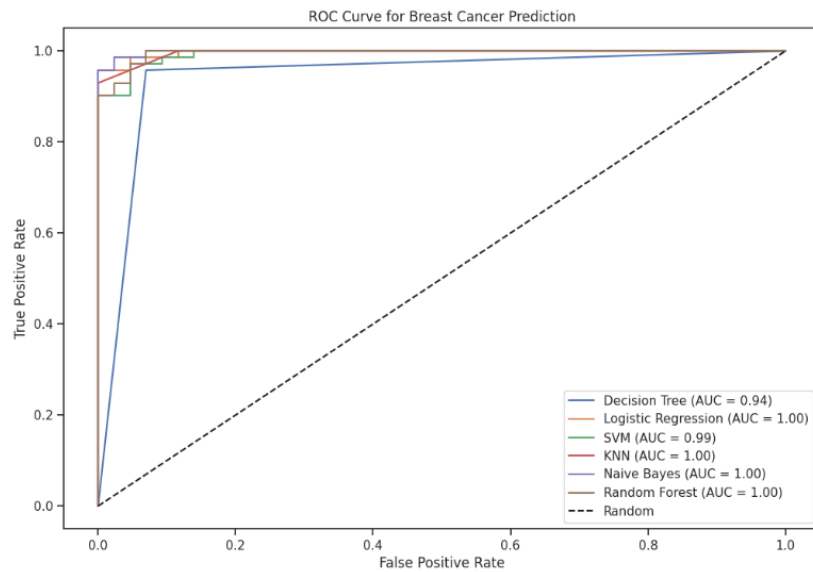


Fig 4.2.3: ROC Curve

4.2.4 AUC Curve

One of the metrics used to evaluate the effectiveness of screening procedures, particularly those for breast cancer, is the AUC (Area Under the Curve). AUC and Receiver Operating Characteristic (ROC) curves are frequently employed in breast cancer. These graphs show how the two parameters are traded off for various diagnostic procedure threshold settings. Curve for Detecting Breast Cancer: A test's diagnostic precision is graphically represented using ROC curves. They demonstrate the degree to which a screening or tests for diagnosis can distinguish between those who have the illness and those who do not. The study's general effectiveness is measured by the AUC value. Optimal biases are represented by a value of 1, while unpredictability is offered by a number near a value of 0. The diagnosis examination's capacity to separate out among positive and negative instances is more robust whenever its area under the curve (AUC) approaches 1. More excellent, reliable diagnosis is indicated by a larger AUC, and avoiding inaccurate results and incorrect results in breast cancer detection is vital.

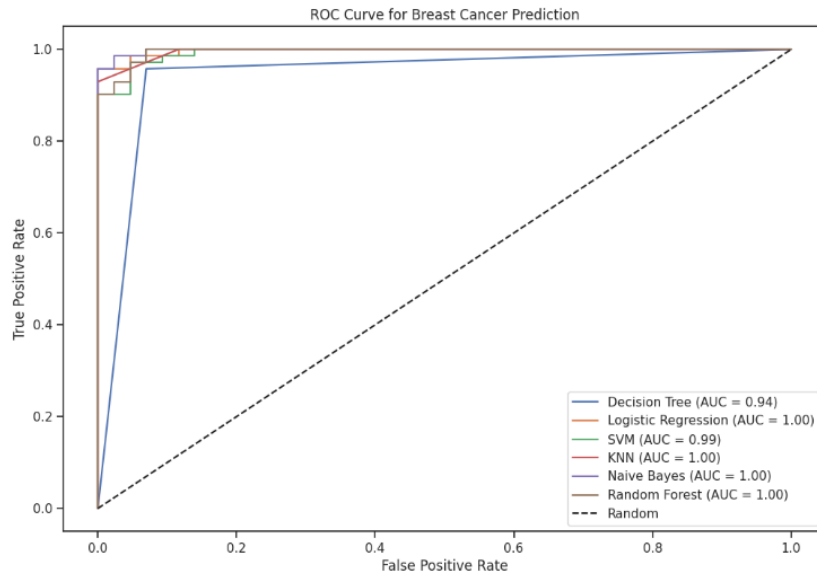


Fig 4.2.4: AUC Score

4.2.5 Jaccard Score

With machine learning, the Jaccard Score—sometimes called the Jaccard Index or Intersection over Union (IoU)—indicates how identical two sets are to one another. For example, the Jaccard Score is often used for assessing discrete classification algorithms when analyzing artificial intelligence algorithms trained on a breast cancer dataset.

