BREAST CANCER CLASSIFICATION USING MACHINE LEARNING

BY

MD. TANZIR AHMED
ID: 191-15-2531

AND

PRANTO BISWAS
ID: 191-15-2687

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

Supervised By

Dr S.M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

MD. Sabab Zulfiker
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled **"Breast Cancer Classification Using Machine Learning"**, submitted by MD. Tanzir Ahmed, ID:191-15-2531, Pranto Biswas, ID:191-15-2687 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment 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 *Wednesday, 25th January, 2023.*

BOARD OF EXAMINERS

Chairman

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

Internal Examiner

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

Shayla Sharmin
25.1.23

Internal Examiner

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

Farid 25-01-23

External Examiner

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

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. S.M. Aminul Haque, Associate Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. S.M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

24.01.23

MD. Sabab Zulfiker
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Tanzir

MD. Tanzir Ahmed
ID: 191-15-2531
Department of CSE
Daffodil International University

Pranto Biswas

Pranto Biswas
ID: 191-15-2531
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We are really grateful and wish our profound indebtedness to **Supervisor Dr. S.M. Aminul Haque, Associate Professor,** 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 drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Co-Supervisor MD. Sabab Zulfiker, Lecturer** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in 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 currently the most common cancer globally and exceedingly threatening pointedly amidst women. Due to the complexity of breast tissues, accurate discerning and categorization of breast cancer is a crucial medical endeavor. “Machine Learning (ML)” approaches are bringing success in a mixed bag of spheres. Discreetly among the health sector, because of their range to automatically cutting attributes. The principal goal in this research is to identify and classify breast cancer and its stages by applying “Machine Learning”. In this study we used variety of “Machine Learning” approaches specifically Logistic Regression (LR), Decision Tree (DT) Classifier, Random Forest classifier, Support vector Classifier (SVC), K-Nearest Neighbor (KNN) Classifier, Adaboost Classifier, Gaussian Naive Bayes (GaussianNB), Gradient Boosting Classifier, Grid Search CV, Extreme Gradient Boosting (XGB) Classifier. The feature selection models used in this study are feature importance and Univariate Selection. Additionally, the proposed methods are using 10-fold cross validation to acquire the finest precision rate, as well as hyperparameter tuning in each classifier to assign the best parameters. This study used the best dataset obtained from the UCI repository. The implemented method’s performance was evaluated to determine “Accuracy”, “Sensitivity” and “Specificity”. When all techniques were compared, the “Logistic Regression Classifier” provided the best accuracy of 98.25%. As a consequence, the suggested technique outperforms existing methods since it classifies the optimal features automatically and experimental findings demonstrate the model outperformed previously published “Machine Learning” methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

Breast cancer is now the utmost familiar type of cancer worldwide, accounting towards 12.5% of new cancers diagnosed worldwide each year. Approximately 13% of women in the United States (around 1 between 8) will grow aggressive breast cancer in their life period. The United States will face around 287800 new cases of women diagnosed with aggressive breast cancer in 2022, and 51,400 will be identified with breast cancer in situ. By 2022, nearly 2,710 new cases of hostile breast cancer are anticipated to be detected in men. A man's life period risk of developing breast cancer is 1 between 833. Till January 2022, there will be over 38 lakh women living in the USA with a past of breast cancer. This contains females presently on treatment besides those who have discontinued medication.

Between American women Breast cancer is the most frequently detected cancer. By 2022, it is estimated that approximately 30% of newly identified cancers in females will be breast cancer. The maximum significant risk factors for breast cancer are female sex and aging. If you are a transgender man or woman, it is important to discuss your personal risk level with your doctor. In the USA, approximately 43,255 women are expected to pass away from breast cancer in 2022. Mortality from breast cancer has been steadily declining since 1989, and by 2020, he will be down 43% overall. These declines are believed to be the result of advances in treatment and early detection through screening. However, the downward trend has slowed down in recent years. Breast cancer is the main cause of cancer related deaths for women in the USA, behind lung cancer.

In the last few decades, many businesses have acquired vast troves of data from a wide variety of sources and in a wide variety of formats. These findings may have medicinal, agricultural, and climate-forecasting uses, among others. Data is growing exponentially, and traditional approaches to analyzing it, sifting through paperwork to find relevant information, and mass-producing choices are becoming increasingly inadequate. Data mined from medical record archives may be analyzed using a variety of machine learning techniques, including classification, clustering, and regression. It has been demonstrated that machine learning algorithms may be used to successfully forecast sickness by gleaning information from medical data sources. Several papers have detailed how machine learning algorithms were used to make breast cancer forecasts. Prophetic models for use in practical decision making for breast cancer predictions have been built using machine learning algorithms to great effect. Predictive models to help in making informed decisions for categorizing breast cancer have been built using machine learning algorithms.

1.2 Motivation

“Machine learning” algorithms have been utilized as tools to develop prediction models for BC in order to assist clinicians' judgments with sufficient exactness. These models, however, have several drawbacks, such as the use of suitable approaches to fit the model based on the dataset without taking into account feature abstraction methods; adequate feature extraction techniques efficiently lessen dimensionality for improved illness estimation. There is also increasing worry about the approaches for dealing with missing values in the dataset. As a result, researchers created an enhanced machine learning classifier to provide accurate breast cancer forecasts and raise women's survival rates.

1.3 Statement of the Problem

This research will help to classify breast cancers into primary problems: (Benign and Malignant tumor). The benign cancer is one that doesn't spread outside of the host or invade the tissue around it. Malignant cancers are a different kind of cancer that can banquet all over the host's body or occupy the tissue around them. On rare occasions, benign cancers can also cause death, but generally speaking, they are not nearly as deadly as the malignant cancers. On the other hand, malignant cancers are similar to those killer bees. They will simply spread out and attack you in mass, and if they are aggressive enough, they may even kill the person if you are not doing anything to them or are anywhere near their hive.

Manually classifying cancers into benign and malignant types can be time consuming, prone to human mistake, and tedious. During construction, the projected technique can automatically classify different types of cancer into safe (benign) categories of risk (malignant). This machine uses machine learning algorithms to perform this function. The scope of this new system is: Classification errors are greatly reduced, early failure analysis can be performed without human error, and devices are never switched off. However, researchers are using machine learning to identify and classify breast cancer.

1.4 Aim and Objective

This study's major objective is to utilize "Machine Learning" for breast cancer detection and categorization.

The goals of this study are to specifically:

- ❖ Use a Machine Learning algorithm to classify breast tumors as benign or malignant.
- ❖ Breast cancer detection in its earliest stages by mammography screening using machine learning.
- ❖ Processing mistakes in breast cancer classification can be eliminated with the use of a Machine Learning system.

1.5 Research Question

The following are the study queries:

- ❖ How does a “Machine Learning” classify breast cancer as benign or malignant?
- ❖ To what degree can a machine learning system classify breast cancer?
- ❖ How do we eliminate breast cancer classification errors while using a machine learning structure?
- ❖ How can we get a more accurate result?

1.6 Significance of Study

Our study has huge significance in the medical world. In this study we tried to exploit “Machine Learning” to classify breast cancer. The study tries to classify breast cancer more accurately and quickly.

1.7 Limitation of the Study

There are some limitations in this study that we would like to work with in the future. The algorithm used is predefined and not updated. The data set is not large. Using the voting classifier in the ensemble method didn't make the accuracy go significantly higher. All ensemble methods were not used.

1.8 Report Layout

The paper is organized as follows:

i. CHAPTER 1: Introduction
ii. CHAPTER 2: Literature Review
iii. CHAPTER 3: Research Methodology
iv. CHAPTER 4: Experimental Results and Discussion
v. CHAPTER 5: Impact on Society, Environment and Sustainability
vi. CHAPTER 6: Conclusion and Future Work
vii. Reference

CHAPTER 2

Background

2.1 Overview of Breast Cancer

A primary site of breast cancer development. The condition manifests itself when a breast lump forms as a result of unchecked growth and transformation of tissue. Breast cancer, like many others, can spread to lymph nodes and other lymph node sites. If it spreads to other organs, it can cause the growth of new tumors there as well. The term "metastasis" describes this phenomenon.

Among women, skin cancer is by far the most common form of cancer, but breast cancer is a close second. As a general rule, it occurs more frequently in women over the age of 50. Males can develop breast cancer, albeit it's quite uncommon. Two thousand six hundred men in the United States are diagnosed with breast cancer per year. It's fewer than 1% of all instances, by the way. Breast cancer often strikes individuals over the age of 50, however anybody can be diagnosed with this disease.

2.2 Risk Factors

Simply put, anything that raises your chances of developing breast cancer is a risk factor for the ailment. Having a family history of breast cancer or other risk factors does not guarantee that you will get breast cancer. There are often no known causes for the development of breast cancer in many women.

Risk of breast cancer is linked to the following factors:

- Compared to men, breast cancer primarily affects females. To put it another way, getting older. You are more likely to grow breast cancer as you become older.

- My family has a history of breast cancer. Those who have undergone breast surgery and had lobular carcinoma in situ (LCIS) or atypical hyperplasia of the breast discovered are at an increased risk of developing invasive breast cancer in the future.

- Having one breast affected by breast cancer increases your risk of developing the disease in the other.

- If you have a mother, a sister, or a daughter who has been diagnosed with breast cancer, especially at a young age, your risk of having the disease is increased. However, the great majority of people who get breast cancer have no family history of the disease.

- Cancer susceptibility in the family Mutations in the genes that increase susceptibility to breast cancer can be passed on from parents to children. Most mutations in these genes occur in BRCA1 and BRCA2. Although having one or more of these genes greatly increases your chance of developing breast cancer or another form of cancer, it is still not a certainty.
- Ill effects from being around radiation increase your chance of developing breast cancer is higher if you were exposed to radiation in your chest when you were a child or a young adult.
- A higher chance of developing breast cancer is associated with being overweight.
- Having your first period before you turn 12 raises your risk of developing breast cancer.
- A complex risk of breast cancer is associated with a later menopause onset age.
- The risk of breast cancer may be advanced in women who have their first child beyond the age of 30.
- In contrast to women who have had one or more pregnancies, those who have never experienced motherhood are at a higher risk of evolving breast cancer.
- Hormone treatment drugs, such as those containing estrogen and progesterone, are used to alleviate menopausal symptoms, but they can raise the risk of breast cancer in women. When women stop using these drugs, they reduce their risk of developed breast cancer.
- Breast cancer risk rises with alcohol consumption.

