UNIVERSITY of RWANDA

COLLEGE OF SCIENCE AND TECHNOLOGY

Regional Centre of Excellence in Biomedical Engineering and e-Health (CEBE)

Ensemble classifier for early detection of breast cancer

By:

ISHIMWE BEZA Aime

Reference Number: 214000872

A Dissertation Submitted to the Regional Centre of Excellence in Biomedical Engineering and e-Health (CEBE), University of Rwanda as partial fulfilment of the requirements for the Master's Degree in Biomedical Engineering.

Supervised by:

Dr. James RWIGEMA

Prof. Lee Hyowon Hugh

February, 2023

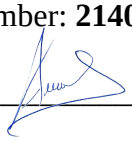
DECLARATION

I, **Aime Ishimwe Beza**, declare that this dissertation entitled **"Ensemble Classifier for Early Detection of Breast Cancer"** is my original work based on research and prototype and has not been submitted for any other degree or professional qualification.

Student Name:

Aime Ishimwe Beza

Student Reference Number: **214000872**

Student Signature: 

Date: 2023-02-08

UNIVERSITY of RWANDA

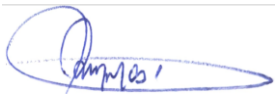
COLLEGE OF SCIENCE AND TECHNOLOGY

Regional Centre of Excellence in Biomedical Engineering and e-Health (CEBE)

Certification

This is to certify that the project entitled "**Ensemble Classifier for Early Detection of Breast Cancer**" is a record of original work done by Aime Ishimwe Beza (Reference number: 214000872), an MSc. Degree student in Biomedical Engineering. This work has been submitted under the guidance of **Dr. James Rwigema** and **Prof. Lee Hyowon Hugh.**

Supervisor:



Dr. James Rwigema

Supervisor:

Prof. Lee Hyowon Hugh

ACKNOWLEDGMENTS

By the name of Jesus most merciful and gracious, I would like to take this opportunity to express my deepest gratitude to my project supervisors, Dr. James Rwigema, and Prof. Lee, Hyowon Hugh. My greatest thanks also go to my family, especially my mum to whom I owe my love and gratitude. I also would like to express my gratitude to CEBE for providing me with all I needed in my project implementation. To all my friends whose names are not mentioned here that have helped and supported me along by the way, I thank you from the bottom of my heart. I wish you all the best in life and hope that our friendship will last forever.

Abstract

Breast Cancer is one of the global illnesses which is not easy to diagnose at an early stage. This study aimed at increasing the patient's awareness and this was achieved under the use of base learners such as Decision Tree Classifier (DT), Naive Bayes (NB), Support Vector Machine (SVM), K Nearest Neighbors (KNN), and ensemble classification algorithms namely Ada-Boost (AB), and Random Forest (RF). We trained and tested those machine learning algorithms on breast cancer datasets created from mammogram images from the King Faisal Hospital (KFH), to classify the patients who are living with breast cancer, and those who are not. With the use of the pre-mentioned dataset, which recorded the past patient's clinical information, the classification algorithms were evaluated by a confusion matrix which led us to classification metrics namely recall, precision, and f1 score. The ensemble machine learning algorithms namely AB, and RF performed well with optimal accuracy of 97%, and 92% respectively as expected compared to the base learners. The python built-in libraries were used on all models during implementation.

List of Abbreviation

AI; Artificial Intelligence

ML; Machine Learning

DT; Decision Tree

NB; Naive Bayes

KNN; K-Nearest Neighbours

SVM; Support Vector Machine

RF; Random Forest

AB; Ada-boost

TP; True Positive

FP; False Positive

FN; False Negative

TN; True Negative

$WBCD$; Wisconsin breast cancer

$CEBE$; Center of Excellence in Biomedical Engineering

KFH; King Faisal Hospital

ML; Machine Learning

$UGHE$; University Global Health Equity

Contents

List of Figures

List of Tables

1. INTRODUCTION

1.1 BACKGROUND

Breast cancer is the most common type of invasive cancer in women all over the world. According to statistics published by the International Agency for Research on Cancer (IARC) in December 2020, breast cancer has now surpassed lung cancer as the most commonly diagnosed cancer in women globally. Over the last two decades, the total number of women diagnosed with breast cancer has more than doubled, growing from an estimated 10 million in 2000 to 19.3 million in 2020, (1).

Nowadays, One out of every five people will be diagnosed with cancer at some point in their lives,(2). According to projections, the number of persons diagnosed with cancer would rise even more in the following years, reaching approximately 50% higher in 2040 than in 2020. The deaths due to cancers have also escalated, with 10 million individuals dying from the disease in 2020 compared to 6.2 million in 2000. The rapid growth in cancer recognition and finding is now being accelerated by the world's improved technology, medicine, and health systems,(3).

In addition, a variety of cancer prevention, detection, and diagnosis strategies have been used in order to reduce the number of new cancer cases in the community. Artificial intelligence (AI) and machine learning (ML) are among the approaches that have been used because they help greatly and unexpectedly in cancer diagnosis and prognosis, much beyond human expectations,(4). The huge amount of data transformed the size of the data as well as the creation of value from it. By evaluating vast amounts of complex medical data, big data has revolutionized business intelligence. It not only anticipates but also aids in taking decisions, and it is instantly being acknowledged as a step forward in continuous betterment with the aim of improving patient care quality while decreasing healthcare expenditures,(5).

Data mining algorithms used in the medical field diagnosis are important because of their great performance in predicting and diagnosing diseases, lowering medical expenditures, and making real-time decisions to stop people from dying,(6). This study primarily classifies, detects, and predicts. breast cancer treatment using ML-based ensemble approaches such as Ada-boost, and Random Forest. The various machine learning techniques precisely supervised machine learning models namely Decision tree (DT), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Support Vector Machine (SVM) were

compared to those ensemble machine learning models.

1.2 PROBLEM STATEMENT

Both men and women are at risk of acquiring a breast tumor. However, it affects a hundred times more women than men. Breast cancer is among the furthermost serious and common malignancies in women's reproductive systems,(7). A breast tumor is an abnormal growth of tissues in the breast that can manifest as a lump, nipple discharge, or a modification in skin texture around the nipple area. Cancers are uncontrolled cell divisions that have the ability to infect other tissues. Cancer cells can get into the blood and lymphatic systems in different regions of the human body. In the last 50 years, it has become a major public health concern, and its prevalence has risen in recent years. The image below depicts breast cancer worldwide among females.

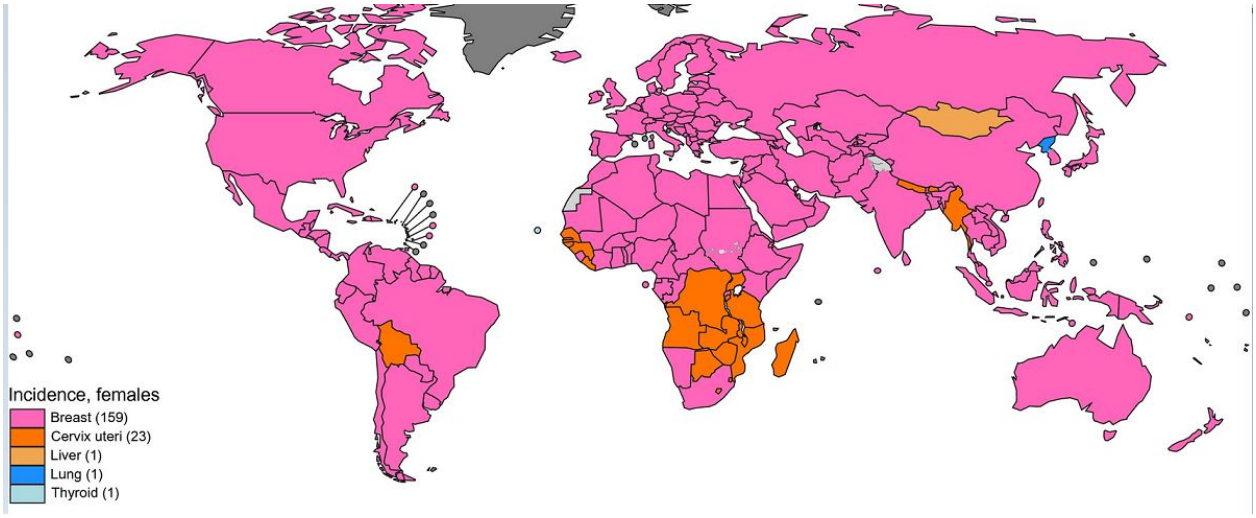


Figure 1.1: Type of cancer incidence in 2020 in each country among women (1).

According to the National Cancer Institute, a tumor is an irregular mass of tissue that grows when cells split faster than they should or do not die when they should,(8). Cells grow, divide, and replace each other in a healthy human body. Old cells die as new one's form. When a person has cancer, new cells form even if the body does not require them,(9). When there are too numerous cells, a tumor might form. In the human body, there are non-cancerous cells that are benign and cancerous cells that are malignant. Malignant tumors are malignant and likely to spread to other regions of the body. Additionally, when cells replicate too quickly, a tumor develops. Tumors can range in dimensions from

a little nodule to a massive mass, and they can appear practically anywhere on the body, depending on the type. There are three main forms of tumors, according to the National Cancer Institute:

Benign, these cells aren't malignant. The majority of benign tumors are harmless. They either don't spread or grow at all, or they do so slowly. They don't usually reappear after being eliminated from the body. However, if they push against nerves or blood arteries, or if they activate the overproduction of hormones, as in the endocrine system, they can cause pain or other issues. Premalignant, the cells in these tumors aren't currently cancerous, but there is a possibility to develop cancer,(10).

