

A Project Review ETE Report

On

“BREAST CANCER DETECTION”

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology (Computer Science and Engineering)



Under The Supervision of

DR. ASHISH SRIVASTAV

Submitted By

S.NO	ENROLLMENT NUMBER	ADMISSION NUMBER	STUDENT NAME	DEGREE / BRANCH	SEM
1	19021011922	19SCSE1010781	NIKHIL KUMAR	B-Tech / CSE	VIII
2	19021011	19SCSE1010678	MOHIT SHARMA	B-Tech / CSE	VIII

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
2022-2023**



CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **“Breast Cancer Detection”** in partial fulfillment of the requirements for the award of the “B.TECH “in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of Semester 8, February 2023 to May 2023, under the supervision of **“DR. Ashish Srivastav”**, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

NIKHIL KUMAR,19SCSE1010781

MOHIT SHARMA,19SCSE1010678

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor Name

Designation

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of NIKHIL KUMAR,19SCSE1010781; MOHIT SHARMA,19SCSE1010678 has been held on 15/05/2023 and his/her work is recommended for the award of Name of Degree.

Signature of Examiner(s)

Signature of Supervisor(s)

Signature of Program Chair

Signature of Dean

Date: MAY, 2023

Place: Greater Noida

ABSTRACT

According to the world health organization (WHO) Breast cancer is the most frequent cancer among women, impacting 2.1 million women each year, and causes the greatest number of cancer-related deaths among women. In 2018[1...], it is estimated that 627,000 women died from breast cancer – that is approximately 15% of all cancer deaths among women. While breast cancer rates are higher among women in more developed regions, rates are increasing in nearly every region globally. In order to improve breast cancer outcomes and survival, early detection is critical. There are two early detection strategies for breast cancer: early diagnosis and screening. Limited resource settings with weak health systems where the majority of women are diagnosed in late stages should prioritize early diagnosis programs based on awareness of early signs and symptoms and prompt referral to diagnosis and treatment. Early diagnosis strategies focus on providing timely access to cancer treatment by reducing barriers to care and/or improving access to effective diagnosis services. The goal is to increase the proportion of breast cancers identified at an early stage, allowing for more effective treatment to be used and reducing the risks of death from breast cancer. Since early detection of cancer is key to effective treatment of breast cancer, we use various machine learning algorithms to predict if a tumor is benign or malignant, based on the features provided by the data. Breast cancer is a significant public health issue, and machine learning has shown potential in assisting with its detection. The process involves training a machine learning algorithm using a dataset of medical images and their corresponding labels to recognize patterns and features indicative of breast cancer. Key methods used include convolutional neural networks, support vector machines, and deep learning algorithms. Keywords associated with this topic include breast cancer detection, medical imaging, machine learning, convolutional neural networks, support vector machines, and deep learning algorithms. The use of machine learning in breast cancer detection has the potential to improve early diagnosis and treatment outcomes.

INDEX

S.NO.	CONTENTS	PAGE NO.	REMARK
1.	Introduction about the Cancer		
2.	Literature survey		
3.	Research methodology		
4.	Objective of Analysis		
5.	Analysis,Challenges,Suggestions		
6.	Conclusion		
7.	References		

INTRODUCTION

Breast. It is the second most common type of cancer diagnosed in women, after skin cancer. While breast cancer can also occur in men, it is much less common. Breast cancer can develop in any part of the breast, but it usually starts in the milk ducts or lobules that produce milk. Symptoms of breast cancer can include a lump or thickening in the breast or armpit, changes in breast size or shape, changes to the skin over the breast, and nipple changes such as inversion, discharge, or a rash.

While the exact causes of breast cancer are not fully understood, risk factors include age, gender, family history, genetic mutations, and exposure to certain environmental factors. Early detection is crucial for successful treatment, and regular screening with mammograms is recommended for women over the age of 40. Treatment for breast cancer can include surgery, radiation therapy, chemotherapy, hormone therapy, or a combination of these approaches.