2.3 Symptoms of Breast Cancer

Understanding how your breasts naturally look and feel is essential for breast health. Mammograms do not detect all breast cancers, despite the necessity of regular breast cancer screenings. In order to detect changes in your breasts, it is essential to be familiar with their typical appearance and texture.

The most prevalent breast cancer symptom is an innovative lump or tumor (though most breast lumps are non-cancer). However, breast cancer can also be pulpy, rounded, tender, and even sore.

Other possible breast cancer signs contain:

- ❖ Enlargement of all or a percentage of the breast (even if no swelling is felt)
- ❖ Pockmarks on the skin (occasionally looks like orange peel)
- ❖ Chest and neck discomfort
- ❖ Skin that is red, dry, scaly, or thickened on your genitalia or breasts
- ❖ Nipple ejection (other than breast milk)
- ❖ Lymph bumps that are enlarged underneath the arm or nearby the clavicle (this may be a sign that the breast cancer has blowout, even earlier the unique breast tumor is palpable).

2.4 Screening/ Diagnosis of Breast Cancer

Screening searches for illness indicators, such as breast cancer, before symptoms manifest. The purpose of cancer screening is to detect the disease at an early stage, when it is treatable and curable. Cancers that are very tiny or very slowly growing may be detected by screening testing. These cancers are unlikely to result in death or illness within an individual's lifetime.

Scientists are attempting to determine who is more prone to get specific types of cancer. For instance, they consider the patient's age, family history, and specific stressors. This information assists physicians in determining who should undergo cancer screening, which tests to employ, and how frequently.

The below tests and approaches are used to stage breast cancer:

- Blood test such as a complete blood count
- Detection of cancer signals in the other breast by mammography
- Breast MRI
- Bone scan
- Computed tomography (CT) scan
- Positron emission tomography (PET) scan

2.5 Treatment of Breast Cancer

There are various therapy options for breast cancer patients. Some treatments are mainstream (already employed) treatments, while others are being evaluated in clinical studies. An investigation designed to discover new remedies for patients. If clinical trials demonstrate that the new treatment is superior to the current standard, the new treatment may replace the current standard. Patients could wish to participate in a clinical trial. Certain clinical studies are restricted to people who have not yet begun treatment. There are numerous forms of breast cancer treatments.

The following forms of treatment are employed:

- surgery
- radiotherapy
- chemical treatment
- hormone therapy
- targeted therapy
- immunotherapy

Clinical trials are being used to test new medicines. The treatments for breast cancer have adverse effects. Patients could wish to participate in a clinical trial. Patients may engage in clinical trials prior to, during, or subsequent to beginning cancer treatment. Additional testing might be necessary.

2.6 Related Work

Numerous new instances of reputation strategy software for screening breast cancer have been published in recent years. Anusha Derangula [2] and associates reportedly used Gradient Boosting Techniques from Machine Learning for Feature Selection of Breast Cancer Data. In this case, we employ the feature-selection methods of the Light Gradient Boosting Model (LGBM), Catboost, and Extreme gradient boosting (XGB) (XGB). With the help of the enhanced features, the Naive Bayes classifier was able to attain a 96.49% success rate (Anusha Derangula et al.,2020).

Linear Regression, Decision Tree, Multi-Layer Perceptron, and K-Nearest Neighbor were only few of the ML techniques used by Muatz Humida and V. Chitraa [3]. (Muatz et al.,2019).

Both of these optimization strategies were employed by Idri, A., and Hosni [12]. Grid search and optimization of particle swarms are two examples. Four BC datasets were chosen to evaluate the predicting performance of the suggested ML approaches.

Six supervised machine learning methods, including k-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine with a radial basis function kernel, are introduced by Gupta, P., and Garg [10]. Also used was Adam Gradient Descent Learning, a form of deep learning that combines adaptive gradient algorithm with root mean square propagation. Each model is demonstrated to have undergone a different hyper parametric adjustment that improves accuracy both internally and when compared to the others.

Bhardwaj, Arpit [6] used the WBCD dataset to categorize patients as benign or malignant using multilayer perceptron (MLP), K-nearest neighbor (KNN), genetic programming (GP), and random forest (RF). When compared to other classifiers, the findings demonstrate that RF has the highest classification accuracy at 96.24%.

In [24], Michael and Epimack offer an automated CAD system that may provide a very efficient method. Thirteen of the 185 features available for use in machine learning training are actually used. To determine whether tumors were malignant or benign, five machine learning classification tools were applied. Using a machine learning predictor for 10-fold cross-validation, the experiments uncovered Bayesian optimization using a tree-structured Parzen estimator.

Using the Weka software, Akbugday and Burak [11] assessed the performance of three distinct machine learning algorithms: k-Nearest Neighbors (k-NN), Naive Bayes (NB), and Support Vector Machine (SVM). Different relevant settings for each method have been used to get the results. k-NN, NB, and SVM algorithms have all been studied for their potential to correctly categorize breast cancer data. After analyzing the data, we find that the most effective classification algorithms are k-NN under the k = 3 condition and C-SVM with parameters C = 215and = 2-15 and that uses a Radial Basis Kernel, achieving an accuracy of 96.85%.

Accuracy of 99.2%, recall of 98.0%, and precision of 98.0% on the WBCD dataset and accuracy of 79.5%, recall of 76.0%, and precision of 59.0% on the WPBC dataset were achieved using LDA-SVM when the median was used to compute missing values (Egwom, Onyinyechi Jessica [14]).

With the goal of better predicting breast cancer, Gopal, V. N., Al-Turjman [20] looked at the usage of machine learning techniques in conjunction with IoT sensors. The suggested classifier achieved 98% precision, 97% recall, 96% accuracy, and 98% F Measure. Minimal rates of MAR, RMSE, and RAE for the classifier, respectively.

Using linear discriminant analysis (LDA) to first decrease the high dimensionality of features, Omondiagbe [23] suggested a hybrid strategy for breast cancer diagnosis by first applying the new reduced feature dataset to Support Vector Machine. Accuracy of 98.82 percent, sensitivity of 98.41 percent, specificity of 99.07 percent, and region underneath the receiver's operating characteristic curve of 0.9994 were achieved using the suggested method.

Mashudi, Nurul Amirah [20] employed 10-fold cross validation and a number of machine learning approaches to predict the survival of breast cancer patients. These methods included k-Nearest Neighbors (k-NN), Random Forest, and Support Vector Machine (SVM), as well as ensemble techniques. Two-fold, three-fold, and five-fold cross validation are also used to classify the recommended procedures for the maximum accuracy rate possible. Accuracy, sensitivity, and specificity were calculated in order to evaluate the performance of the suggested approaches. AdaBoost ensemble approaches achieved the highest accuracy (98.77%) across all evaluation metrics, with 2-fold cross validation yielding 98.41% and 3-fold cross validation yielding 98.24%. Five-fold cross-validation results, however, demonstrate that SVM achieved the highest accuracy (98.60%) with both the greatest error rate.

Breast cancer was first recommended for classification using the K-Star system by Mohamed Sakr and his group [26]. (Mohamed Sakr et al., 2020). The results proved that the K-star algorithm was superior than its contemporaries. Based on experimental results, the K-star approach was determined to have the best levels of accuracy (97.142%), sensitivity (95.244%), specificity (93.3%), and area under the curve (0.998). Different classifier algorithms include K-star, Clonal Selection Algorithm (CLONALG), and Artificial Immune Recognition System (AIRS).

2.7 Uses of Machine Learning in Breast Cancer Classification

There is an burning need for novel approaches to breast cancer screening, diagnosis, and therapy because breast cancer is one of the leading causes of mortality among women globally. In this research, we present a novel ML-based framework for assisting in the analysis of breast cancer. In specifically, we introduced ML and discussed how it may be used for breast cancer categorization. A total of 569 patients were analyzed using the ML method, and the outcomes presented that 37% of the patients had been diagnosed with malignant tumors and 63% had been identified with benign tumors. Breast tumors can be classified as either benign or malignant based on their cell pattern, texture, perimeter, area, compactness, concavity, and convexity. Smoothness, symmetry, and fractal dimension of cell pictures are not predictive of a more favorable diagnosis. Tumors that start out as benign can sometimes progress to malignant forms. We anticipate that ML's use in healthcare settings, particularly hospitals, will increase significantly over the coming years. The outcomes of ML-based research can be implemented in clinical decision-making tools.

CHAPTER 3
Research Methodology

3.1 Design Language and Tools of the Proposed System

Python was the programming language used for this study. Python contains a modular machine learning library called PyBrain that provides simple-to-use machine learning techniques. The most effective and trustworthy coding solution necessitates a well-structured environment and established Python frameworks and modules. In addition, this paper utilized Google Colab, Kaggle, UCI, and Microsoft Excel. The majority of the code was performed with Google Colab. The data were retrieved from the UCI repository and placed in an Excel spreadsheet.