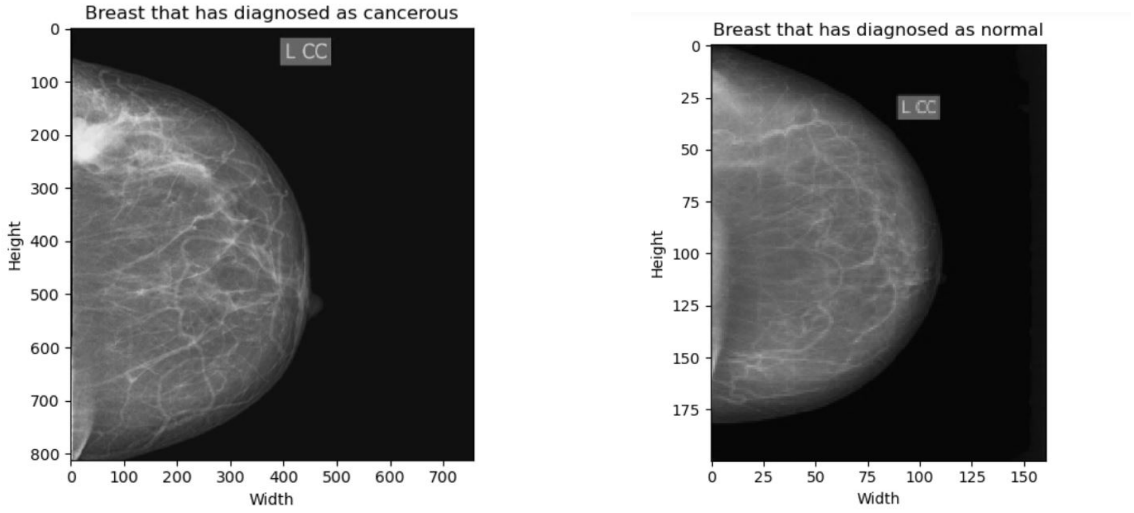
Malignant, Tumors that are cancerous are known as malignant tumors. Through the process of metastasis, cells can develop and spread to different areas of the body,(11). When cells proliferate out of control, they become tumors. The condition will become life-threatening if it continues to grow and spread. Surprisingly, cancer cells that spread to other parts of the body are comparable to those that originated cancer in the first place. They also have the ability to infiltrate other organs in the body as a whole. If breast cancer spreads to the lungs, malignant cells that emerge in the liver are still breast cancer cells. It's not always easy to predict how a tumor will behave in the future. Some benign tumors can progress from benign to premalignant to malignant. As a result, it is preferable to keep track of any growth by visiting a specialist physician or doctor on a regular basis for health checks.

1.2.1 Case Study: Breast Cancer in Rwanda. In Rwanda, a number of studies on breast cancer have been done. Breast cancer diagnoses and fatalities are rising in low- and middle-income nations. Over 50% of breast cancers are discovered at an advanced stage, and this is a major factor in the high breast cancer mortality rates in low-income nations in particular. Finding the best methods to reduce breast cancer in low- and middle-income nations is a serious issue worldwide. In low-income countries, has some limitations between knowing the onset of breast signs and the early diagnosis of breast cancer,(12).

Although the patients get the treatment at a later stage, it appears that an inadequate and unsuitable healthcare system is a factor in this issue. The inclusion of patients with breast cancer symptoms in early detection efforts is a vital step in building early detection capacity in low-income nations,(13; 14). According to the study (14), around 75% of breast cancer patients in Rwanda are suffering from stage III or stage IV disease when they are first diagnosed and are incurable. Although Rwanda is still dealing with the breast cancer problem, a lot has been done to raise awareness of breast cancer. Rwanda

has implemented a breast cancer early detection program $(BCED)$ and tested it in a random way at a clinical center in the Burera district in the northern province where the Butaro Cancer Centre of Excellence $(BCCOE)$ is located, and it is the one in charge of cancer.

The program is carrying out two distinct initiatives to raise community awareness of breast cancer and train key personnel in breast examination and breast ultrasound. The program makes a significant contribution to the improvement of community health workers and nurses' skills in dealing with patients with breast cancer-related concerns. At this point, Rwanda has reached the benchmark for implementing $BCED$ and moving it up a level so that it can be deployed in other districts. However, due to the cost of $BCED$, it might be difficult to implement the initiative in Rwanda and other low- and middle-income nations. In this study, we made use of information gathered from King Faisal Hospital, and we were able to obtain data on patients who had been given benign or normal diagnoses, as depicted in fig.1.2 below, respectively.



(a) Patient was diagnosed as benign (b) patient was diagnosed as normal

Figure 1.2: Breast Patient examination

.

Concerned with the virulence of breast cancer, we endeavor to find out a systematic way of early finding breast cancer using various machine learning techniques. Different investigations have been conducted on breast cancer with the use of different machine-learning techniques. But few of them didn't use the Rwanda breast cancer data set. Since, our main focus is to make a classification in the breast cancer clinical dataset collected from one Rwandan hospital so that we can predict the patient who has breast

cancer and doing so, avoid patients 'death and more importantly the spread of cancer, the recall score turns out to be of higher importance.

1.3 Research Hypotheses

1. What causes the increase in breast cancer?
2. Which characteristics are most important in predicting breast cancer?
3. Which classification algorithm is best for developing a model of breast cancer on the Rwandan dataset?
4. How well the suggested model perform in predicting breast cancer on the Rwandan dataset?

1.4 AIMS AND OBJECTIVES

1.4.1 Main objectives. The main objective of this project is to build on the breast patients' datasets (mammogram images) collected from Rwandan hospitals more precisely King Faisal hospital a high-performance machine learning ensemble model for early prediction of breast cancer.

1.4.2 Specific Objectives. This project has the following specific objectives, toward the fulfillment of the main objective:

- understanding the essence of ML algorithms such as DT, NB,KNN, SVM
- compares various ML algorithms using different evaluation metrics and against the python built-in ensemble models such as AdaBoost (AD), and Random Forest (RF).
- evaluating models using confusion matrix as is one of the machine learning evaluation metrics.

1.5 Study Scope

Detecting breast cancer in patients' breasts has its own set of difficulties. It is often recommended to women aged 40 and above go for a mammography screening for early detection,(15). Diagnosis of this cancer is made by self-examination and mammography. The fastest and most effective examination is mammography, which is used to look for abnormalities of the mammary gland. A mammogram is an

X-ray picture of the breast. However, doctors can wrongly diagnose a negative case when the patient has breast cancer, due to the difficulty of interpreting the screening images correctly. A mammogram has a black background color and the breast is shown in white and gray color. The white area shows more dense tissue in the breast while the gray area shows less dense tissue.

It is often harder for radiologists to detect breast cancer in the breast with dense tissue. These attributes need the development of automated systems capable of detecting non-trivial patterns of breast behavior that may indicate breast cancer in large mammogram image data sets ahead of time. These features drive the employment of machine learning techniques, which provide ensemble and supervised methods that have been shown to learn non-trivial patterns in data without the need for human involvement and generalize well to historical data.

1.6 Significance of the Study

The project will create a tremendous positive impact on the community, and insurance companies, without forgetting the whole country in general. The implementation of the project will reduce the budget used to buy different medical equipment for testing breast cancer.

1.7 Organization

Throughout the implementation of this study, we made some desktop reviews in order to make sure we have enough concepts to write the first chapter, which is titled Introduction. The introduction gives an overview of the challenge that we addressed and the objectives used in order to solve all study hypotheses. In addition, we examined several related breast cancer research studies, write Chapter 2 which is a literature review, and learned from their failures. As a result, it assisted us in identifying the most effective machine learning models for better-predicting breast cancer before it progresses to a critical stage. To accomplish so, we looked into each of the listed machine learning models that were used in this study, and as a result, we got to the end of our third chapter, which is a research methodology.

Moreover, we used mammogram breast cancer data collected from King Faisal Hospital to make a prediction of breast cancer. Additionally, after gathering data from King Faisal Hospital, we used

the various machine-learning models indicated above. We were able to complete Chapter 4, which is the project results from implementation as a result of applying the various machine learning models to the breast cancer dataset collected from King Faisal Hospital. Referring to chapter 4 results, we determined the most effective machine learning strategies for dealing with our problem statement. Finally, we concluded our study by providing a summary of the entire project as well as a future research recommendation.

1.8 Summary

The first chapter provides an outline of how breast cancer has evolved into a global issue. It also depicts the various approaches that have been utilized and are currently being used to address the problem of breast cancer, which primarily affects women. We conducted this research in light of the aforementioned difficulties. We are employing several machine learning algorithms to raise the patient's breast cancer awareness in order to discover a systematic strategy to address breast cancer.

2. LITERATURE REVIEW

When it comes to the breast cancer problem, which is killing a lot of people all over the world, there are numerous techniques that can be applied. In order to determine the best solution to the problem, various studies have been conducted. Various machine-learning approaches were employed to raise patient awareness about breast cancer, which was one of the many recommended solutions.

Many researchers have conducted breast cancer research using a variety of datasets, including the SEER dataset, mammogram pictures as a dataset, the Wisconsin Data-set, and data from other hospitals. Authors extract and select various aspects from these datasets to complete their research. These are some important findings.

(16) uses 3D photos to show the application of numerous machine learning algorithms precisely supervised models in breast cancer classification and finds that SVM performed well based on the predicted result of the other used models. (17) proved that Support Vector Machine (SVM) is operative in predicting and diagnosing breast cancer, achieving the greatest result a 97.13 percent accuracy.

We have found that (18) proposed a comparison of different weak learners algorithms, and showed that linear support vector machine (SVC) is the considered machine learning model with an optimal accuracy of 97.9% when contrasted to KNN, RF, and NB. Using integrating a clustering approach with an effective probabilistic vector support machine, (19) offered a strategy for the screening of the breast cancer with a 99.10% as accuracy determined by the SVC technique.