The formula for Jaccard Score is given by:

$$Jaccard\ Score = \frac{\text{Intersection of Predicted and Actual Positives}}{\text{Union Predicted and Actual Positives}} \quad (16)$$

Crossover of Expected and Real Positives (True Positives): When the algorithm accurately detects cancerous tumors, like cancer in breast prognosis. A compilation of all occurrences that the system correctly or incorrectly predicts as malignant; this includes both true positives and false negatives. It is called the combination of anticipated and actual positives. Having 1.0 indicating an ideal fit among anticipated and actual positive events, a higher Jaccard Score implies superior modeling efficacy. To measure a model's accuracy in identifying malignant tumors in a breast cancer dataset, this metric helps gauge the proportion of true positives among all optimistic forecasts. DT & SVM achieved the highest Jaccard Score of 0.88, and KNN earned the worst performance, 0.78. On the other hand, the scores of the remaining two algorithms were 0.85 in NB and RF and 0.86 for LR score.

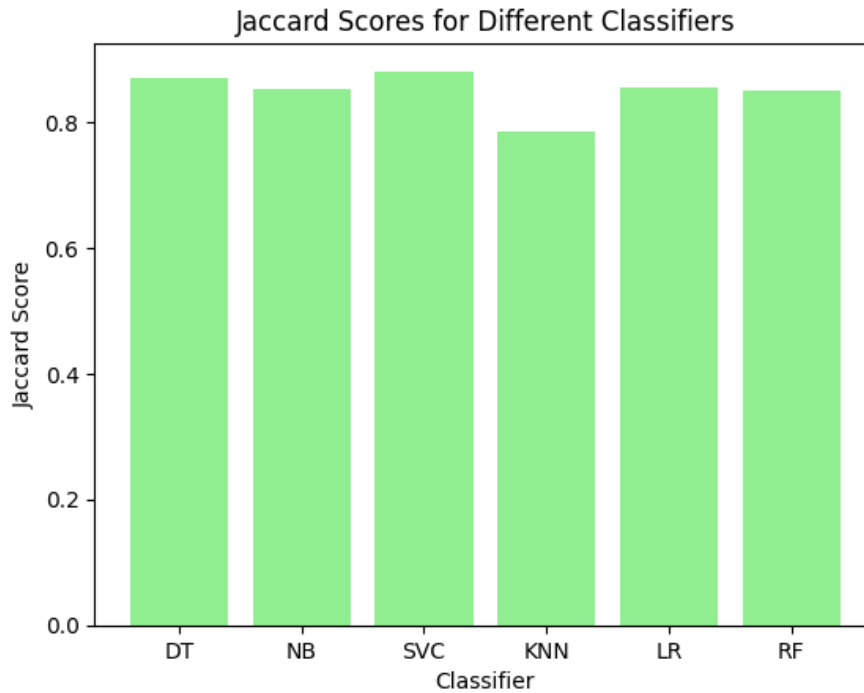


Fig 4.2.5: Metrics of Different Machine Learning Algorithms Jaccard Score

4.2.6 Mean Cross-Validation Score

Machine learning relies heavily on the Mean Cross-Validation Score, an indicator of effectiveness that measures how well a model can be applied to new data. Indeed, the algorithm works well with multiple data collection and is often applied within a breast cancer sample. A predictive algorithm's efficacy may be evaluated using the cross-validation approach, which involves dividing the information across numerous groups. We use half of those groups to train the model and to verify it, we use the others. An additional rigorous assessment is achieved by repeating this procedure repeatedly instead of only doing a single train-test split. The mean cross-validation score is the average of the assessment measure (e.g., accuracy) overall folds, determined after running cross-validation numerous times. This offers a more trustworthy and precise evaluation of the algorithm's efficacy. Information scientists and academics frequently employ k-fold cross-validation and other cross-validation procedures to train and test statistical models on various parts of the breast cancer sample. Being a fraction, the Mean Cross-Validation Score shows how effectively the equation performs, generally spanning different divisions; this number indicates that it can generalize exciting, unreliable information.