3.2 Implementation Analysis and Design of Our Proposed System

At the beginning of this process, a comprehensive diagram is created representing the typical workflow to be followed. As the diagram makes evidence, the first step in our methodology is to collect data. This study used Wisconsin-diagnosed breast cancer (WDBC) Dataset obtained from UCI Machine Learning repository. After we obtained the dataset, we preprocessed the dataset to make the data set more accurate and suitable for our work. The next step was feature selection. Feature selection is a way to reduce the number of input variables in a model by using only relevant data and leaving out noise. After that we used many machine learning classification techniques to get the best accuracy. Next step is Hyperparameter tuning. Hyperparameter tuning consists of finding the optimal set of hyperparameter values for a learning algorithm while applying the optimized algorithm to an arbitrary dataset. This combination of hyperparameters maximizes model performance and minimizes the defined loss function to produce better results with fewer errors. And finally, after cross validation and performance evaluation we get the final result.

figure 3.2.1 we can see the full workflow diagram of our work

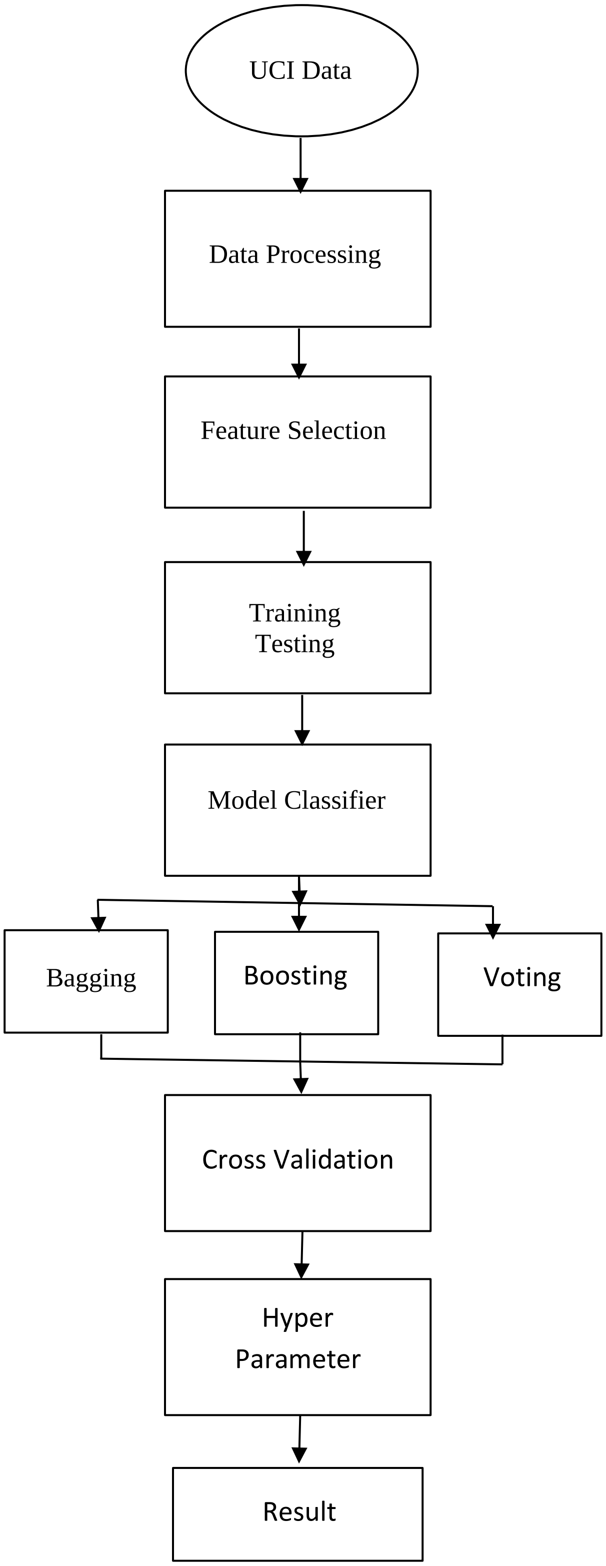


Figure 3.2.1 Procedural Framework

3.3 Data Collection

In our research we used “Wisconsin-diagnosed breast cancer (WDBC)” Dataset acquired from UCI Machine Learning repository. This dataset is multivariate and regrouped 569 instances in total. Attributes are calculated from digitized fine-needle aspiration images of breast masses. They designate the information of the cell nuclei existing in the image. It contains a total of 31 features or attributes. Here 357 cases are benign and 212 are malignant. 31 features included in this dataset have information as given, “ID number, diagnosis (malignant or benign), radius, texture, perimeter, area, smoothness, compactness, concavity, concave point, symmetry, fractal dimension”. In figure 3.3.1 we can see the combination of two stages value and predicted attribute,

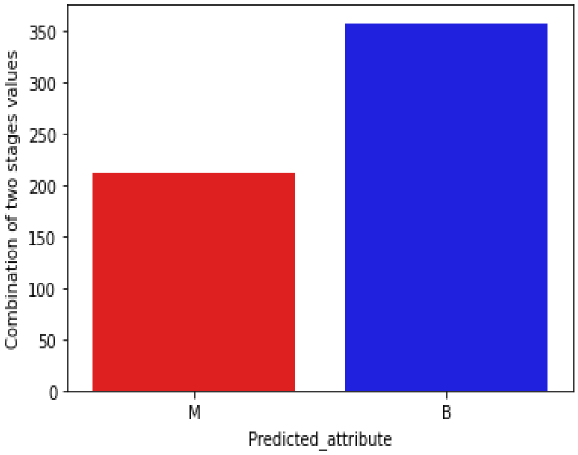


Figure. 3.3.1 Combination of two stage values and predicted attributes

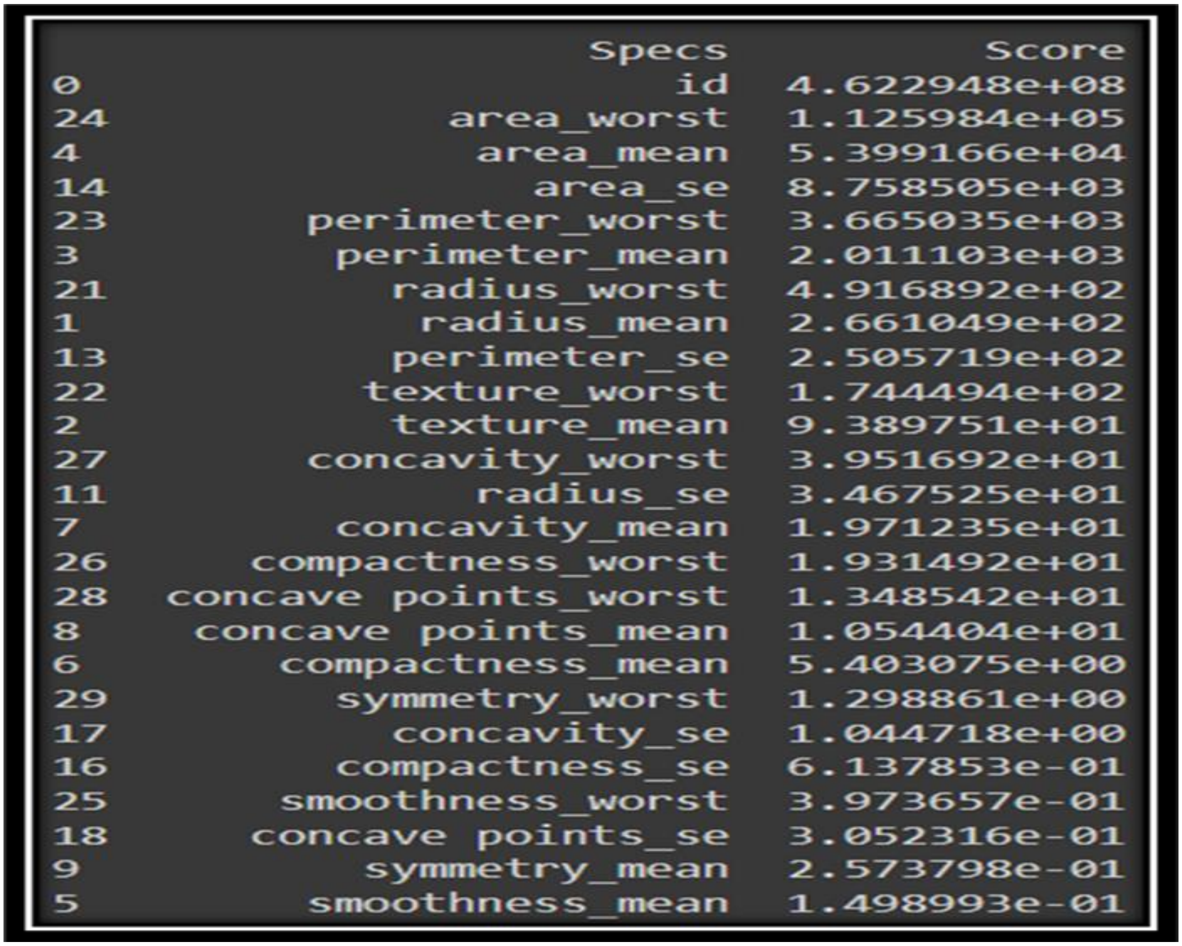
3.4 Data Processing

In all model’s, preparing data is the first step before running a simulation. The first step in the preprocessing phase is to guarantee that BC records have no absent data. Then we select malignant (M) and benign (B) from diagnosis. After that we renamed ‘Diagnosis’ into ‘target’. Then we changed malignant (M) and benign (B) from categorical (M & B) to numerical (1 & 2). At last, we put all the attribute data into a standard scaler. And then the data is ready for classification.

3.4.1 Feature Impact Analysis:

The idea of 'feature impact' is used to find which features in a dataset have the greatest impact on the conclusions that a machine learning model draws from analyzing the data. Additionally, feature influence is used for feature selection, which is one of the best ways to improve model accuracy.