Additionally, K-Means and Convolutional Neural Network techniques were utilized by (20) for the identification of Breast cancer. They suggested that the Mammography Image Analysis Society MIAS dataset was used to assess the proposed approach. 95.8% accuracy was achieved using the method. Similar to this, (21) created a classification of breast cancer from histological pictures using a recurrent residual convolutional neural network architecture. The study put a lot of emphasis on contrasting the chosen algorithms. Concerning feature selection, the writers said nothing.

(22) conducted an assessment of numerous Machine Learning algorithms that had been used to classify breast cancer in earlier studies. Furthermore,(3) classified benign and aggressive breast cancers using genetic programming and machine learning techniques. To choose the optimal qualities, the genetic

programming technique was used, together with the high parameter tuning of the machine learning techniques. According to the authors, the proposed method was founded on the roc curves, sensitivity, specificity, precision, and accuracy. The paper claimed that by combining feature preprocessing techniques with the selected classifier algorithms, genetic programming may automatically discover the optimal model.

(23) proposed two alternative machine-learning techniques for classifying breast cancer. Naive Bayes (NB) classifier and K-nearest neighbor (KNN) are two of the techniques used in the study. The study's main objective is to compare the two new implementations, and it uses cross-validation to assess each one's accuracy. Results indicate that KNN outperforms NB classifier (96.19%) in the accuracy of (97.51%) and has a lower error rate as well. Additionally, (24) suggested a method for breast cancer that made distinctions between several kinds of breast cancer.

Moreover, (25) applied the random forest classifier to predict breast cancer. The Wisconsin breast cancer (Diagnostic) dataset was the one used in the study. The University of Wisconsin Hospitals, Madison, originally made the breast cancer dataset available. The Python programming language was employed in the research, and several of its libraries were utilized for the experimental analyses. The dataset was divided into training and testing sets of 75%, and 25%. respectively. The following metrics were used to assess how well the model performed: accuracy, precision, recall, f1-score, and Cohen's Kappa Statistics. The model was able to have a 96% accuracy rate. The precision score was 98%. And 96% of the recall was achieved. The study managed to get the F1-score of 97% and 91% of Cohen's Kappa achieved statistics.

Thus, the main goal of our study is to identify the best methods for detecting and predicting breast cancer using the data set gathered from one of the Rwandan hospitals. We do this by using the aforementioned machine learning models in previous studies and approaches.

2.1 Summary

It is usually essential to read as much as possible about previous projects that are similar to the new research to be conducted. In this regard, we have read a number of sources that were conducted with the help of a machine learning algorithm applied to a breast cancer data set. Different researchers

were able to conduct research in accordance with their objectives. However, none of the studies were conducted using machine learning algorithms on local data sets, whereas the data set acquired from Rwandan hospitals was. As a result, we were excited to apply several machine-learning techniques to the Rwandan data set and compare the findings to those obtained from another online data source.

3. RESEARCH METHODOLOGY

3.1 Research Process

Throughout this conducted research, the Scikitlearn package, and Python programming language was used to conduct all machine learning techniques tests. The different python built-in libraries were applied in order to be able to implement our project. All machine learning models used in this study followed several processes as it is shown in figure 3.1 below.

Process of Breast Cancer Diagnosis using Machine Learning Models approach

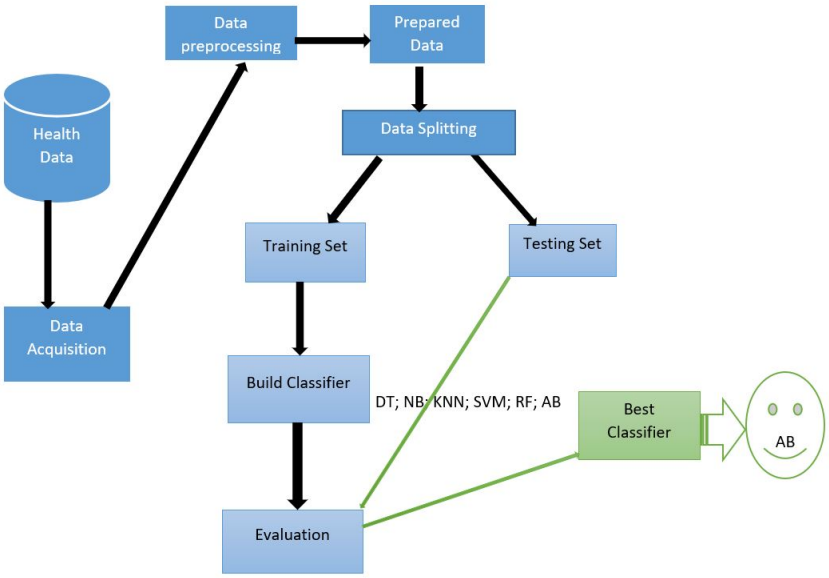


Figure 3.1: Flowchart of the project on machine learning process

Data collection is depicted in the flowchart. The essential dataset was gathered from the KFH radiology department system at this time, and the data sources have been identified (PACS). Data pre-processing comes in at stage two. The data is now cleaned to make it ready for analysis and modeling. Last but not least, we fitted a variety of ML models with chosen traits that seemed to have a strong connection with the target variable during the model prediction and testing process.

3.2 Data Collection

In this study, data from King Faisal Hospital were used. After receiving the King Faisal Rwanda Institutional Review Board (KFH IRB)'s clearance, the data were gathered at the KFH radiology department. We were able to obtain every type of necessary data thanks to the approval letter. The radiology department kept all of the recorded data on a PACS system.

We examined every mammography image pertaining to breast examination during our research. We were able to locate each patient's report and associated mammography photos in PACS. The $1,017$ mammography images were successfully gathered. As we solely utilized machine learning-based models and an ensemble method in our research, the data we collected appeared to be sufficient. Both patients with breast cancer who had been screened and confirmed they have breast cancer as well as healthy patients were recorded during the data collection. It is also crucial to recognize that the information that gathered must be kept confidential because it includes medical records. In this regard, we are disclosing information on the plan for ethical protection.

Ethical Protection Plan

1. **Ethical considerations**

   The Regional Center of Excellence in Biomedical Engineering and E-Health (CEBE) gave its approval to this project. The King Faisal Hospital (KFH) is used for collecting patients' mammography breast images. The research's core ethical tenet is beneficence. There was no harm done to the patients and the outcomes of this study will be beneficial to future patients of healthcare as the research moves into binary breast cancer image classification and modification.

2. **Vulnerable populations**

   Since we simply needed to collect the patient's recorded breast-related data for this study, no one was involved in it.

3. **Information and consent process**

   According to the research objectives, the required data were for both those who have been diagnosed with breast cancer and those who are not. As a result, there won't be a process of completing questionnaires and signing consent forms.

4. **Protection of privacy and confidentiality**

   We conducted the research activities in accordance with all ethical standards to protect the confidentiality of patient information and enhance the quality of services provided to patients in the future. We kept the privacy of the supplied information as a researcher. In KFH's system, we found that the stored patient image was having a name and age and we removed that information. All given information is stored on password-protected laptops and in an encrypted document that is accessible only to myself, and two supervisors.

5. **Conflict of interest**

   There are no declared conflicts of interest among the study's authors or supervisors.

3.3 Research Design Method

Machine Learning (ML) is the study of creating computations and algorithms that allow software systems to become more accurate in predicting outcomes without having to be explicitly coded, hence making human work easier. ML is a broad category of statistical analytic algorithms that develop models for autonomous predictions by iteratively improving in response to training data. Data is at the crucial of ML, which allows computers to learn from it, and make incredible data driven decisions, and predictions. Machine learning is frequently divided into three categories: supervised, unsupervised, and reinforcement learning. In this study, supervised learning algorithms employed to distinguish between benign, and malignant breast cancer, and their combination was used to accurately predict cancer diagnosis labels.

3.3.1 Weak Learners. The primary goal of our research was to find the most usable, and classification machine learning models for the early detection of breast cancer. To do so, we used the supervised machine learning models namely Decision tree (DT), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Support Vector Machine (SVM), on the mammogram images data set gathered from King Faisal Hospital (KFH) and compared the results to see which model has the best accuracy.

1. **Decision Tree Classifier**

   Decision tree (DT) is a supervised machine learning algorithm widely used. Decision tree can be

used for both regression, and classification problem. It has three main components: root ,and internal nodes, branches, and leaves. At each node, an attribute value undergoes a conditional test. The outcome of that test makes the value follows the True or False data-driven branch connects two leaves. The first node of the tree is called root, and the last nodes are called terminal nodes.They determine the class that the value belongs to. The evaluation of the degree of disorder in the two groups that result from a split gives an idea of how good is that split. For this, one uses the Gini Index and Information Gain. The Gini Index is calculated by subtracting from one, the sum of the squared probabilities of each class and is given by

$$GINI\,index = 1 - \sum_{i}^{C} p_i^2, \tag{3.3.1}$$

Where,

p_i is the probability of class i and C is the target variable class.

Small partitions with many distinct values are supported by Information Gain which is the entropy before the split (parent node) minus the entropy after the split (child nodes), the entropy, denoted S, is defined by

$$Entropy(S) = -\sum_{i}^{C} p_i \log_2 p_i \tag{3.3.2}$$

The following decision trees are the operation of the decision tree classifier using problem data. as shown in fig. 3.2.