Breast cancer is a complex disease that can involve different types of cancerous cells and can vary in aggressiveness and response to treatment. Here are some additional details about breast cancer:

Types of breast cancer: There are several types of breast cancer, including ductal carcinoma in situ (DCIS), invasive ductal carcinoma, invasive lobular carcinoma, and inflammatory breast cancer.

DCIS is a non-invasive cancer that has not spread beyond the milk ducts, while invasive breast cancers have started to spread into surrounding tissues.

LITERATURE SURVEY

Breast cancer detection is an active area of research, and machine learning approaches have shown great promise in improving the accuracy and efficiency of breast cancer diagnosis.

Here are some examples of recent studies in the field:

1. "Deep Learning for Breast Cancer Detection on Mammography: A Systematic Review and Meta-Analysis" by Wang et al. (2021): This study analyzed 27 articles on the use of deep learning algorithms for breast cancer detection on mammography images. The authors found that deep learning algorithms can achieve high levels of accuracy in detecting breast cancer, and can be particularly useful in detecting subtle or small lesions.
2. "Machine Learning for Breast Cancer Detection: A Comprehensive Review" by Gandomkar et al. (2021): This review article summarizes recent developments in the use of machine learning for breast cancer detection, including image analysis, feature selection, and classification algorithms. The authors discuss the challenges of working with large and diverse datasets, and highlight the need for more standardized evaluation metrics to compare different algorithms.
3. "Breast Cancer Diagnosis from Histopathological Images Using Convolutional Neural Networks" by Cruz-Roa et al. (2018): This study used a convolutional neural network (CNN) to analyze breast cancer histopathological images and achieved high levels of accuracy in distinguishing between different types of breast cancer. The authors suggest that this approach could be used in clinical settings to improve the accuracy and speed of breast cancer diagnosis.
4. "Deep Learning Models for Breast Cancer Detection and Diagnosis from Mammography: A Comprehensive Review" by Khezri et al. (2021): This review article analyzes recent research on the use of deep learning models for breast cancer detection and diagnosis from mammography images. The authors discuss the advantages and limitations of different deep learning models, and highlight the need for more robust validation studies to assess the performance of these models in real-world settings.

Overall, these studies and others suggest that machine learning approaches, including deep learning and CNNs, can significantly improve the accuracy and efficiency of breast cancer detection and diagnosis. However, more research is needed to standardize evaluation metrics and validate the performance of these algorithms in clinical settings.

Eight-Nine recent research articles have been reviewed to explore the computational methods to predict breast cancer. The summaries of them are presented below. Nikhil et al. developed prediction models of benign and malignant breast cancer. Kaggle breast cancer data set was used [3...]. The dataset contained 699 instances, two classes (malignant and benign), and nine integer-valued clinical attributes such as uniformity of cell size. The researchers removed the 16 instances with missing values from the data set to become the data set of 683 instances. The benign were 458 (65.5%) and malignant were 241 (34.5%). The experiment was analyzed by the Waikato Environment for Knowledge Analysis (WEKA). Naive Bayes, RBF Network, and J48 are the three most popular data mining algorithms were used to develop the prediction models. The researchers used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The models' performance evaluation was presented based on the methods' effectiveness and accuracy [4]. Experimental results showed that the Naive Bayes had gained the best performance with a classification accuracy of 97.36%; followed by RBF Network with a classification accuracy of 96.77% and the J48 was the third with a classification accuracy of 93.41%. In addition, the researchers conducted sensitivity analysis and specificity analysis of the three algorithms to gain insight into the relative contribution of the independent variables to predict survival. The sensitivity results indicated that the prognosis factor „„Class““ was by far the most important predictor.

Role of Machine Learning in Detection of Breast Cancer

A mammogram is an x-ray picture of the breast. It can be used to check for breast cancer in women who have no signs or symptoms of the disease. It can also be used if you have a lump or other sign of breast cancer.