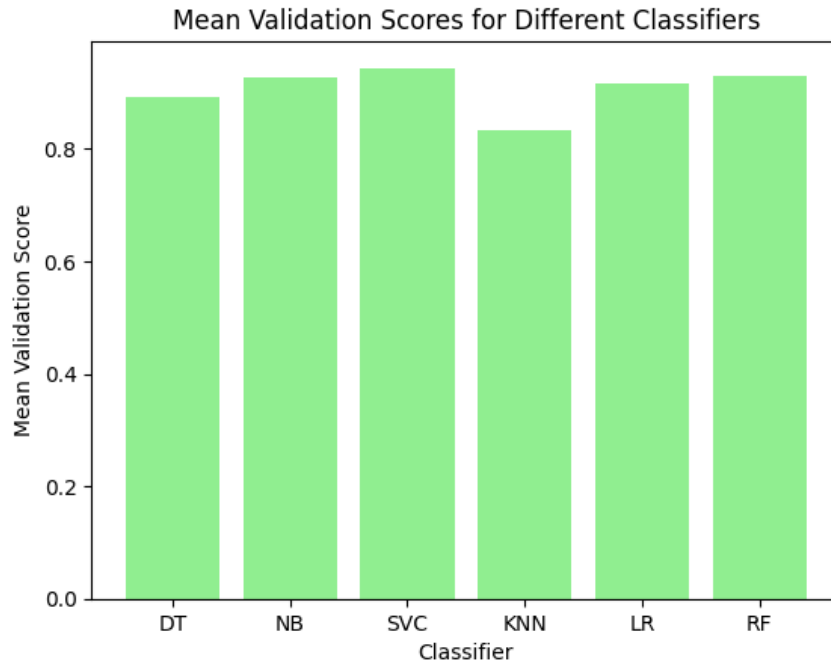


Fig 4.2.6: Metrics of Different Machine Learning Algorithms Mean Cross-Validation Score

4.2.7 Classification Error

One factor to consider when assessing the efficacy of machine learning (ML) models for breast cancer categorization is classifying error, which can be referred to as misperception frequency. The statistic shows how many cases were misclassified as a percentage of all occurrences. An identification mistake happens in the setting of breast cancer detection using machine learning techniques when the algorithm wrongly identifies a benign tumor as malignant or vice versa. Misclassification of cancer types could end up resulting in incorrect therapeutic options, which could have profound health implications. Statistical methods, including logistic regression, random forest, and K-nearest neighbor, are applied to evaluate classification errors. Using these techniques, we want to build algorithms that reliably identify malignant or benign cancerous breast cancer. It is essential to reduce classification mistakes to make breast cancer prediction models more reliable. Datasets like the Wisconsin Breast Cancer dataset are often enhanced in terms of accurate classification by academics via machine learning methods.