For our study, in the dataset there are a total of 30 features or attributes. In Figure 3.4.1.1. we can see we used the best 25 features



```
                          Specs         Score
0                            id  4.622948e+08
24                   area_worst  1.125984e+05
4                     area_mean  5.399166e+04
14                      area_se  8.758505e+03
23              perimeter_worst  3.665035e+03
3                perimeter_mean  2.011103e+03
21                 radius_worst  4.916892e+02
1                   radius_mean  2.661049e+02
13                 perimeter_se  2.505719e+02
22                texture_worst  1.744494e+02
2                  texture_mean  9.389751e+01
27              concavity_worst  3.951692e+01
11                    radius_se  3.467525e+01
7                concavity_mean  1.971235e+01
26            compactness_worst  1.931492e+01
28         concave points_worst  1.348542e+01
8           concave points_mean  1.054404e+01
6              compactness_mean  5.403075e+00
29               symmetry_worst  1.298861e+00
17                 concavity_se  1.044718e+00
16               compactness_se  6.137853e-01
25             smoothness_worst  3.973657e-01
18            concave points_se  3.052316e-01
9                 symmetry_mean  2.573798e-01
5               smoothness_mean  1.498993e-01
```

Figure. 3.4.1.1 Feature Selection

Figure. 3.4.1.2. represents the Feature Importance Bar Chart. Feature importance describes how a model assigns weights to each of its input features. Each feature's score merely reflects its "importance." An increased score indicates that this attribute has a larger effect on the variable prediction model.

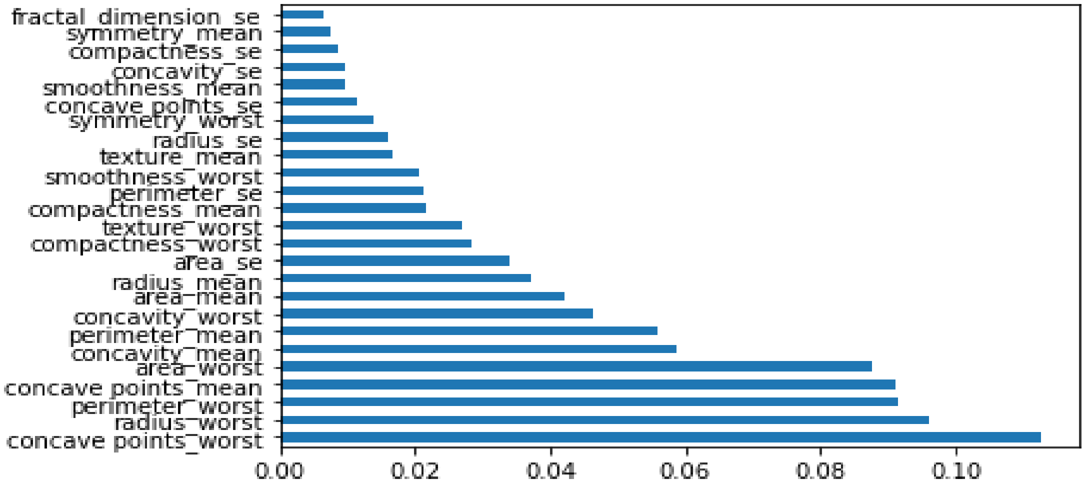


Figure.3.4.1.2 Feature Importance Bar Chart

3.4.2 Pearson Correlation

The linear link between two variables can be quantified using the Pearson correlation coefficient (r). The goal of the Pearson correlation is to construct a line that best fits the available data for the two variables. Distance from each data point to the line of best fit is represented by the Pearson correlation coefficient, r.

The Pearson correlation coefficient permits variables to be measured in distinct units. Height and weight, for instance, can be connected. It is constructed so that the unit of measurement has no effect on covariation investigations.

Pearson's correlation coefficient (r) is a unitless correlation measure that has no effect on origin or scale shift measures. It disregards the classification of the variable as either dependent or independent. Treat all variables equally. You may wish to determine whether basketball performance correlates with height. The same result might be drawn, though, if we tried to determine if a person's height was determined by their basketball prowess (which is completely absurd).

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad \text{---(1)}$$

where x_i, y_i, are the variables and x bar, y bar, are the mean, respectively. Figure 3.4.2 represents Pearson Correlation Heat Map of our study.

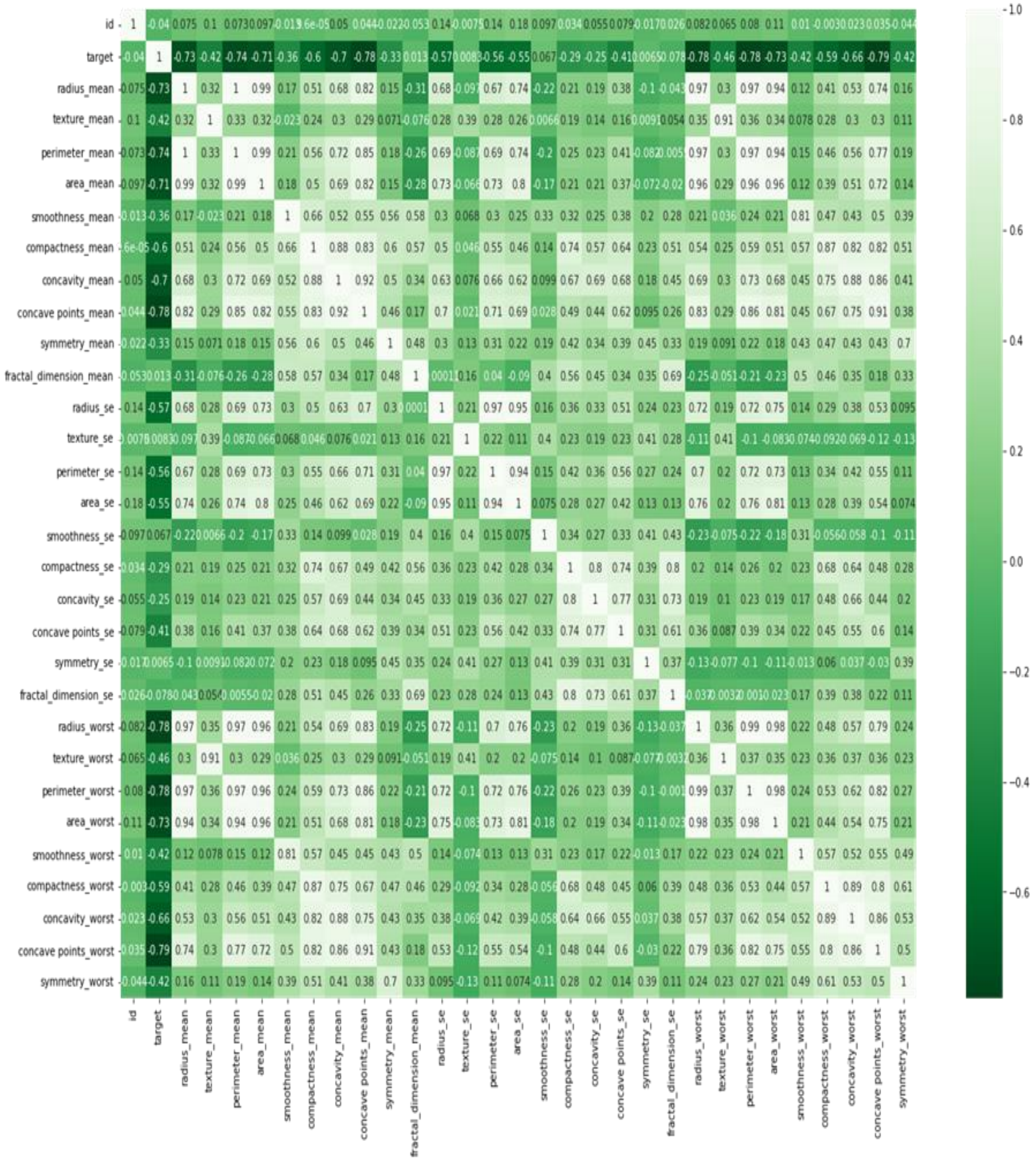


Figure 3.4.2 Pearson Correlation Heat Map

3.5 Techniques of Proposed System

Logistic Regression Classifier (LRC), Random Forest Classifier (RFC), Gaussian Nave Bayes Classifier (NBC), Decision Tree Classifier (DTC), K-Nearest Neighbors Classifier (KNNC), Gradient Boosting Classifier (GBC), AdaBoost Classifier (ABC), XGBoost Classifier (XGBC), and Support Vector Classifier (SVC) were all implemented using standard machine learning classifier parameters in this study.

3.5.1 Logistic Regression (LR)

Due to its applicability in studies involving only two groups, logistic regression was selected. In logistic regression, you can use either the L1 or L2 parameter, but the L2 parameter is the one that is typically used because it gives equal weight to all of the feature vectors. L2 regression, also known as ridge regression, can be derived from the equation's regularization formula, which estimates the sum of squared errors and can display the restrictions.

L2(C)=w*=argminy$\sum_i$In[log(1+exp(-z))] + λ*$\sum(w_j)^2$ -------------------------------------- (2)

Where, $\sum(w_j)^2$ is a regularization term,

$\sum$ [log (1+exp (-z_i))] is the Loss term.

λ is a hyper parameter.

‘C’=coefficient of regularization is used a$\sum(w_j)^2$ is a regularization s a parameter

3.5.2 Random Forest Classifier

To perform tasks like classification, regression, and differentiation, random forests can be used as an ensemble method in place of individually trained decision trees. The problem of overfitting in decision trees can be solved by employing random forests. In random forests, individual trees are grouped together to form a collective wisdom.

3.5.3 Decision Tree (DT) Classifier

A decision tree uses a tree diagram to depict alternatives to a given decision and the results of that choice. The decision tree's root algorithm is an iterative dichotomy, which calculates the Entropy or Information Gain for each characteristic. Here, a decision tree's maximum depth, minimum leaf sample size, and maximum leaf node count are the tuning parameters.

3.5.4 Gradient Boosting Classifier

The goal of gradient boosting is for each successive predictor to improve upon its predecessor by making fewer mistakes. Instead of fitting a predictor to the data in each iteration, the novel idea behind Gradient Boosting is to fit a new predictor to the residual errors created by the prediction that came before it. For the sake of precision, we do this.

3.5.5 Support Vector Classifier (SVC)

Support Vector Classifier (SVC) is a popular supervised machine learning approach used to solve classification problems. SVC operates by mapping data points to a high-dimensional space and then identifying the optimal hyperplane for classifying the data.