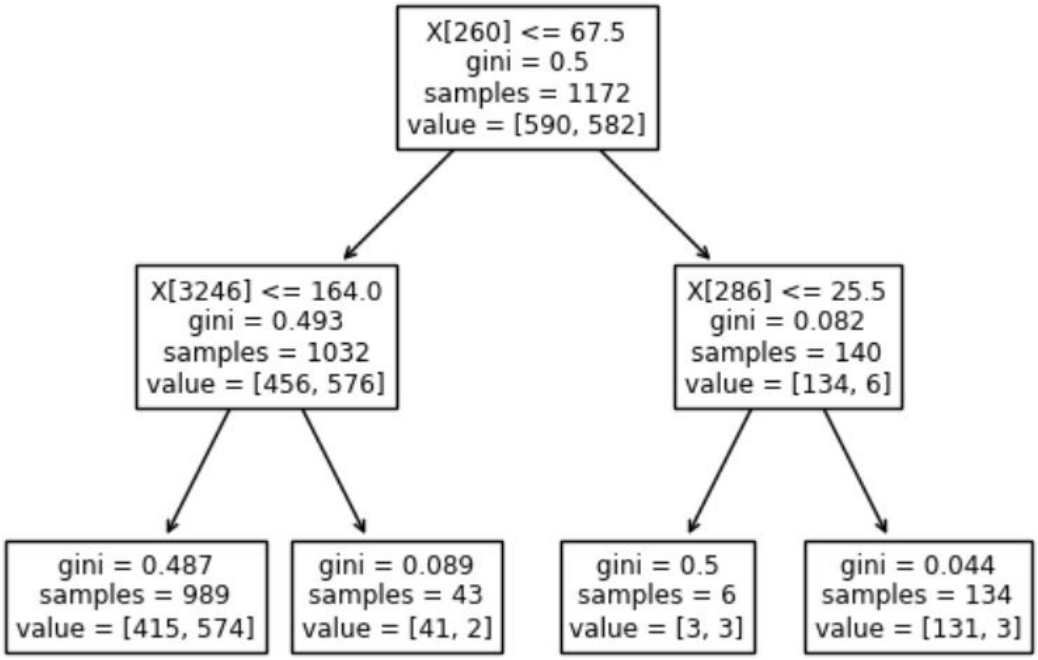


Figure 3.2: Decision tree classifier

.

2. **Naive Bayes Classifier**

The naive Bayes classifier uses the Bayes theorem. For the Naive Bayes (NB) classifier, one assumes that the attributes are conditionally independent of the label. This is not always the case in real life since some attributes are correlated and others have no contribution or contribute differently to the target variable.

For any observation $x = (x_1, ..., x_n)$, the NB classifier uses the Bayes theorem to compute $(P = j|X = x)$ for all the class labels and assigns to x the class with the maximum posterior. Therefore, the Bayes theorem appears to be the principal component of the NB classifier. Indeed, the theorem states that:

$$P(Y = j|X = x) = \frac{P(Y = j)P(X = x|Y = j)}{P(X = x)} \tag{3.3.3}$$

,

where $X = x$ is the random variable that generates the features and Y is the target or class variable. $P(X)$ is the prior probability of the features, $P(Y)$ is the prior of the class, and $P(X = x|Y = j)$ is the likelihood which is the probability of attributes (predictor X) given class (target variable), $P(Y = j|X = x)$ the posterior class probability for given attributes (X).

The full calculation of Bayes theorem for classification problems is in practice challenging. So one needs to simplify the calculation if willing to use the theorem. Using the independence assumptions, we write

$$P(Y = j|X = x) = \frac{P(Y = j)P(X_1 = x_1|Y = j)P(X_2 = x_2|Y)...P(X_n = x_n|Y = j)}{P(X_1 = x_1)P(X_2 = x_2)...P(X_n = x_n)} \quad (3.3.4)$$

which can be expressed as

$$P(Y = j|X = x) = \frac{P(Y = j)\prod_{i=1}^{n} P(X_i = x_i|Y = j)}{P(X_1 = x_1)P(X_2 = x_2)...P(X_n = x_n)} \quad (3.3.5)$$

Since we can observe that the denominator in 3.3.5 does not change for a given input. It is just a scaling factor that depends only on the well-observed dataset. Hence dropping down the denominator gives

$$P(Y = j|X = x) \propto P(Y = j)\prod_{i=1}^{n} P(X_i = x_i|Y = j). \quad (3.3.6)$$

The Bayes classifier is the function that assigns a class label y$\hat{}$ which is expressed as follows.

$$\hat{y} = argmaxP(Y = j)\prod_{i=1}^{n} P(X_i = x_i|Y = j) \quad (3.3.7)$$

3. **K Nearest Neighbors Classifier**

The K Nearest Neighbours (KNN) is a simple classification algorithm that does not make any assumptions. KNN classifier uses the distance metric when classifying the data. This classifier depends on the number of K used. If for example, $K = 3$ for a given two classes problem, we have to find three neighbors of a new data point and classify it by the use of a majority vote. This is shown in Fig. 3.3, below.

(a) Data point. (b) Majority Votes.

Figure 3.3: KNN Classification process (26).

It is worth mentioning that the choice of K is important in the KNN algorithm. For a small number of K, there is a weak bias, but a higher variance, i.e., the small correlation between the two classes, but contrary to a larger number of K. Moreover, as the number of K increased, the decision boundary which divides two classes is being affected. Anytime this happens, each data point tends to belong where it supposes to be. A trick is, if the predictive class has two elements, the number of K to be chosen is an odd number. These details are shown in Fig.3.4, below.

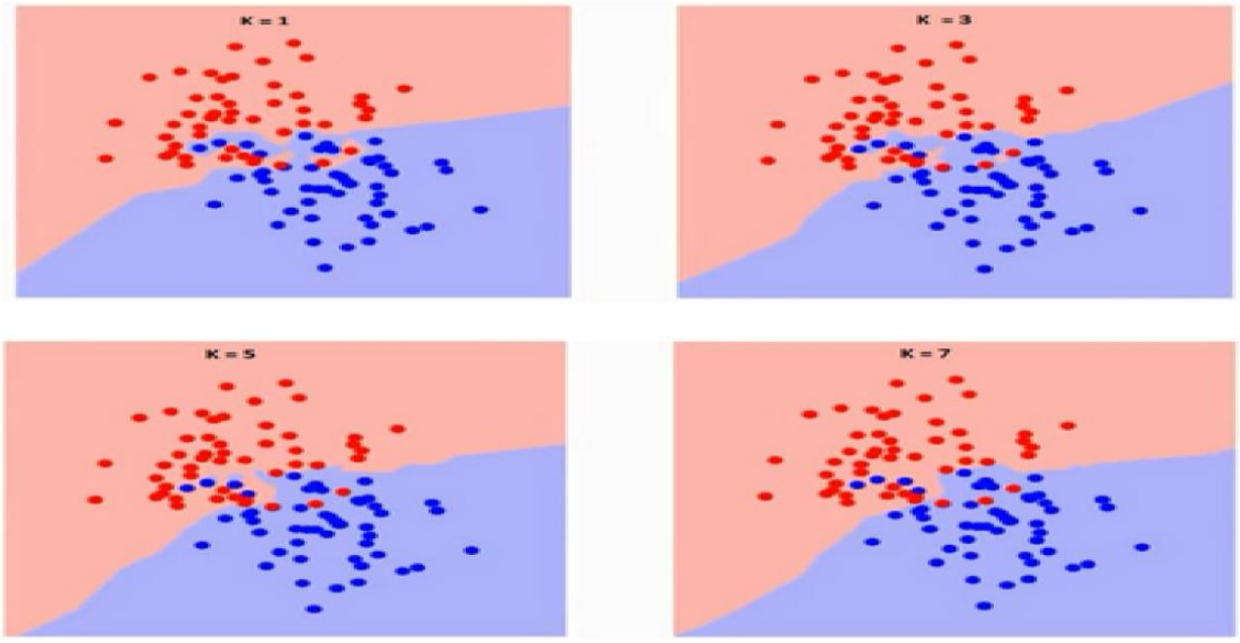


Figure 3.4: KNN dependency on K-parameter (26).

The K- Nearest Neighbours algorithm works under the following steps:

- Standardize the data
- Select the number of K of the neighbors for each example in the data points to be classified
- calculation of the distance between the test group and each train instance with the use of different distance measurements, such as the Euclidean distance which is the most popular one. The euclidean distance can be mathematically expressed as follow.

$$d(x, x_i) = \sqrt{\sum_{i=1}^{k}(x - x_i)^2}, \tag{3.3.8}$$

where,

x is a new data point,

x_i is an existing data point throughout all features.

- To choose the distance, we sorted it in ascending order
- Choose a class with a smaller distance with the use of the majority vote.

4. **Support Vector Machine**

Support Vector Machine (SVM) is a machine learning algorithm for classification and regression problems. The principle of SVM is to separate the data point into two classes using a line or a hyperplane. For illustration, Fig. 3.5a below shows a group of data points on the left represented by a red square $(class1)$, and a data point on the right represented by a circle in blue $(class2)$. Here, we can draw several lines or hyperplanes to classify these data points into two classes, Fig.3.5b. The challenge comes from selecting the best line or hyperplane. As it is clearly shown in Fig. 3.5b, the green line is getting closer to the data point. We can say that the yellow line is the optimal hyperplane but the challenge is how to get it.

(a) Data point in two classes. (b) Data point with lines or hyperplane.

Figure 3.5: Working principal of SVM (27).

Based on the SVM working principle, we found the closest points in the two classes to the line or hyperplane, and they are called support vectors. The distance between the support vectors and the hyperplane is called the margin. With SVM, our goal is to maximize the margin. So, the hyperplane which allows us to have a higher margin is called the optimal hyperplane, see Fig.**??**. As we can see with the support vector moving away, there is an increase in the margin and when other points move, nothing changes in the margin.