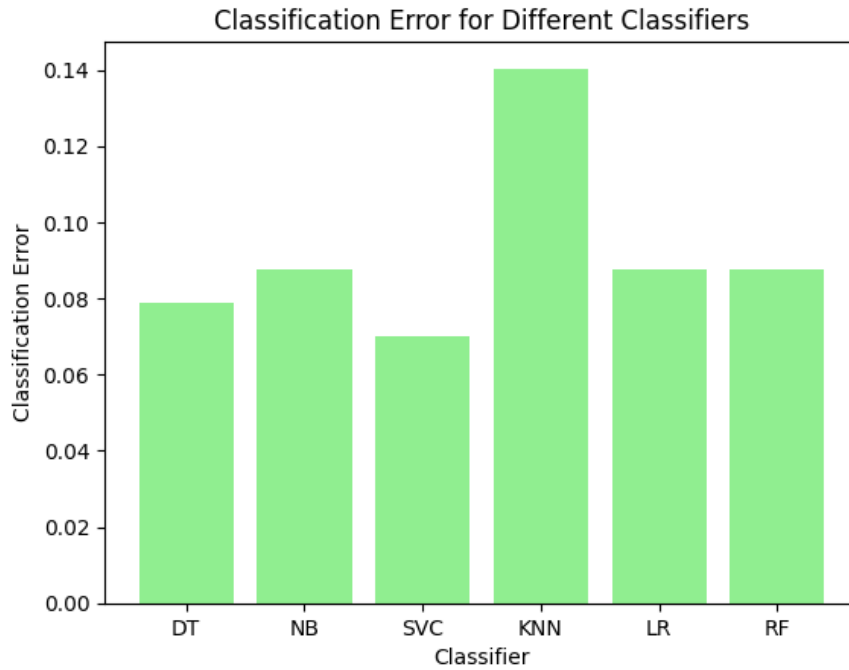


Fig 4.2.7: Metrics of Different Machine Learning Algorithms Classification Error Score

4.3 Comparative Analysis

The entire study is conducted using Google Collab [21], a free service for cloud computing. We use a ten-fold cross-validation method to efficiently evaluate the accuracy of ML techniques and avoid excessive fitting. For training and testing, we have divided the dataset 80:20. In this study, we evaluated four different algorithms—DT, LR, NB, SVM, KNN, and RF—based on their prediction abilities. Those methods are used for classification once data pretreatment procedures are completed without using the feature selection methodology. The figure shows that the accuracy rates for DT, LR, NB, SVM, KNN, and RF are 0.95, 0.97, 0.95, 0.81, 0.97 & 0.97, respectively. These algorithms may have enhanced prediction performance even more with AdaBoost's boosting technique. Figure 4.3.2 illustrates that the efficiency of such techniques was somewhat improved. Table 4.3.1 compares efficiency and precision, F1-score, recall & specificity scores, among other assessment parameters.

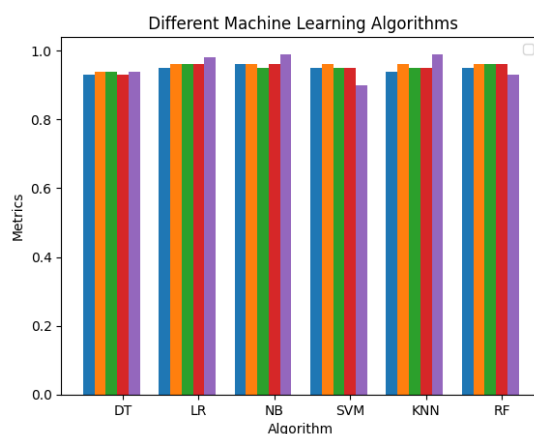


Fig 4.3.1: Performance using all Algorithm

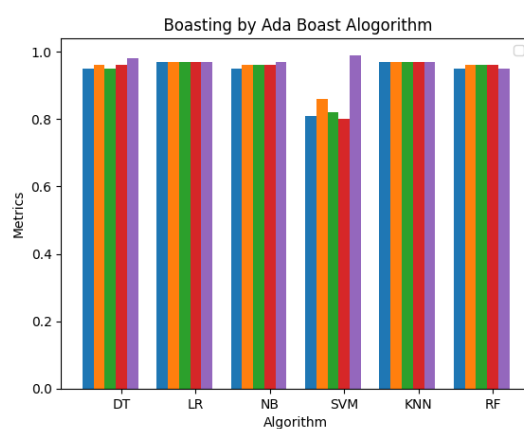


Fig 4.3.2: Algorithms (Boasting) Performance all Features

Table4.3.1: Applied Different Machine Learning Algorithms with Boasting by AdaBoast