The objective of a linear SVC (Support Vector Classifier) is to fit the input data and deliver the "best" subdividing or classifying hyperplane. After obtaining the hyperplane, you can next feed certain features into a classifier to determine the predicted class.

3.5.6 K-Nearest Neighbors (KNN) Classifier

The K-nearest neighbor algorithm is an example of a non-parametric lazy algorithm. Calculating the Euclidean distance between the x and y vectors that are given in the equation enables one to discover which entity is the closest neighbor. The output of KNN shifts about in response to changes in the value of K. A smaller value for K will result in fewer calculations, but a greater proportion of class overlap will occur if K is sufficiently large.

Euclidean Distance = $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ --(3)

3.5.7 AdaBoost Classifier

Ada-Boost, also known as Adaptive Boosting, was proposed by Yoav Freund and Robert Schapire in 1996 as one of several ensemble boosting classifiers. Combining different classifiers into one that performs better is one way to achieve this. In other words, AdaBoost is an iterative ensemble technique. The AdaBoost classifier combines several weak classifiers to create a single powerful one, resulting in a highly accurate strong classifier with a high degree of precision. Adaboost central idea is to accurately forecast out-of-the-ordinary observations by setting classifier weights and training data samples at each iteration. Classification can begin with any machine learning technique that uses weights from the training data.

3.5.8 Gaussian Naive Bayes (GaussianNB)

Gaussian Naive Bayes is a probabilistic classification technique that applies Bayes' theorem under the strong condition of independence. In the context of categorization, independence refers to the notion that the existence of one property value does not affect the existence of another property value (as opposed to independence in probability theory). Using the premise that object functions are independent of one another is naive. In the area of machine learning, it is known that naïve Bayesian classifiers are highly expressive, scalable, and reasonably accurate, but their performance falls fast as the size of the training set increases. The efficiency of naive Bayesian classifiers relies on a wide variety of factors. Also, it can easily deal with continuous data and scales effectively as the size of the training set increases.

3.5.9 Extreme Gradient Boosting (XGB) Classifier

As a machine learning technique, gradient boosting employs a collection of algorithms to address classification and regression modeling issues.

Models based on decision trees can be used to construct ensembles. Iterative model improvements are made by adding individual trees to the ensemble and modifying the ensemble to account for previous models' shortcomings. In the realm of machine learning, this is known as a boosting model.

The model is fit using a loss function that can be changed in any way and an optimization method called "gradient descent." This is like neural networks because when the model is fit, the loss gradient is minimized, which is where the name "gradient boosting" comes from.

3.6 Ensemble Learning

The goal of ensemble learning is to solve a given computer intelligence problem by systematically creating and merging multiple models, such as classifiers or experts. The results of machine learning are enhanced by ensemble learning since several models are combined. With this strategy, we can provide more precise forecasts than would be possible with a single model. The core idea is to single out a group of classifiers (experts) and give them a say in the matter.

3.6.1 Voting Classifier

To forecast the output (class) based on the likelihood that the chosen class is the most probable outcome, a voting classifier is an ensemble-trained machine learning model. The Voting Classifier does nothing more complicated than taking an average of the predictions made by the individual classifiers that were given to it. The goal is to train on these models and produce a single model that predicts output based on the total majority of votes for each output class, rather than constructing many models to evaluate the accuracy of each.

3.6.2 Bagging

Bagging, also recognized as Bootstrap aggregation, is an ensemble learning approach used to improve the accuracy and speed of machine learning algorithms. A balance between bias and variance in a prediction model is achieved by its utilization. Using bagging, which prevents overfitting of the data, improves the performance of decision tree algorithms for regression and classification problems.

3.6.3 Boosting

Boosting is an approach to ensemble modeling that uses multiple, less effective classifiers to create a single, more effective one. To do this, a model is built from a series of simplified models. At first, a model is built with the help of the training data. After identifying the problems with the first model, the second model is built to correct them. This process is repeated until all of the training data set can be predicted with high accuracy, or till the maximum number of models is reached.

3.7 Cross Validation

3.7.1 K-10 Fold Cross Validation

Cross-validation after model development is crucial. We utilized stratified K-fold cross-validation as a result. Each time, the dataset was divided into 10 equal folds. The model was created using nine convolutions and evaluated using one. Therefore, the accuracy of the model was compared to its mean after 10 iterations. A stratified K-fold ensures that the attribute classes of interest are uniformly distributed throughout all folds. We have tried several methods. We delve as deeply as possible to achieve our objectives.

One way to test how well a machine learning model can make predictions is through cross-validation. Issues like overfitting and selection bias can be uncovered, and information about the model's ability to generalize to new data can be gleaned.

3.7.2 Stratified K-Fold Cross Validation

Compared to traditional k-fold cross validation, stratified k-fold cross validation adds additional features. However, the ratio between the target classes is kept at the same level in each fold as it is in the whole dataset, so the splits are not fully random.

3.8 Hyper Parameter Tuning

The tweaking of hyperparameters is an essential component of managing the behavior of machine learning models. If you do not set the hyperparameters properly, the estimated model parameters will produce suboptimal results because the loss function will not be minimized. This means our model will make more errors. In practice, important metrics such as accuracy and confusion matrix get worse.

3.9 Performance Matrix Evaluation

3.9.1 Confusion Matrix

A confusion matrix is a very common way to figure out how to classify something. This works for both problems with two classes and problems with more than two classes. In Table 3.9.1, you can see an example of a confusion matrix for a two-way classification.

Table 3.9.1. Confusion matrix for binary classification.

		Predicted	
Actual		**Negative**	**Positive**
	Negative	**TN**	**FP**
	Positive	**FN**	**TP**

The quantity of times the actual values were tallied versus the expected values is displayed in the confusion matrix. "True Negatives," or "TN," output indicates the total number of samples that were correctly identified as negative. Indicates the proportion of true positives that were identified. The letters "FP" denote a false positive result. H. The percentage of false-negative cases that turned out to be correct. A "false negative" (or "FN") is the number of true positives that were incorrectly classified as false negatives.

Accuracy

One frequent metric used to evaluate the efficacy of a categorization is its accuracy. The following formula can be used to calculate the model's accuracy using the confusion matrix:

$$\text{Accuracy}=\frac{\text{TN+TP}}{\text{TN+FP+FN+TP}} \text{ -- } (4)$$

Accuracy can be deceiving when applied to unbalanced datasets; hence, there are additional metrics based on confusion matrices for evaluating performance. Using the 'confusion matrix ()' function in Python, the confusion matrix can be obtained. Python users can include this feature into their code by typing from sklearn metrics import exposed matrix. The operator of the function must input both the actual and expected values so as to generate the confusion matrix.

Precision

It refers to the proportion of positive observations that were accurately predicted. Precision identifies the actual true portion of the total number of occasions in which the prediction was accurate.

Precision=$\frac{TP}{(TP+FP)}$ --(5)

Recall

It represents the proportion of accurately predicted positive observations.

Recall=$\frac{TP}{(TP+FN)}$ -- (6)

F1-score

In its most basic form, it is the harmonic average of the recall and the precision.

F1 Score=$\frac{2(\text{Precision} * \text{Recall})}{(\text{Precision}+\text{Recall})}$ --- (7)

3.9.2 AUC Score & ROC CURVE

AUC Score

Area below the curve (AUC) is used to measure performance based on many criteria. This measures how far apart the results are. It also shows how the model classifiers the data.

ROC curve

A graphical representation of the overall classification level of the classification model by Receiver Operating Characteristic (ROC) curve. This curve shows two variables. Really positive and false positive rate.

CHAPTER 4

Experimental Result and Analysis

Experimental result and analysis

For this study we used ML clarification methods to build a model that could classify the correct stages of breast cancer. Table 5.1.1 shows the models' classification performance. In this study we used 9 classifier algorithms. In the Table 5.1.1. along with accuracy we displayed Model Classifier's F1-Score, precision and recall. Among them we got the best result from the Logistic Regression classifier with an accuracy rate of 98.245%, F1-score of 0.98, precision of 0.99 and Recall rate of 0.98.

Table 4.1.1 Comparison of Accuracy, F1-score, Precision, Recall

Serial	Model Classifier	Accuracy	F1-score	Precision	Recall
1	Logistic Regression	98.245	0.98	0.99	0.98
2	Random Forest	93.859	0.94	0.94	0.94
3	Decision Tree	87.719	0.88	0.88	0.88
4	Gradient Boosting	96.491	0.96	0.97	0.96
5	Support Vector Classifier	92.982	0.93	0.93	0.93
6	k-Nearest Neighbors	96.491	0.96	0.96	0.96
7	Adaboost Classifier	97.368	0.97	0.97	0.97
8	Gaussian NB	95.614	0.96	0.96	0.96
9	XGB Classifier	95.614	0.96	0.96	0.96