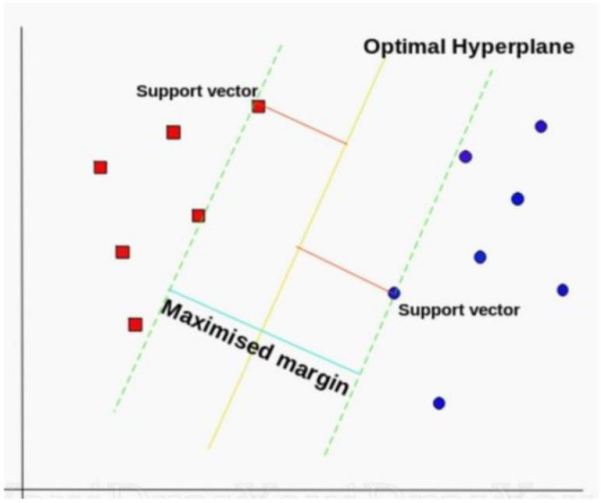


Figure 3.6: Optimal Hyperplane (27).

3.3.2 Ensemble learning techniques. The ensemble method is the act of combining different weak learners to improve stability, and model prediction to make a strong learner. The ensemble model takes into consideration the three types of errors such as bias, variance, and irreducible error that any weak learner. Moreover, bias error stands for how much on average the predicted values are not related to the existing values in the class, meanwhile having high bias shows poor model performance which

leads to misclassification. Thus, variance error shows the difference in the prediction made on the same observation. Having a high variance leads to over-fitting on the training set and makes a poor prediction on the testing set. In this essay, ensemble machine learning techniques such as random forest, and adaboost are used to take into account these pre-mentioned errors made by weak learners.

1. **Random Forest Classifier**

   Random Forest Classifier works by taking multiple results from the decision trees and then aggregating them by vote. The following constitutes its pseudo-code:

   - Let N be the number of instances in the training set, and then create the sample of n instances randomly with replacement, i.e, each instance can be chosen more than once.
   - If P represents the features variables, a number $p < P$ is specified such that at each node, p variables are randomly selected from the P features. The best split of those nodes is used to split the node. The forest will be grown despite the value of p will be constant.
   - Each decision tree will be grown as much as possible without any pruning, i.e, deletion of the branch.
   - Finally predicts new data points by aggregating the different number of trees' predictions, i.e, by making the majority vote.

   The figure below clearly summarizes the steps mentioned above.

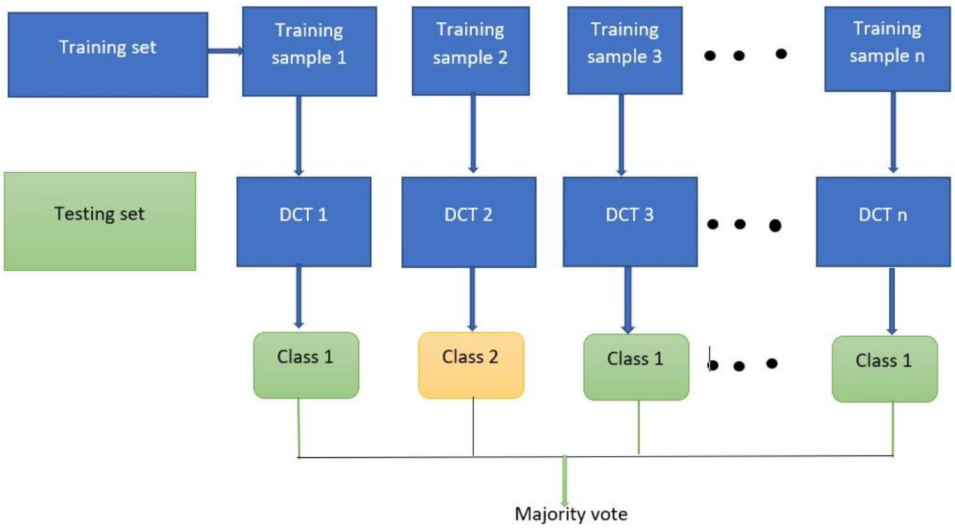


Figure 3.7: Performance of random forest.

2. **Adaboost**

The Adaboost ensemble model reduces errors during classification. It is flexible in the use of different weak learners, such as the decision stump. The predictions made by a decision stump are taken into account by the next decision stump, and repeatedly until it gets to the final classification likely a small error. This shows that the poor performance of a model is improved by the following prediction until most data points are well classified.

Adaboost works under the following procedures.

Consider the training set

$$(x_1, y_1), ..., (x_n, y_n),$$

Where,

$$\begin{cases} x_i \in \mathcal{D}\, is\, an\, instance\, and \\ y_i \in \{-1, +1\}\, its\, class\, , for\, i = 1, 2, 3, , ., .n \end{cases}$$

The objective is to use the adaptive boosting method to boost T, weak performant learners. This consists of assigning and adjusting weights vector $\mathcal{D}_t$ to data points that are misclassified by the learners.

Let first initialize the weights vector by n_1, that is $\mathcal{D}_1(i) = \frac{1}{n}$ for $i = 1, ..., n$. For $t = 1, ..., T$, the $t'th$ learner is trained, and tested. The error is calculated as follows

$$\sum_{i=1}^{n} \chi \left[h_t(x_i) \neq y_i \right] \tag{3.3.9}$$

Where h_t : represents the $t'th$ weak learner. The weights vector the $(t+1)'$ th learner is calculated as:

$$\mathcal{D}_{t+1}(i) = \mathcal{D}_t(i) \exp -\alpha \times t_y \times ih \times t(x_i) \tag{3.3.10}$$

and normalized afterward. The factor α, defined by

$$\alpha_t = \frac{1}{2} \log \left(\frac{1 - \epsilon_t}{\epsilon_t} \right) > 0 \tag{3.3.11}$$

(28) is the estimator weight.

After T iterations over the weak learners, the classifier predicts the class of the observation x as follows:

$$H_y(x) = sign \sum_{t=1}^{T} \alpha_t h_t(x) \tag{3.3.12}$$

3.4 Model Evaluation

Evaluation of machine learning models is crucial since it shows how well they can predict the future. A specific machine learning task, in this case, a classification problem, determines the choice of evaluation measures. Precision, recall, and an F1 score were obtained in our study through the application of the confusion matrix as shown in fig.3.8 below.

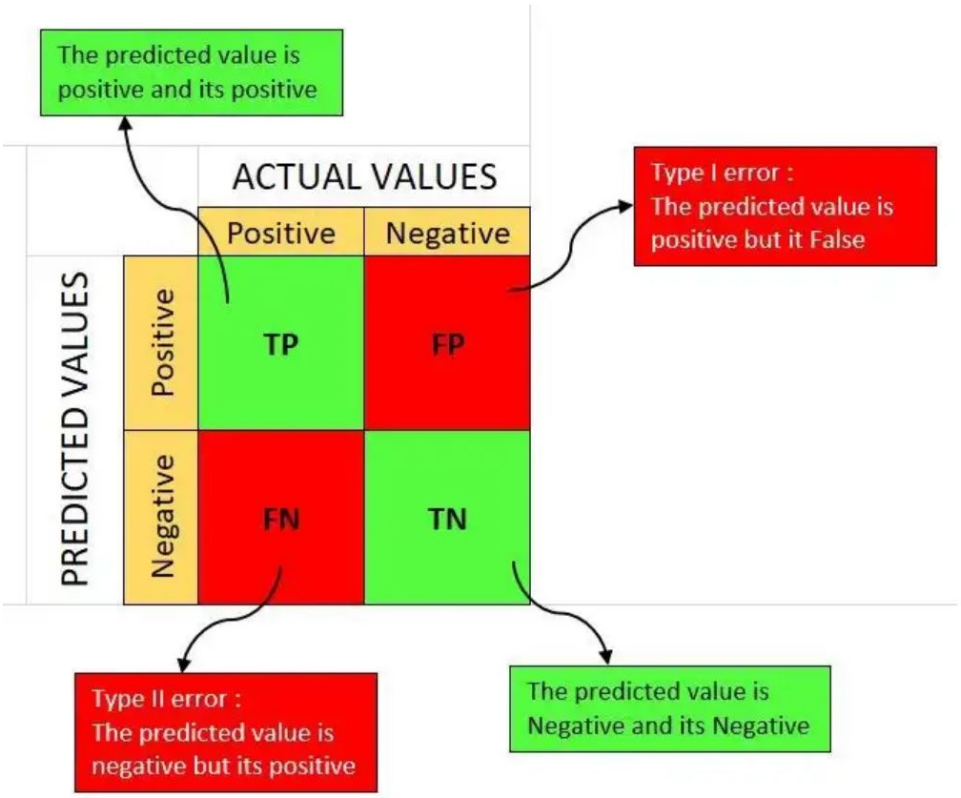


Figure 3.8: Confusion matrix.
(29)

A Confusion matrix is a $N \times N$ matrix used to assess the effectiveness of a classification model, where N is the total number of target classes. The matrix compares the actual target values to those predicted by the machine learning model. High TP and TN rates as well as low FP and FN rates define how well the model performed during the prediction. In addition, it is good to use the confusion matrix when you have an imbalanced data set, (30) as in our case we have normal as the majority class and benign as the minority class. A confusion matrix is a two-by-two matrix that shows a classifier's accurate and inaccurate predictions. It is more helpful when evaluating the effectiveness of a model during classification. It can be used to calculate performance metrics like accuracy, precision, recall, and F1-score in order to assess the effectiveness of a classification model.

The fundamental terms listed below will enable us to identify the different evaluation metrics.

- **True Positives** (TP): when both the predicted and actual values are positive.
- **True negatives** (TN): when both the prediction and the actual value are negative.

- **False positives** (FP): when the prediction is positive but the actual is negative. It is known as the **Type 1 error**
- **False negatives** (FN): when the prediction is negative but the actual is positive. It is known as the **Type 2 error**

For a better understanding and study of our model and its performance, we can basically obtain the additional measures for model evaluation by referring to the confusion matrix.

1. **Accuracy**; simply put, indicates how frequently the classifier predicts correctly. It serves as a gauge of how accurately true predictions are made. Simply said, it informs us of the proportion of how many actual positive predictions are among all total positive predicted. Accuracy can be calculated mathematically as follow.