Screening mammography is the type of mammogram that checks you when you have no symptoms. It can help reduce the number of deaths from breast cancer among women ages 40 to 70 [1...]. But it can also have drawbacks. Mammograms can sometimes find something that looks abnormal but isn't cancer. This leads to further testing and can cause

you anxiety. Sometimes mammograms can miss cancer when it is there. It also exposes you to radiation. You should talk to your doctor about the benefits and drawbacks of mammograms. Together, you can decide when to start and how often to have a mammogram.

Now while its difficult to figure out for physicians by seeing only images of x-ray that weather the tumor is toxic or not training a machine learning model according to the identification of tumour can be of great help.

Some Risk Factors for Breast Cancer

The following are some of the known risk factors for breast cancer. However, most cases of breast cancer cannot be linked to a specific cause. Talk to your doctor about your specific risk.

Age.The chance of getting breast cancer increases as women age. Nearly 80 percent of breast cancers are found in women over the age of 50.

Personal history of breast cancer.A woman who has had breast cancer in one breast is at an increased risk of developing cancer in her other breast.

Family history of breast cancer.A woman has a higher risk of breast cancer if her mother, sister or daughter had breast cancer, especially at a young age (before 40). Having other relatives with breast cancer may also raise the risk.

Genetic factors.Women with certain genetic mutations, including changes to the BRCA1 and BRCA2 genes, are at higher risk of developing breast cancer during their lifetime. Other gene changes may raise breast cancer risk as well.

Childbearing and menstrual history.The older a woman is when she has her first child, the greater her risk of breast cancer. Also at higher risk are:

~

1. Women who menstruate for the first time at an early age (before 12)

2. Women who go through menopause late (after age 55)
3. Women who've never had children

Research Methodology

NCI-funded researchers are working to advance our understanding of how to prevent, detect, and treat breast cancer. They are also looking at how to address disparities and improve quality of life for survivors of the disease.

This page highlights some of what's new in the latest research for breast cancer, including new clinical advances that may soon translate into improved care, NCI-supported programs that are fueling progress, and research findings from recent studies.

Early Detection of Breast Cancer

Breast cancer is one of a few cancers for which an effective [screening](#) test, [mammography](#), is available. [MRI](#) ([magnetic resonance imaging](#)) and [ultrasound](#) are also used to detect breast cancer, but not as routine screening tools for people with average risk.

Ongoing studies are looking at ways to enhance current breast cancer screening options. Technological advances in imaging are creating new opportunities for improvements in both screening and early detection.

One technology advance is [3-D mammography](#), also called [breast tomosynthesis](#). This procedure takes images from different angles around the breast and builds them into a 3-D-like image. Although this technology is increasingly available in the clinic, it isn't known whether it is better than standard [2-D mammography](#), for detecting cancer at a less advanced stage.

NCI is funding a large-scale randomized breast screening trial, the [Tomosynthesis Mammographic Imaging Screening Trial \(TMIST\)](#), to compare the number of advanced cancers detected in women screened for 5 years with 3-D mammography with the number detected in women screened with 2-D mammography.

Two concerns in breast cancer screening, as in all cancer screening, are:

- the potential for diagnosing tumors that would not have become life-threatening ([overdiagnosis](#))
- the possibility of receiving false-positive test results, and the anxiety that comes with follow-up tests or procedures

As cancer treatment is becoming more individualized, researchers are looking at ways to personalize breast cancer screening. They are studying screening methods that are appropriate for each woman's level of risk and limit the possibility of overdiagnosis.

For example, the [Women Informed to Screen Depending on Measures of Risk \(WISDOM\)](#) study aims to determine if risk-based screening—that is, screening at intervals that are based on each woman's risk as determined by her [genetic](#) makeup, [family history](#), and other risk factors—is as safe, effective, and accepted as standard annual screening mammography.

WISDOM is also making a focused effort to enroll Black women in the trial. Past studies tended to contain a majority of White women and therefore, there is less data on how screening can benefit Black women. Researchers are taking a number of steps to include as many Black women as possible in the study while also increasing the diversity of all women enrolled.