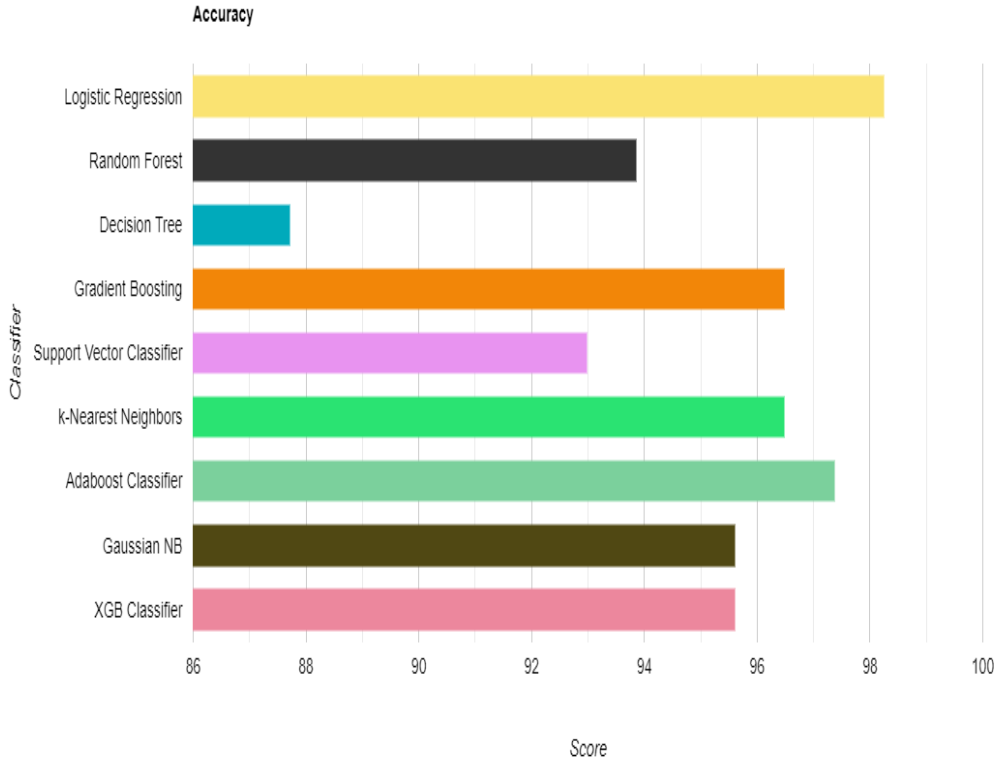


Figure 4.1.1 Accuracy Diagram

In Figure 4.1.1 is our accuracy diagram. Here each bar represents a certain classifier. The bars were scored up to 100. Among them Logistic Regression had the highest bar which shows highest score of 98.245%.

Table 4.1.2 Displays some of the other important numbers regarding performance measurements of the classifiers. Some of the information’s are Testing Accuracy, sensitivity, Specificity, False Positive Rate, False Negative, Negative Predictive Value, False Discovery rate, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Log Loss, Cohen Kappa Scorer.

Table 4.1.2 Performance Measurements of the ML Algorithms

SN	Classifier	Testing Accuracy	Sensitivity	Specificity	False Positive Rate	False Negative	Negative Predictive Value	False Discovery rate	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Log Loss	Cohen Kappa Scorer
1	Logistic Regression	98.245	1.0	0.972	0.027	0.0	1.0	0.046	0.017	0.017	0.132	13.028	0.962
2	Random Forest	93.859	0.891	0.970	0.029	0.108	0.929	0.046	0.061	0.061	0.247	13.028	0.871
3	Decision Tree	87.719	0.837	0.901	0.098	0.162	0.901	0.162	0.122	0.122	0.350	13.028	0.738
4	Gradient Boosting	96.491	0.975	0.958	0.041	0.024	0.985	0.069	0.035	0.035	0.187	13.028	0.924
5	Support Vector Classifier	92.982	0.926	0.931	0.068	0.073	0.957	0.116	0.070	0.070	0.264	13.028	0.849
6	k-Nearest Neighbors	96.491	0.953	0.971	0.028	0.046	0.971	0.046	0.035	0.035	0.187	13.028	0.925
7	Adaboost Classifier	97.368	0.976	0.972	0.027	0.023	0.985	0.046	0.026	0.026	0.162	13.028	0.943
8	Gaussian NB	95.614	0.975	0.945	0.054	0.025	0.985	0.093	0.043	0.043	0.209	13.028	0.905
9	XGBClassifier	95.6140	0.975	0.945	0.054	0.025	0.985	0.093	0.043	0.043	0.209	13.028	0.905

Table 4.1.3. AUC and Cross Validation Score

Serial	Classifier	AUC Score	Cross Validation
1	Logistic Regression	0.997	0.982
2	Random Forest	0.971	0.956
3	Decision Tree	0.869	0.885
4	Gradient Boosting	0.995	0.947
5	Support Vector Classifier	0.993	0.982
6	k-Nearest Neighbors	0.990	0.947
7	Adaboost Classifier	0.993	0.947
8	Gaussian NB	0.995	0.956
9	XGB	0.987	0.964

Table 4.1.3 represents AUC and Cross Validation Scores of each classifier. AUC measures how far apart the results are. Cross-validation is used to evaluate a machine learning model's ability to predict new data. For Logistic Regression AUC score is 0.997 and Cross Validation score is 0.982 which is the highest. The lowest number is for Decision Tree with AUC score of 0.869 and Cross Validation score of 0.885.

Figure 4.1.2 and 5.1.3 represent ROC and AUC curve diagrams for each classifier. ROC curve is a graphical representation of the overall classification level of the classification model. AUC shows the model classifiers the data.

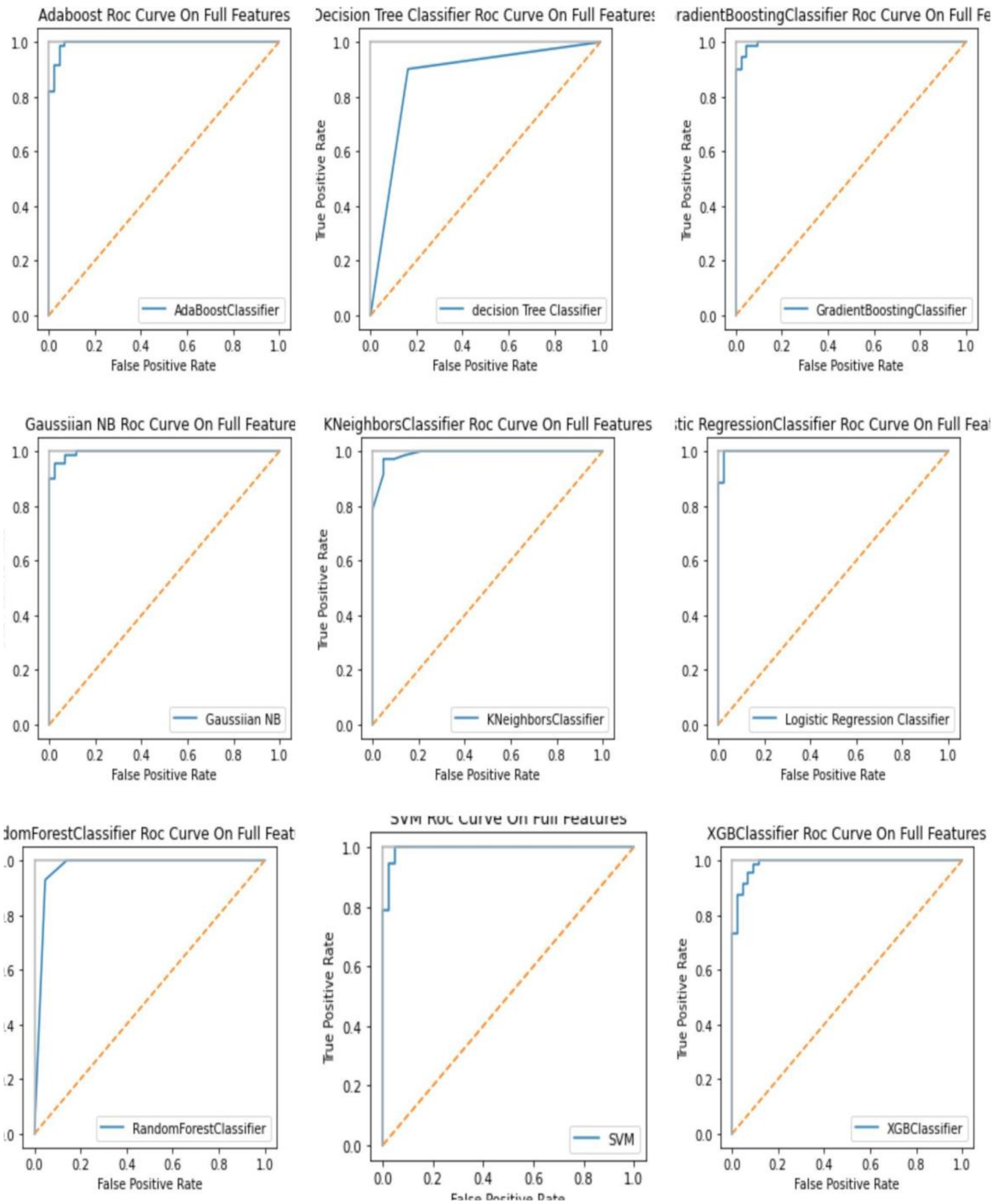


Figure 4.1.2 ROC Curve Diagram

Figure 4.1.2 represents the ROC curve graph. Each graph performs as a virtual representative for all nine classifiers. The vertical axis represents True Positive rate and the horizontal axis displays the False Positive rate.