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{3.4.1}$$

2. **Precision**; it is an indicator of correctness that is obtained in true prediction. In brief, it demonstrates the proportion of how many predictions are actually positive out of all the total positive predicted. Precision is obtained by taking true-positive predictions about the number of true-positive and false-positive predictions. This is the optimal way to measure the exactness of the model. Having low precision clearly shows the highest number of false positives. Precision can be written mathematically as follow.

$$Precision = \frac{TP}{TP + FP} \tag{3.4.2}$$

3. **Recall**; It measures how many instances of the positive class are genuinely anticipated to be positive, or how many actual instances are successfully predicted as positive. Recall, obtained by taking the number of true positives divided by the number of true positive and false-negatives predictions. This metric is the optimal way of measuring the completeness of the classification model. Having classification under low recall clearly shows the highest number of false negatives.

We can calculate the recall mathematically as follow.

$$Recall = \frac{TP}{TP + FN} \tag{3.4.3}$$

4. **F1-score**, is also one of the metrics which is obtained by taking the weighted average of precision and recall. We can write F1- score mathematically as follow.

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{3.4.4}$$

3.5 Summary

The project's implementation necessitates the usage of a specific approach. We employ numerous machine learning algorithms in this work, specifically weak learners DT, NB, KNN, and SVM, as well as ensemble machine learning classifier boosting (AD) and bagging classifier RF. All of the models mentioned above were created using Python programming libraries. The study was successfully carried out using the mathematics underlying the machine learning techniques. All of the machine learning models were trained using a local data set derived from mammography images obtained from King Faisal Hospital.

The used data set was created with an ethical protection plan in mind. Furthermore, it is always important to ensure that the model correctly classified the mammography pictures. In this regard, we used the confusion matrix to gain a better picture of where the machine learning classifier performed well and where it struggled.

4. PROJECT RESULTS

4.1 Data Presentation

The data set used in this study was prepared after collecting 1017 mammogram images from the King Faisal Hospital system. After receiving approval from KFH 's IRB, we were able to access the hospital's PACS system. In the two formats JPG, and $JPEG$, we gathered 284 benign patients and 733 normal patients among 1017 mammogram images collected. Our study involves a binary classification problem, and the data set is unbalanced because cancerous is the minority class and normal is the majority class. The figure 4.1 below illustrates how unevenly distributed our data set is. The majority class's share is 72.1%, and the minority class' share is 28.8%.

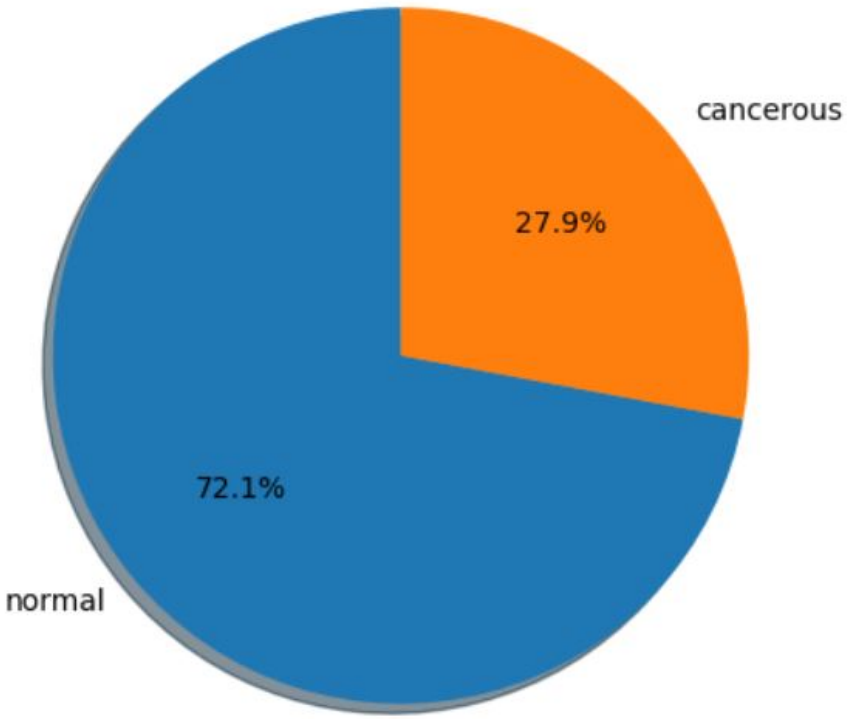


Figure 4.1: Breast examination overview

.

4.2 Data Preprocessing

Data preparation is an essential step in the data mining process. thus they have a direct impact on task success rates. It has to work with information that is unimportant, noisy, and untrustworthy. It includes data conversion, if necessary. The collected data had varied sizes, therefore we resized them to the same size. We resized each image to the dimensions depicted in table 4.1, below. where,

Image Dimension	:	(200, 161, 3)
Image Height	:	200
Image Width	:	161
Number of Channels	:	3

Table 4.1: Image size

.

- **Height** shows the number of image pixel rows or the number of pixels in each image array column.
- **Width** shows the number of image pixel columns or the number of pixels in each row of the image array.
- **The number of channels** corresponds to the number of components required to represent each pixel. In our case, the channel number$= 3$ represents Red, Green, and Blue (RGB).

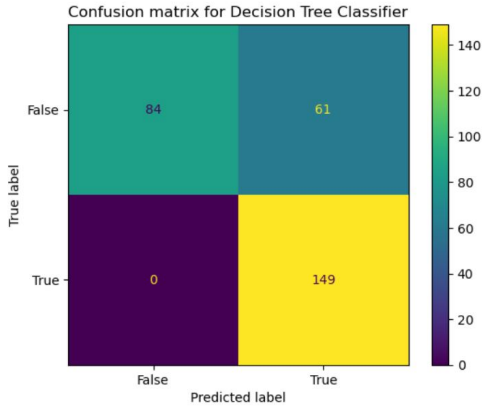
4.2.1 Oversampling. Oversampling can be thought of as the opposite of undersampling since, instead of removing points from the majority class to balance out the classes, we simply add additional points to the minority class. Oversampling can be done using a variety of techniques, like random sampling and smote which stands for Synthetic Minority Oversampling Techniques. In our work, we employ the Smote technique since duplicate cases in random sampling can lead to model overfitting. By inserting a point between existing observations from the original dataset, the smote approach creates new observations. The smote technique uses the ideal of the K-Nearest Neighbors approach.

4.3 Results Analysis

4.3.1 Base Learner's Models Results Analysis. The research used the application of four base learners model classification algorithms on the data set and then made a comparison of the base learner's model results and also predicted the outcome of whether a patient has breast cancer or not from the given data. The results of those selected base learners namely Naïve Bayes Classifier, Decision Tree Classifier, Linear Support Vector Machine, and K Nearest Neighbours Classifier were compared by using a classification report as shown in fig. 4.2 and fig. 4.3. We got results with the use of python and implementation was done by using the sklearn library. The model performance was evaluated by the Confusion Matrix. Under consideration of that matrix, the true positives, true negatives along with false positives and false negatives were used to find Recall, Precision, and F1 score.

Model : Decision Tree Classifier

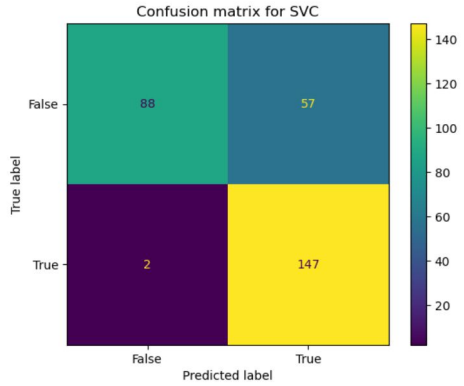
	precision	recall	f1-score	support
benign	1.00	0.58	0.73	145
normal	0.71	1.00	0.83	149
accuracy			0.79	294
macro avg	0.85	0.79	0.78	294
weighted avg	0.85	0.79	0.78	294



(a) Evaluation of Decision tree classifier

Model : SVC

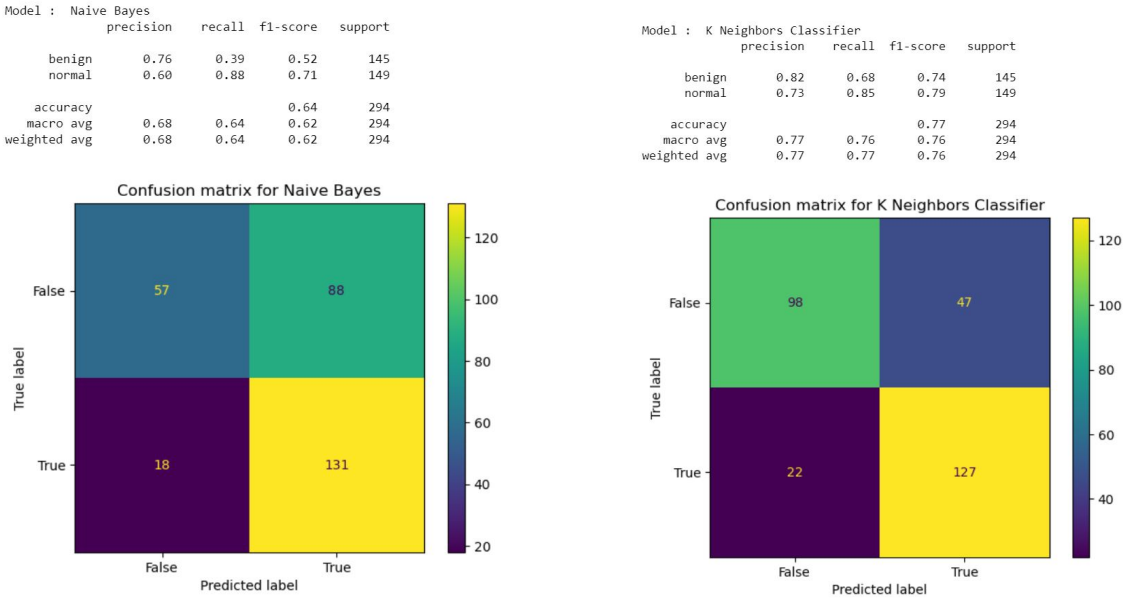
	precision	recall	f1-score	support
benign	0.98	0.61	0.75	145
normal	0.72	0.99	0.83	149
accuracy			0.80	294
macro avg	0.85	0.80	0.79	294
weighted avg	0.85	0.80	0.79	294



(b) Evaluation of Support Vector Machine

Figure 4.2: Model Evaluation

.