	Algorithm	Accuracy	precision	Recall	F1-Score	Specificity
Classified Algorithm	DT	0.92	0.93	0.93	0.93	0.94
	LR	0.95	0.96	0.96	0.96	0.98
	NB	0.96	0.96	0.95	0.96	0.99
	SVM	0.95	0.96	0.95	0.95	0.90
	KNN	0.94	0.96	0.95	0.95	0.99
	RF	0.95	0.96	0.96	0.96	0.93
Boasted Model Classifier	DT	0.93	0.94	0.94	0.94	0.94
	LR	0.97	0.97	0.97	0.97	0.97
	NB	0.95	0.96	0.96	0.96	0.97
	SVM	0.81	0.86	0.82	0.80	0.99
	KNN	0.97	0.97	0.97	0.97	0.97
	RF	0.95	0.96	0.96	0.96	0.95

When the two factors are linearly related, Pearson's correlation coefficient will quantify the trajectory and intensity of that link. Subsequently, it may take on values between -1 and 1, with 1 being an ideal positive linear connection. -1 denotes a linear connection that is perfectly negative. No linear relationship is indicated by 0. To calculate the coefficient, take the product of the two variables' deviations from regular and divide it into the overlap.

$$r = \frac{cov(X,Y)}{\sigma_X * \sigma_Y} \quad (17)$$

Nous has previously demonstrated that algorithmic predictability is negatively impacted by less significant characteristics. Because TP is superfluous, we have used the Pearson Correlation Coefficient to eliminate it as a species. Such sixth methods are re-used for breast cancer patient prediction after PCC-FS data preprocessing. The accuracy values given by DT, LR, NB, SVM, KNN, and RF are 70.94, 0.96, 0.95, 0.93, and 0.97, as shown in Figure 4.3.3. Furthermore, as shown in Figure 4.3.4, the AdaBoost algorithm generated efficiency a value of 0.96, 0.97, 0.92, 0.78, 0.95, and RF when applied to DT, LR, NB, SVM, KNN, and RF, respectively. The table evaluates not just accuracy but also F1-score, precision, and recall scores, among other assessment metrics. To justify our work to the previous ones. The table confirm that we have got better accuracy than the mention earlier work.

Table 4.3.2: Comparison of Accuracy with Earlier Studies

Author	Methods	Accuracy
Sharma et al (2018) [5]	KNN	0.95
Obaid et al (2018) [6]	KNN	0.95
Gupta et al (2018) [8]	SVM	0.93
Chaurasia et al (2020) [10]	LR	0.94
Bayrak et al (2019) [4]	SVM	0.95
Our Model	PCC + Ada Boast	0.97

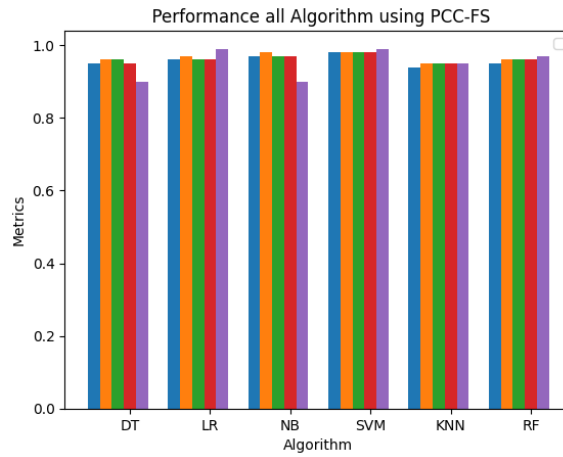


Fig 4.3.3: Performance all Algorithm using PCC-FS

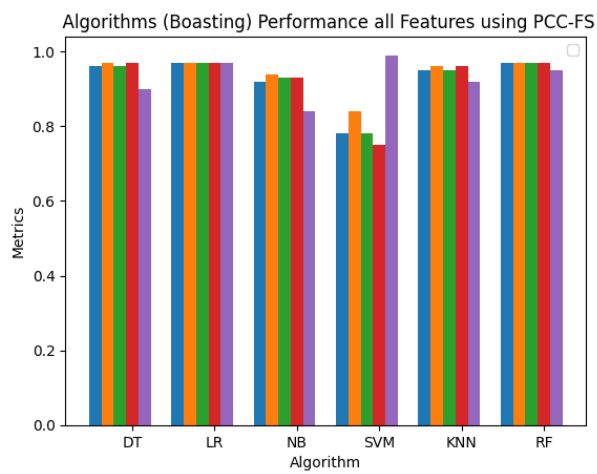


Fig 4.3.4: Algorithms (Boasting) Performance all Features using PCC-FS

Table4.3.3: Applied Different Machine Learning Algorithms with Boosting by AdaBoost using PCC-FS