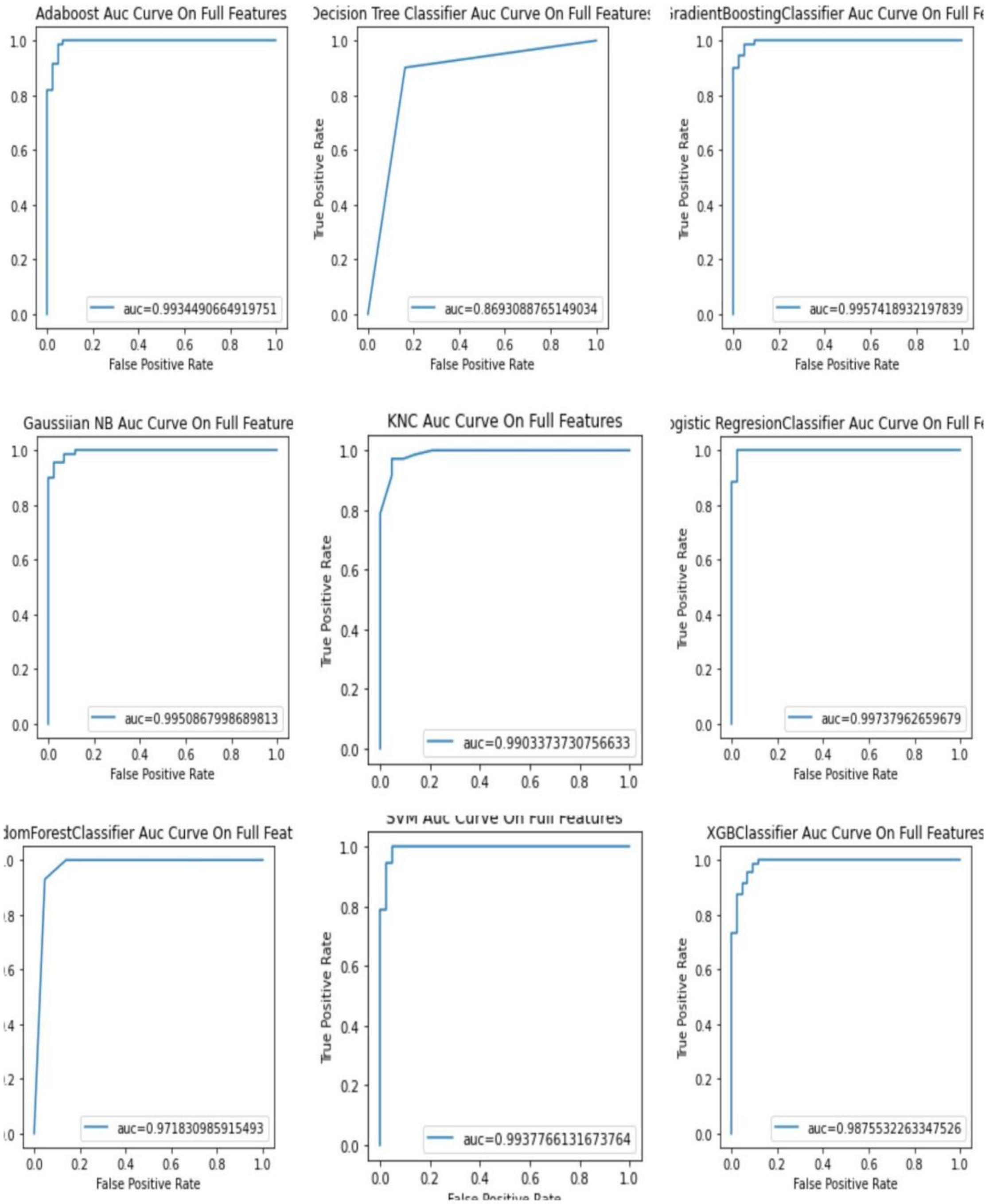


Figure 4.1.3 AUC Curve Diagram

Similarly, The AUC curve is displayed in Figure 4.1.3. Each graph acts as a proxy for the other eight classifiers. False positives are shown on the horizontal axis and True positives are shown vertical.

Table 4.1.4 represents Hyper Parameter Tuning's best parameter and score. Here for each classifier, we used several parameters. Using the parameters, we achieved the best score.

Table 4.1.4 Hyper Parameter Tuning (Best parameter & Score)

Serial Number	Model Classifier	Used Parameter	Best Parameter	Hyper Parameter Score
01	Logistic Regression	{'class_weight' : ['dict','balanced'] }	'class_weight': 'balanced'	0.913
02	Random Forest	{ 'bootstrap': [True], 'max_features': ['auto', 'log2'], 'criterion' : ['entropy','gini'], 'n_estimators': [1,2,4,6,8,5,10,15,25,30,50] }	bootstrap': True,	0.909
03	Decision Tree	{'criterion' : ['gini', 'entropy', 'log_loss'] }	'criterion': 'entropy'	0.719

04	Gradient Boosting	{
		"n_estimators":[30,40,50,60],
		"max_features":[1,2,5,9],
		"learning_rate":[0.1,1.5,1.75]
		}
			learning_rate': 0.1	0.929

05	Support Vector Classifier	{'gamma' : ['scale','auto'] }	'gamma': 'scale'	0.946
06	k-Nearest Neighbors	{ 'n_neighbors' : [7,9,11,13,15,10,20], 'weights' : ['uniform','distance'], 'metric' : ['minkowski','euclidean','manhattan'] }	metric': 'minkowski'	0.929
07	Adaboost Classifier	{'n_estimators' : [50,70,75,40,100], "learning_rate":[1,1.25,1.45,1.75], 'algorithm' : ['SAMME','SAMME.R'] }	'algorithm': 'SAMME'	0.765
08	Gaussian NB	{'var_smoothing' : [10, 15, 20] }	'var_**smoothing**': 10	0.848
09	XGB	{ 'booster' : ['gbtree','linear'], 'max_depth' : [1,2,3,4,5,7], 'n_estimators' : [5,10,20,30,50], "learning_rate":[0.1,1.5,1.75]	'booster': 'gblinear'	0.909

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Whenever it comes to cancers affecting humans, breast cancer is both the most commonly diagnosed and the deadliest. Most people in Bangladesh have never heard of breast cancer. Unfortunately, this may happen again when doctors miss diagnosing a patient's condition and they end up without the care they need. This contributes to the premature deaths of many patients. With the use of machine learning techniques, we were able to develop a model that correctly identified the cancer stage of 98.25% of patients. Our approach will aid doctors in determining the extent of cancer and the likelihood of survival. By doing this, we can perhaps assist lower the mortality rate among those with breast cancer.

5.2 Impact on Environment

As things stand now, it’s clear that when a person is sick from a sickness, they have to go through more pain than necessary before their illness is diagnosed and therapy begins. They have to go to the doctor’s office and wait in line for quite some time before they can see the doctor. Time wasted in this manner is of no use to the patient. Patients don’t always get the care they need in time, and that can be fatal. However, if patients employ an AI-based system to diagnose their illness, they will have a far better chance of making a full recovery. Once again, this aids in the quick diagnosis of ailments by doctors so that patients may receive timely advice and treatment. Once again, this technique helps keep diagnostic costs down. Huge volumes of trash are created daily in the pathology lab. Using this technique, pathology labs may cut down on their trash. They can go back to their treatment faster. It is clear that our country’s public healthcare facilities, including hospitals and physicians’ offices, are overburdened. Patients here report a heightened sensitivity to sickness because to the close quarters. That’s bad for the patients’ health. Government hospitals are notorious for their filth, which may make patients feel much worse. A machine learning-based system like this would allow for at-home illness testing.

5.3 Ethical Aspects

When conducting our study, we need to be truthful. In order to improve their study outcomes, some academics have been caught fabricating data sets or manipulating existing ones. It goes against every principle of ethics. Because in the field of medicine, this sort of study has the potential to endanger the lives of people. Patients risk their lives by receiving substandard care. We conducted our research using information obtained from the data repository at UCI. The data provided by UCI are considered to be among the most reliable in the world. In order to lessen the burden that cancer places on the country as a whole, the UCI data repository compiles cancer statistics. In order to accurately forecast the stage of cancer, we analyzed data from 31 attributes. When we were building our models, we used a substantial quantity of data in order to ensure that they were as precise as possible in their predictions.

5.4 Sustainability Plan

Whenever we desire to get something done, we need to consider whether or not it can be accomplished in a way that is long-lasting. During the course of our study, we endeavored to develop methods based on machine learning that are capable of classifying the cancer stages of breast cancer survivors. Research along these lines has the potential to have a profound influence on people. Because of this, we gave some consideration to the long-term strategy for our study. The signs and symptoms of any disease seem to be evolving on a daily basis. This also applies to breast cancer in its many forms. Because of this, we have devised a strategy for dealing with this issue. We plan to use reinforcement learning in our model in the not-too-distant future. This will train our model with new sorts of data, as well as acquire new types of data, and it will continually deploy new models. In the course of this procedure, there won't be any complications even if the breast cancer symptoms shift.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary

Some of the highest rates of early death in the world are attributable to breast cancer, which is especially devastating in underdeveloped regions. In the study our goal was to classify the different stages of breast cancer. In order to do that we used some machine learning classifiers. After using "Hyperparameter Tuning" to increase model accuracy and reduce model runtime we evaluated the performance and tested the ability of our machine learning model using cross validation, we found out that the "Logistic Regression methods" provide the highest accuracy at 98.25%.

6.2 Conclusion

Among humans, breast cancer is by far the most frequent form of the disease. One of the most difficult challenges in the field of medical informatics is the classification of medical data. Nonetheless, it is widely recognized as a classic among scholarly literature. Consequently, many solutions have been presented by diverse groups. And yet, there are some enhancements. Ten machine learning models were applied to the Wisconsin Breast Cancer Dataset, and their performance as a function of their hyperparameters was documented. Hyperparameter tuning ensures that the best parameters are used, allowing the models to outperform their predecessors. Our proper analysis of the dataset has been completed. So far, several performances using a wide range of techniques have stood out. We provided evidence that our model is superior to existing alternatives. If implemented, our method would revolutionize how breast cancer is categorized for medical research. Finally, our staging approach would aid breast cancer patients in understanding the various phases of their disease and receiving the most effective treatment possible. This has the potential to save many lives and allow people to live healthy, happy lives while exploring our beautiful world.

6.3 Future Work

This section outlines future research avenues that can be applied to the classification of breast cancer, and major efforts are required to increase the classification's efficacy. In spite of the positive outcomes of the examined literature, there are still substantial limitations and hurdles associated with Machine Learning (ML) approaches for breast cancer detection and classification. Future research could concentrate on the development of more effective machine learning algorithms and the application of unsupervised machine learning to samples that are unknown. Additionally, we would like to use Deep Learning (DL) for breast cancer classification. There are also more scopes to consider, such as a larger data set and more preprocessed datasets.

Reference:

[1] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithms for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28.

[2] Derangula, A., Edara, S., & Karri, P. K. (2020). Feature selection of breast cancer data using gradient boosting techniques of machine learning. *European Journal of Molecular & Clinical Medicine*, *7*(2), 3488-3504.