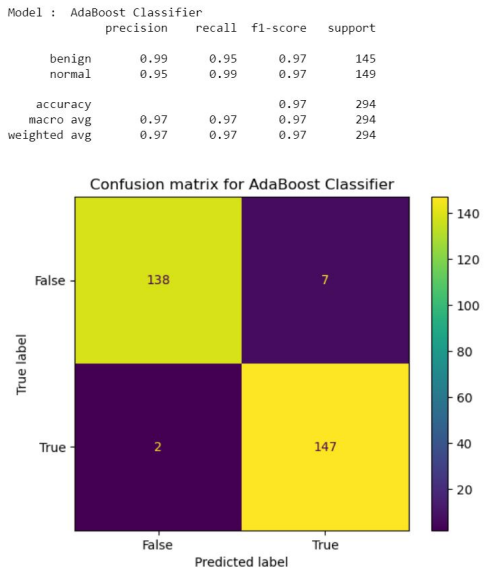
(a) Evaluation of Naive Bayes (b) Evaluation of K Nearest Neighbours

Figure 4.3: Performance of each base learner

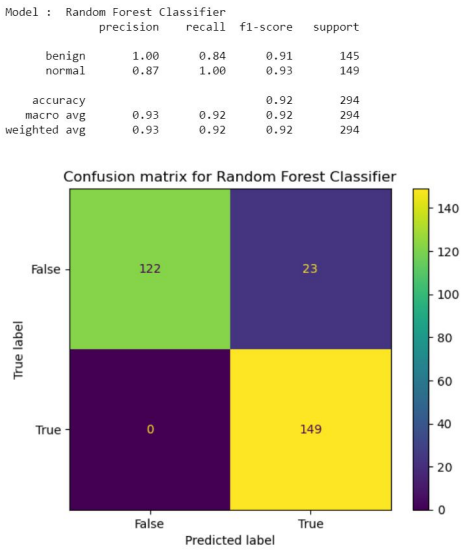
.

According to what we are seeing in fig. 4.2, and 4.3 above, Linear Support Vector Machine (SVC) is the first model among base learners models with optimal accuracy of 79.93%followed by Decision Tree (DT) with optimal accuracy of 79.25%. Decision Tree is followed by K Nearest Neighbours (KNN) with optimal accuracy of 68.42%, and Naive Bayes (NB) was the last one with optimal accuracy of 63.95%.This is how those base learner's models classify whether a patient has breast cancer or not and the error made during classification. Linear Support Vector Machine Classifier predicts 88 for patients with breast cancer and 147 for patients without breast cancer. Besides, 2 patients were misclassified as patients without breast cancer and 57 patients were misclassified as a patient with breast cancer. All of the models follow that way in the confusion matrix.

4.3.2 Ensemble Model Result Analysis. Referring to errors made by base learner's models, as shown in 4.2, and 4.3 above, we found other machine learning techniques to improve the accuracy and reduce the error during the classification. An ensemble machine learning classification algorithm called Adaboost and Random Forest was used to increase the robustness of the model. The predicted results were compared among themselves which were implemented with the python sklearn library as shown in fig.4.4 below.



(a) Performance of Adaboost Classifier



(b) Performance of Random Forest Classifier

Figure 4.4: Ensemble Model Performance

.

Fig. 4.4 shows that the Adaboost Classifier classified the patient well according to the actual data point in the target variable. Adaboost is the first and the selected model for our research with optimal accuracy of 97%. This model has a recall of 95%, hence recall represents the completeness of the model and a 138 of patients with breast cancer were well classified as required. The performance of Adaboost was followed by Random Forest Classifier with an optimal accuracy of 92%,

Models were evaluated by Confusion Matrix and Adaboost Classify 138 patients with breast cancer and 147 without Breast cancer. 7 patients were misclassified as with breast cancer and 2 were misclassified as a patient without breast cancer. Below is a graphical representation of ensemble models and base learners models as shown in Fig 4.2.

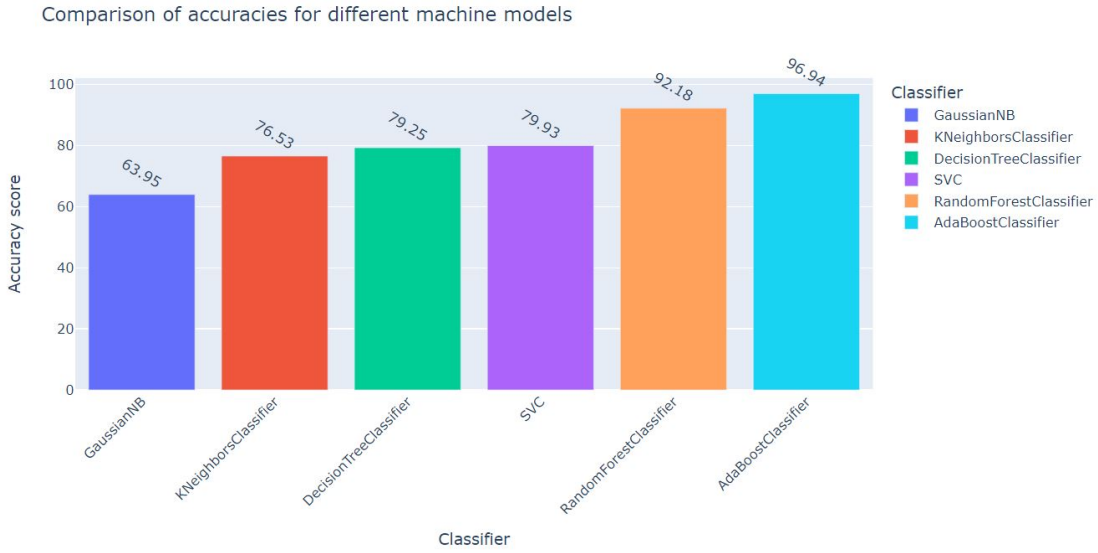
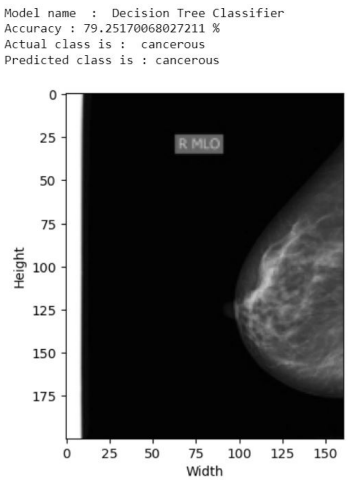


Table 4.2: Graphical Representation of models performance

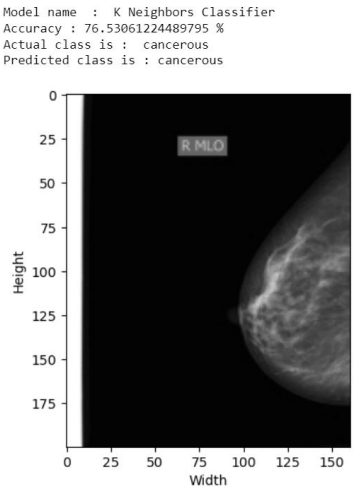
.

.

All of the machine learning algorithms employed in this study went through multiple cycles of development in order to classify every mammogram image (1017) that was gathered from the King Faisal Hospital. The figures below were taken during the prediction attempt by the model and contrasted with the actual class of the image.



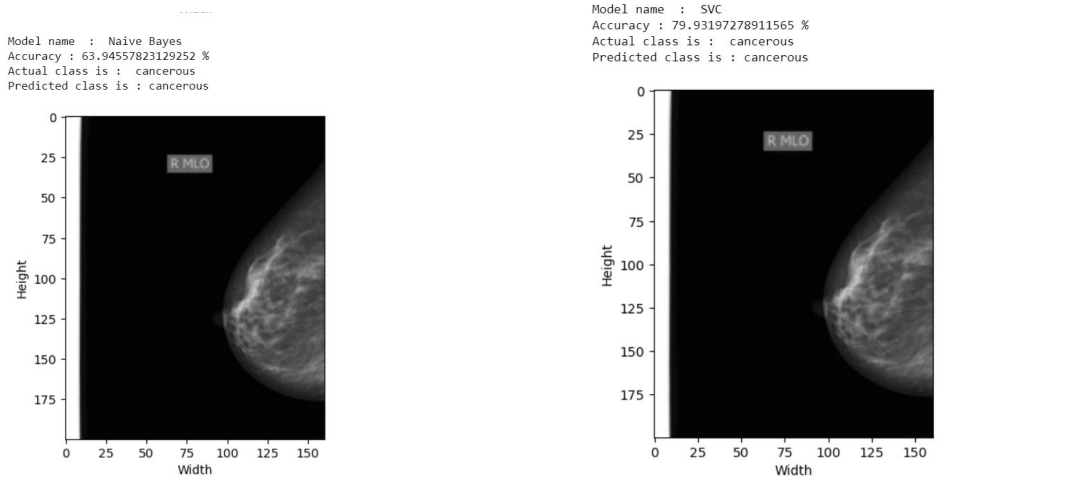
(a) Prediction of Decision Tree Classifier