	Algorithm	Accuracy	precision	Recall	F1-Score	Specificity
Classified Algorithm	DT	0.95	0.96	0.96	0.95	0.90
	LR	0.96	0.97	0.96	0.96	0.99
	NB	0.97	0.98	0.97	0.97	0.90
	SVM	0.98	0.98	0.98	0.98	0.99
	KNN	0.94	0.95	0.95	0.95	0.95
	RF	0.95	0.96	0.96	0.96	0.97
Boasted Model Classifier	DT	0.96	0.97	0.96	0.97	0.90
	LR	0.97	0.97	0.97	0.97	0.97
	NB	0.92	0.94	0.93	0.93	0.84
	SVM	0.78	0.84	0.78	0.75	0.99
	KNN	0.95	0.96	0.95	0.96	0.92
	RF	0.97	0.97	0.97	0.97	0.95

4.4 Discussion

This research primarily uses statistics for characteristics extraction to remove the superfluous qualities of data collection. Topics as diverse as emotion evaluation, medicine categorization, recognizing faces, and automobile driving may be compared using a questionnaire, as can scenarios like obesity, cancer of the cervical cavity perseverance percentage, and detectable signs of illness tumor cells. In cases when each information set of properties is obviously unimportant or insignificant, such as when identifying pedestrians, calculating credit scores, or discovering spam, a deviation from the requirements is apparent. The framework may be flexible by feeding the crucial classification data from stacking devices, a group, and modes. Data sets with few characteristics and binary categorization may implement this research approach after fundamental pretreatment. A few things could still be improved with the design now. More data specifically designed for categorization in medical and health records is needed, which increases the likelihood of missing outliers and values and adds extra data points that might impact the precision of classification. Considerations like the confusion matrix, specificity, accuracy, and additional indications should be part of any highly complex data set management strategy. Because of these issues, the suggested paradigm is not immediately applicable in a clinical setting. Memory utilization isn't the only factor that could impact performance; feature selection technique, patterns classifier type and quantity, and other similar decisions might have an impact. In the future, we may implement a method to verify whether the standard classifier is perfect and, if not, to create it from scratch. Deep learning algorithms may improve the accuracy of classification given more instances as more significant dimensions.

CHAPTER 5

Impact on Society, Environment & Sustainability

5.1 Impact on Society

Through analyzing information relating to unconscious trends, predictive methods may diagnose breast cancer earlier, increasing the chances of survival and improving the results of treatment. To maximize the efficacy of medication while avoiding negative consequences, statistical techniques aid in developing treatment regimens derived from specific patient data. Machine learning and intense learning methods enhance patient care by improving the diagnosis process, decreasing the occurrence of misdiagnosis, and cutting down on needless treatments. Saving money and better managing healthcare assets are two outcomes of automated breast cancer detection. By allowing for more precise examinations and preventative actions, better breast cancer model predictions aid society campaigns. With machine learning techniques, researchers can sift through enormous amounts of data in search of previously unknown patterns and relationships, which in turn leads to improved comprehension of the biology and therapy of breast cancer. With the help of forecasting algorithms, patients are better able to take an active role in their healthcare, make educated choices, and become more invested in their own recovery. Incorporating automated learning within breast cancer prediction aids worldwide initiatives to lessen the disease's impact, promoting better wellness for all people.

5.2 Impact on Environment

By optimizing the allocation of resources, machine learning models for breast cancer prediction help reduce the number of tests and treatments that are not essential. By lowering energy usage and trash output, this decrease in treatment procedures helps to minimize the environmental effect. Improvements in healthcare efficiency and accuracy in breast cancer detection are made possible by machine learning algorithms. Efficiency improvements can lessen the financial and ecological impact of healthcare delivery. Various machine-learning approaches to breast cancer detection are the subject of ongoing investigations on their potential environmental consequences. Clinically effective and ecologically sustainable approaches may be found via

comparative investigations. Continuous research in machine learning considers ecological factors, encouraging the creation of eco-friendly algorithms and models for breast cancer prediction.