[3] Alzubier, Muatz Humida, and V. Chitraa. "Classification of Breast cancer Using Machine Learning Techniques." *International Journal of Research and Analytical* 58 (2019).

[4] Sakr, M., Saber, A., M Abo-Seida, O., & Keshk, A. (2020). Machine learning for breast cancer classification using the k-star algorithm. *Applied Mathematics & Information Sciences*, *14*(5), 855-863.

[5] Abdulla, S.H., Sagheer, A.M. and Veisi, H., 2021. Breast Cancer Classification Using Machine Learning Techniques: A Review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(14), pp.1970-1979.

[6] Bhardwaj, Arpit, Harshit Bhardwaj, Aditi Sakalle, Ziya Uddin, Maneesha Sakalle, and Wubshet Ibrahim. "Tree-Based and Machine Learning Algorithm Analysis for Breast Cancer Classification." *Computational Intelligence and Neuroscience* 2022 (2022).

[7] Chang, Chi-Chang, and Ssu-Han Chen. "Developing a novel machine learning-based classification scheme for predicting SPCs in breast cancer survivors." *Frontiers in Genetics* 10 (2019): 848.

[8] Ting, W.C., Chang, H.R., Chang, C.C. and Lu, C.J., 2020. Developing a novel machine learning-based classification scheme for predicting SPCs in colorectal cancer survivors. *Applied Sciences*, *10*(4), p.1355.

[9] T. Chandrasekaran S, Hua R, Banerjee I, Sanyal A. A fully-integrated analog machine learning classifier for breast cancer classification. Electronics. 2020 Mar 20;9(3):515.

[10] Gupta, P., & Garg, S. (2020). Breast cancer prediction using varying parameters of machine learning models. *Procedia Computer Science*, *171*, 593-601.

[11] Akbugday, Burak. "Classification of breast cancer data using machine learning algorithms." In *2019 Medical technologies congress (TIPTEKNO)*, pp. 1-4. IEEE, 2019.

[12] Idri, A., Hosni, M., Abnane, I., Carrillo de Gea, J. M., & Fernández Alemán, J. L. (2019, April). Impact of parameter tuning on machine learning based breast cancer classification. In *World Conference on Information Systems and Technologies* (pp. 115-125). Springer, Cham.

[13] Li, Yuqian, Junmin Wu, and Qisong Wu. "Classification of breast cancer histology images using multi-size and discriminative patches based on deep learning." *IEEE Access* 7 (2019): 21400-21408.

[14] Egwom, Onyinyechi Jessica, Mohammed Hassan, Jesse Jeremiah Tanimu, Mohammed Hamada, and Oko Michael Ogar. "An LDA–SVM Machine Learning Model for Breast Cancer Classification." *BioMedInformatics* 2, no. 3 (2022): 345-358.

[15] Hazra, R., Banerjee, M., & Badia, L. (2020, November). Machine learning for breast cancer classification with ann and decision tree. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0522-0527). IEEE.

[16] Mashudi, Nurul Amirah, Syaidathul Amaleena Rossli, Norulhusna Ahmad, and Norliza Mohd Noor. "Comparison on Some Machine Learning Techniques in Breast Cancer Classification." In *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pp. 499-504. IEEE, 2021.

[17] Laghmati S, Cherradi B, Tmiri A, Daanouni O, Hamida S. Classification of patients with breast cancer using neighbourhood component analysis and supervised machine learning techniques. In2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet) 2020 Sep 4 (pp. 1-6). IEEE.

[18] Zhang, Lihao, Chengjian Li, Di Peng, Xiaofei Yi, Shuai He, Fengxiang Liu, Xiangtai Zheng, Wei E. Huang, Liang Zhao, and Xia Huang. "Raman spectroscopy and machine learning for the classification of breast cancers." *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 264 (2022): 120300.

[19] Huang, Wenhua, Qixin Shang, Xin Xiao, Hanlu Zhang, Yimin Gu, Lin Yang, Guidong Shi et al. "Raman spectroscopy and machine learning for the classification of esophageal squamous carcinoma." *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 281 (2022): 121654.

[20] Gopal, V. N., Al-Turjman, F., Kumar, R., Anand, L., & Rajesh, M. (2021). Feature selection and classification in breast cancer prediction using IoT and machine learning. *Measurement*, *178*, 109442.

[21] Mohammed, Siham A., Sadeq Darrab, Salah A. Noaman, and Gunter Saake. "Analysis of breast cancer detection using different machine learning techniques." In *International Conference on Data Mining and Big Data*, pp. 108-117. Springer, Singapore, 2020.

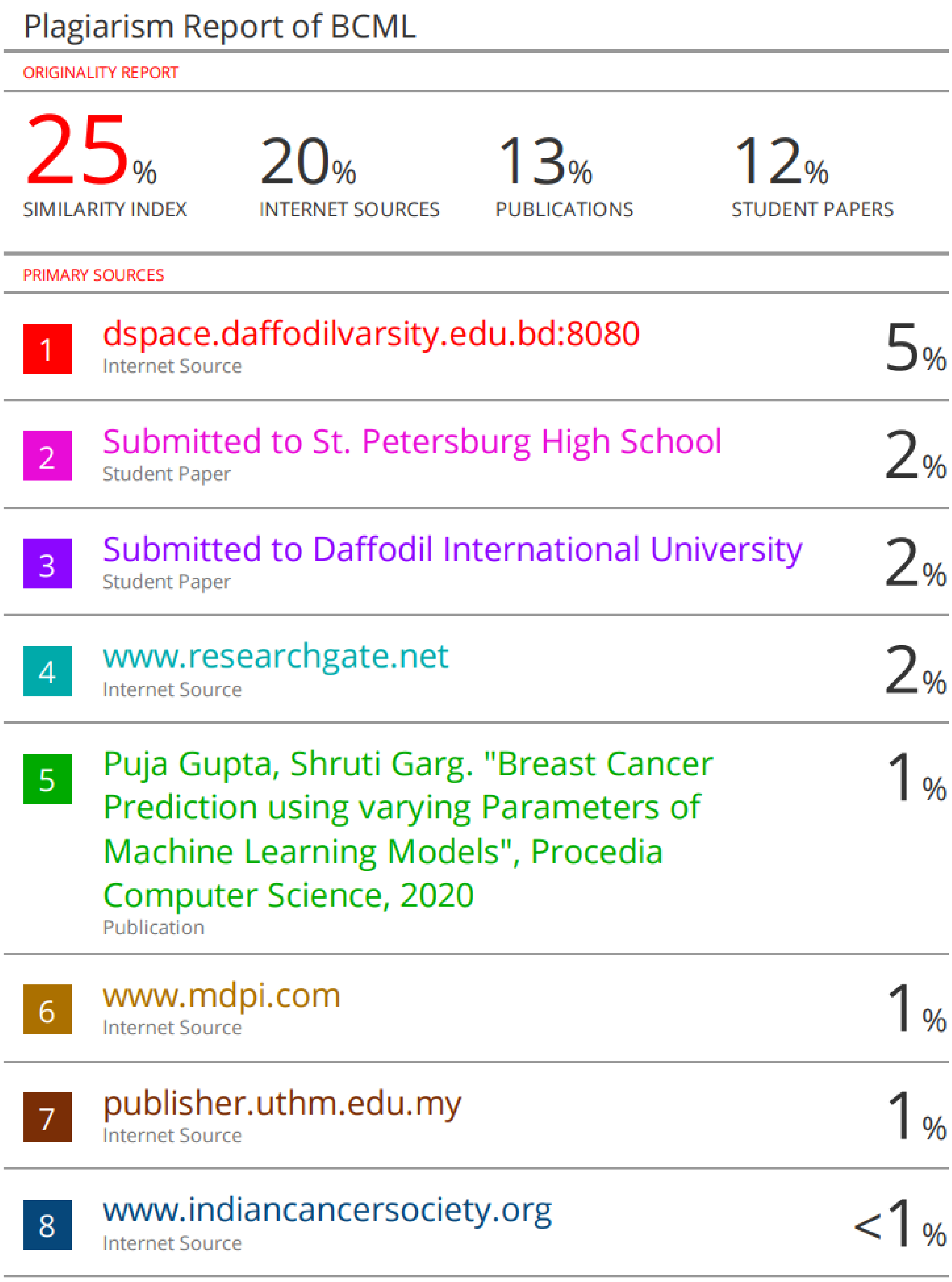
[22] Houssein, Essam H., Marwa M. Emam, Abdelmgeid A. Ali, and Ponnuthurai Nagaratnam Suganthan. "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review." *Expert Systems with Applications* 167 (2021): 114161.

[23] Omondiagbe, D.A., Veeramani, S. and Sidhu, A.S., 2019, April. Machine learning classification techniques for breast cancer diagnosis. In *IOP Conference Series: Materials Science and Engineering* (Vol. 495, No. 1, p. 012033). IOP Publishing.

[24] Michael, Epimack, He Ma, Hong Li, and Shouliang Qi. "An optimized framework for breast cancer classification using machine learning." *BioMed Research International* 2022 (2022).

[25] Rane, Nikita, Jean Sunny, Rucha Kanade, and Sulochana Devi. "Breast cancer classification and prediction using machine learning." *International Journal of Engineering Research and Technology* 9, no. 2 (2020): 576-580.

[26] Sakr, Mohamed, et al. "Machine learning for breast cancer classification using k-star algorithm." *Applied Mathematics & Information Sciences* 14.5 (2020): 855-863.



Plagiarism Report of BCML

ORIGINALITY REPORT

25%	20%	13%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to St. Petersburg High School Student Paper	2%
3	Submitted to Daffodil International University Student Paper	2%
4	www.researchgate.net Internet Source	2%
5	Puja Gupta, Shruti Garg. "Breast Cancer Prediction using varying Parameters of Machine Learning Models", Procedia Computer Science, 2020 Publication	1%
6	www.mdpi.com Internet Source	1%
7	publisher.uthm.edu.my Internet Source	1%
8	www.indiancancersociety.org Internet Source	<1%