(b) Prediction of K Nearest Neighbours Classifier

Figure 4.5: Base Learner Predictions

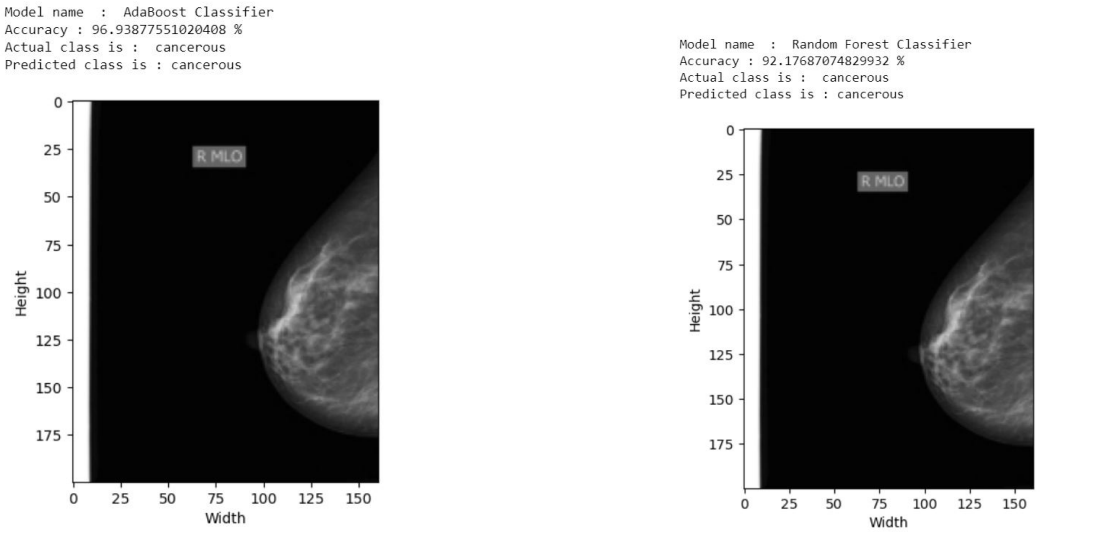
.



(a) Prediction of Naive Bayes Classifier (b) Prediction of Linear Support Vector Classifier

Figure 4.6: Base Learner Predictions

.



(a) Prediction of Adaboost Classifier (b) Prediction of random forest Classifier

Figure 4.7: Ensemble Predictions

.

4.4 Summary

After training each of these machine-learning models, the prediction was completed, and evaluation metrics revealed that Ada-Boost (AB) and Random forest (RF) outperformed base learners models, with an accuracy of 97%, and 92% respectively.As shown in Table.4.3 below, are all prediction results

from base learners and ensemble models were recorded.

Classifier	Accuracy score
GaussianNB	63.945578
KNeighborsClassifier	76.530612
DecisionTreeClassifier	79.251701
SVC	79.931973
RandomForestClassifier	92.176871
AdaBoostClassifier	96.938776

Table 4.3: General performance of models performance

.

.

5. CONCLUSION AND RECOMMENDATION

5.1 Conclusion

Six machine learning classification algorithms, including Random Forest (RF), Ad-boost (AD), Naive Bayes (NB), Decision Trees (DT), K Nearest Neighbors (KNN), and Support Vector Machine (SVC), were employed in this study to determine whether or not a patient is likely to have breast illness. The outcomes demonstrated that the chosen classifiers can help radiologists make decisions. Adaboost performed better than the other machine learning algorithms used, with an ideal accuracy of 98%. In a nutshell, the ensemble-based model provided as the proposed method outperformed the other ML models.

5.2 Recommendations

Our research suggests that the accuracy was good, but it would be preferable to increase the accuracy as much as feasible in the health industry. The mathematical underpinnings of both the Random Forest classifier and Ad-Boost should be developed in order to enable the creation of those models from scratch in future works, which might improve the performance of our ensemble models.

Acknowledgements

This is optional and should be at most half a page. Thanks Ma, Thanks Pa. One paragraph in normal language is the most respectful.

Do not use too much bold, any figures, or sign at the bottom.

References

[1] P. Mathur, K. Sathishkumar, M. Chaturvedi, P. Das, K. L. Sudarshan, S. Santhappan, V. Nallasamy, A. John, S. Narasimhan, F. S. Roselind *et al.*, "Cancer statistics, 2020: report from national cancer registry programme, india," *JCO Global Oncology*, vol. 6, pp. 1063–1075, 2020.

[2] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*. IEEE, 2018, pp. 1–4.

[3] H. Dhahri, E. Al Maghayreh, A. Mahmood, W. Elkilani, and M. Faisal Nagi, "Automated breast cancer diagnosis based on machine learning algorithms," *Journal of healthcare engineering*, vol. 2019, 2019.

[4] R. R. Janghel, A. Shukla, R. Tiwari, and R. Kala, "Breast cancer diagnosis using artificial neural network models," in *The 3rd International Conference on Information Sciences and Interaction Sciences*. IEEE, 2010, pp. 89–94.

[5] S. J. Miah, E. Camilleri, and H. Q. Vu, "Big data in healthcare research: a survey study," *Journal of Computer Information Systems*, vol. 62, no. 3, pp. 480–492, 2022.

[6] N. Jayasri and R. Aruna, "Big data analytics in health care by data mining and classification techniques," *ICT Express*, vol. 8, no. 2, pp. 250–257, 2022.

[7] WHO, "Breast cancer," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[8] N. C. Institute, "What is a cancer," 2021. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[9] H. K. Matthews, C. Bertoli, and R. A. de Bruin, "Cell cycle control in cancer," *Nature Reviews Molecular Cell Biology*, vol. 23, no. 1, pp. 74–88, 2022.

[10] A. Stachs, J. Stubert, T. Reimer, and S. Hartmann, "Benign breast disease in women," *Deutsches Ärzteblatt International*, vol. 116, no. 33-34, p. 565, 2019.

[11] A. Fanizzi, T. Basile, L. Losurdo, R. Bellotti, U. Bottigli, R. Dentamaro, V. Didonna, A. Fausto, R. Massafra, M. Moschetta *et al.*, "A machine learning approach on multiscale texture analysis for breast microcalcification diagnosis," *BMC bioinformatics*, vol. 21, no. 2, pp. 1–11, 2020.

[12] M.-L. Vázquez, I. Vargas, M. Rubio-Valera, I. Aznar-Lou, P. Eguiguren, A.-S. Mogollón-Pérez, A.-L. Torres, A. Peralta, S. Dias, and S. S. Jervelund, "Improving equity in access to early diagnosis of cancer in different healthcare systems of latin america: protocol for the equitycancer-la implementation-effectiveness hybrid study," *BMJ open*, vol. 12, no. 12, p. e067439, 2022.

[13] E. Androulakis, E. Kordi, A. Mprouziotis, S. Xesfiggi, and V. Salatas, "Evaluating the cancer control programmes through the study of perception and behaviour of greek students." *INFORMATION COMMUNICATION TECHNOLOGIES IN HEALTH*, p. 318.

[14] R. Nambaziira, L. C. Niteka, J. M. V. Dusengimana, J. Ruhumuriza, K. P. Bhangdia, J. C. Mugunga, M. L. Uwineza, V. Rugema, P. Erfani, C. Shyirambere *et al.*, "Health system costs of a breast cancer early diagnosis programme in a rural district of rwanda: a retrospective, cross-sectional economic analysis," *BMJ open*, vol. 12, no. 6, p. e062357, 2022.

[15] A. Yala, P. G. Mikhael, C. Lehman, G. Lin, F. Strand, Y.-L. Wan, K. Hughes, S. Satuluru, T. Kim, I. Banerjee *et al.*, "Optimizing risk-based breast cancer screening policies with reinforcement learning," *Nature Medicine*, vol. 28, no. 1, pp. 136–143, 2022.

[16] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for rf-based breast cancer detection," in *2017 Computing and Electromagnetics International Workshop (CEM)*. IEEE, 2017, pp. 13–14.

[17] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[18] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in *2018 International conference on electronics, control, optimization and computer science (ICECOCS)*. IEEE, 2018, pp. 1–5.

[19] A. H. Osman, "An enhanced breast cancer diagnosis scheme based on two-step-svm technique," *Int. J. Adv. Comput. Sci. Appl*, vol. 8, no. 4, pp. 158–165, 2017.

[20] S. Shamy and J. Dheeba, "A research on detection and classification of breast cancer using k-means gmm & cnn algorithms," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 6S, pp. 501–505, 2019.

[21] M. Z. Alom, C. Yakopcic, M. Nasrin, T. M. Taha, V. K. Asari *et al.*, "Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network," *Journal of digital imaging*, vol. 32, no. 4, pp. 605–617, 2019.

[22] A.-A. Nahid and Y. Kong, "Involvement of machine learning for breast cancer image classification: a survey," *Computational and mathematical methods in medicine*, vol. 2017, 2017.

[23] A. Eleyan, "Breast cancer classification using moments," in *2012 20th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2012, pp. 1–4.

[24] H. T. T. Thein, K. M. M. Tun *et al.*, "An approach for breast cancer diagnosis classification using neural network," *advanced computing: an international journal (acij)*, vol. 6, no. 1, pp. 1–11, 2015.

[25] A. M. Oyelakin, "A model for the classification of breast cancer using random forest algorithm," 2021.

[26] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, no. 4, pp. 580–585, 1985.

[27] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[28] R. E. Schapire, "Explaining adaboost," in *Empirical inference*. Springer, 2013, pp. 37–52.

[29] A. Vidhya, "What is a confusion matrix?" [Online]. Available: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

[30] J. Liang, "Confusion matrix," *POGIL Activity Clearinghouse*, vol. 3, no. 4, 2022.