5.3 Sustainability Plan

A machine learning-based strategy for the long-term prediction of breast cancer Create green computing-focused machine learning techniques to lessen human influence on the environment. To guarantee longevity after inaugural implementation, it is essential to systematically evaluate the long-term environmental effect of machine learning models used to detect breast cancer. Make a difference in reducing the impact of climate change by making efficient use of resources by adopting sustainable techniques in data gathering, model training, and estimating procedures. Encourage the development of sustainable healthcare information technology practices by funding continuing studies to find and use ecologically sound machine learning methods for breast cancer prediction. Insist on using only ethically sourced data. Verify that the data used to train the models is sourced sustainably and aligned with environmental objectives. Prompt scientists, doctors, and environmentalists to work together to find solutions that take responsibility for the environment and therapeutic efficacy into account. Get people involved in the conversation about long-term efforts to use machine learning for forecasting breast cancer so that everyone feels more accountable for their actions.

CHAPTER 6

Summary, Conclusion, Recommendation & Implementation for the Future Study

6.1 Summary of the Study

This machine-learning research on breast cancer prediction hopes to improve early diagnosis and prognosis by analyzing a wide range of clinical and genetic data. By using sophisticated algorithms, the model is very accurate in spotting possible instances, allowing for prompt intervention and ultimately leading to better patient outcomes. The study trains and validates the prediction model using a large dataset that includes imaging findings, genetic markers, and patient histories. The findings point to encouraging progress in the area, demonstrating how machine learning might be essential to breast cancer-tailored treatment. To improve diagnosis methods and, by extension, treatment plans, the research stresses the need to use state-of-the-art technology.

6.2 Conclusion

The principal goal of this work is to create an effective diagnosis system for breast cancer patients utilizing six distinctive supervised machine learning classifiers. We researched all the class executions on patient information parameters, and the LR and DT classifier gives the most elevated order exactness 0.97 dependent on the F1 measure to predict breast cancer, and SVM gives the most miniature precision 0.84. From now on, the outperform classification procedure will provide the decision support system disease. The application will have the option to predict breast cancer prior and advise the patient's well-being condition. This application can be gainful in low-salary nations where our absence of medicinal foundations and just as particular specialists. In our study, there are a few bearings for future work in this field. We just explored some popular supervised machine learning algorithms; more algorithms can be picked to assemble an increasingly precise model of liver disease prediction, and performance can be progressively improved. Additionally, this work likewise is ready to assume a significant role in health care research and just as restorative focuses to anticipate breast cancer.

6.3 Implementation of Future Study

To strengthen the prediction methods' resilience, future research on machine learning-based breast cancer identification should investigate the integration of many data sources like biological data, images, and medical records. To foster a sense of confidence between patients and physicians and to guarantee that procedures for making choices are understood, develop frameworks that are easier to interpret. Evaluate customized machine learning algorithms based on the unique traits of every individual, accounting for variations in genetics among other customized elements. Switch to continuous surveillance structures, which evaluate data continually to allow for rapid identification and prompt response. Improved results for patients. To make sure that machine learning models work well for various groups and ethnicities, test them on multiple groups of people. If you want to ensure justice and equity in the identification and prognosis of breast cancer, overcome moral challenges with limitations in machine learning models. Perform research to assess efficacy and applicability. Incorporating machine learning algorithms throughout healthcare processes to promote their broad use. Promote interdisciplinary cooperation among academics, physicians, and computational scientists to use their combined knowledge for deeper approaches.

Reference

- [1] Schneider, A. P., Zainer, C. M., Kubat, C. K., Mullen, N. K., & Windisch, A. K. (2014). The breast cancer epidemic: 10 facts. *The Linacre Quarterly*, 81(3), 244–277
- [2] Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, 1, 1-14.
- [3] Yassin, N. I., Omran, S., El Houby, E. M., & Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine*, 156, 25-45.
- [4] Bayrak, E. A., Kırıcı, P., & Ensari, T. (2019, April). Comparison of machine learning methods for breast cancer diagnosis. In 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT) (pp. 1-3). Ieee.
- [5] Sharma, S., Aggarwal, A., & Choudhury, T. (2018, December). Breast cancer detection using machine learning algorithms. In *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS)* (pp. 114-118). IEEE.
- [6] Obaid, O. I., Mohammed, M. A., Ghani, M. K. A., Mostafa, A., & Taha, F. (2018). Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. *International Journal of Engineering & Technology*, 7(4.36), 160-166.
- [7] Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th international conference on electronic devices, systems and applications (ICEDSA) (pp. 1-4). IEEE.
- [8] Gupta, M., & Gupta, B. (2018, February). A comparative study of breast cancer diagnosis using supervised machine learning techniques. In 2018 second international conference on computing methodologies and communication (ICCMC) (pp. 997-1002). IEEE.
- [9] Osmanović, A., Halilović, S., Ilah, L. A., Fojnica, A., & Gromilić, Z. (2019). Machine learning techniques for classification of breast cancer. In *World Congress on Medical Physics and Biomedical Engineering 2018: June 3-8, 2018, Prague, Czech Republic (Vol. 1)* (pp. 197-200). Springer Singapore.
- [10] Chaurasia, V., & Pal, S. (2020). Applications of machine learning techniques to predict diagnostic breast cancer. *SN Computer Science*, 1(5), 270.
- [11] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, 8, 150360-150376.

- [12] Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J., & Clarke, M. F. (2003). Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences*, 100(7), 3983–3988.
- [13] Sinn, H. P., & Kreipe, H. (2013). A brief overview of the WHO classification of breast tumors. *Breast Care*, 8(2), 149–154
- [14] B.M.Gayathri, C.P.Sumathi, and Santhanam. Breast Cancer Diagnosis Using Machine Learning Algorithms –A Survey, *International Journal of Distributed and Parallel Systems (IJDPS)* Vol.4, No.3, May 2013
- [15] G. Williams, “Descriptive and Predictive Analytics”, *Data Min. with Ratt. R Art Excav. Data Knowl. Discov. Use R*, pp. 193-203, 2011.
- [16] Kavitha R, Kannan E. An efficient framework for heart disease classification using feature extraction and feature selection technique in datamining. in: *IEEE Int. Conf. on Emerging Trends in Engineering Technology and Science (ICETETS)*, 2016, pp 1–5.
- [17] Uysal AK, Gunal S, Ergin S. The impact of feature extraction and selection on SMS spam filtering. *Electronics and Electrical Engineering*. 2013;19(5):67–72
- [18] Dua D, Graff C. *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science. 2019.
- [19] James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*. 1st ed. New York: Springer; 2013.
- [20] L. F. Carvalho, G. Fernandes, M. V. O. De Assis, J. J. P. C. Rodrigues, and M. Lemes Proença, “Digital signature of network segment for healthcare environments support,” *Irbm*, vol. 35, no. 6, pp. 299-309, 2014
- [21] "Google Colaboratory," [Online]. Available: colab.research.google.com. [Accessed: 12-May-2020].
- [22] James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*. 1st ed. New York: Springer; 2013

Plagiarism Report

plagiarism report final Arnab

ORIGINALITY REPORT

24% SIMILARITY INDEX	22% INTERNET SOURCES	12% PUBLICATIONS	14% STUDENT PAPERS
--------------------------------	--------------------------------	----------------------------	------------------------------

PRIMARY SOURCES

1	www.coursehero.com Internet Source	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
3	Submitted to Daffodil International University Student Paper	2%
4	link.springer.com Internet Source	2%
5	Submitted to Damonte Ranch High School Student Paper	2%
6	Submitted to Grambling State University Student Paper	1%
7	www.mdpi.com Internet Source	1%
8	Submitted to Higher Education Commission Pakistan Student Paper	1%
9	www.researchgate.net Internet Source	1%