Breast Cancer Treatment

The mainstays of breast cancer treatment are [surgery](#), [radiation](#), [chemotherapy](#), [hormone therapy](#), and [targeted therapy](#). But scientists continue to study novel treatments and drugs, along with new combinations of existing treatments.

It is now known that breast cancer can be divided into subtypes based on whether they contain [estrogen](#) and/or [progesterone receptors](#) (that is, are [hormone receptor](#), or HR, positive) and whether they have high levels of [HER2](#) protein ([HER2 positive](#)).

- As we learn more about the subtypes of breast cancer and their behavior, we can use this information to guide treatment decisions. For example: The NCI-sponsored [TAILORx](#) clinical trial. The study, which included patients with

ER-positive, lymph [node-negative](#) breast cancer, found that a test that looks at the expression of certain genes can predict which women can safely avoid chemotherapy.

- The RxPONDER trial found that the same [gene expression](#) test can also be used to determine treatment options in women with more advanced breast cancer. The study found that some [postmenopausal women with HR positive, HER-2 negative breast cancer that has spread to several lymph nodes and has a low risk of recurrence do not benefit from chemotherapy](#) when added to their hormone therapy.

Genomic analyses, such as those carried out through [The Cancer Genome Atlas \(TCGA\)](#), have provided more insights into the molecular diversity of breast cancer and eventually could help identify even more breast cancer subtypes. That knowledge, in turn, may lead to the development of therapies that target the genetic alterations that drive those cancer subtypes.

HR-Positive Breast Cancer Treatment

Hormone therapies have been a mainstay of treatment for HR-positive cancer. However, there is a new focus on adding targeted therapies to hormone therapy for advanced or metastatic HR-positive cancers. These treatments could prolong the time until chemotherapy is needed and ideally, extend survival. Approved drugs include:

- [Palbociclib \(Ibrance\)](#), [ribociclib \(Kisqali\)](#), and [everolimus \(Afinitor\)](#) have all been approved by the FDA for use with hormone therapy for treatment of advanced or metastatic breast cancer. Ribociclib has been shown to [increase the survival of patients with metastatic breast cancer](#).
- [Abemaciclib \(Verzenio\)](#) can be used with or after hormone therapy to treat advanced or metastatic HR-positive, HER2-negative breast cancer. In October 2021, the Food and Drug Administration (FDA) approved [abemaciclib in combination with hormone therapy to treat some people who have had surgery for early-stage HR-positive, HER2-negative breast cancer](#).
- [Alpelisib \(Piqray\)](#) is approved to be used in combination with hormone therapy to treat advanced or metastatic HR-positive, HER2-negative breast cancers that have a mutation in the [PIK3CA gene](#).

HER2-Positive Breast Cancer Treatment

The FDA has [approved a number of targeted therapies to treat HER2-positive breast cancer](#), including:

- [Trastuzumab \(Herceptin\)](#) has been approved to be used to prevent a relapse in patients with early-stage HER2-positive breast cancer.
- [Pertuzumab \(Perjeta\)](#) is used to treat metastatic HER2-positive breast cancer, and also used as both [neoadjuvant](#) and [adjuvant therapy](#).
- Trastuzumab and pertuzumab together can be used in combination with chemotherapy to prevent relapse in people with early-stage HER2-positive breast cancer. Both are also used together in metastatic disease, where they delay progression and improve overall survival.
- Trastuzumab deruxtecan (Enhertu) is approved for patients with advanced or metastatic HER2-positive breast cancer who have previously received a HER2-targeted treatment. A 2021 clinical trial showed that the drug [lengthened the time that people with metastatic HER2-positive breast cancer lived](#) without their cancer progressing. The trial also showed that it was better at shrinking tumors than another targeted drug, trastuzumab emtansine (Kadcyla).
- [Tucatinib \(Tukysa\)](#) is approved to be used in combination with trastuzumab and [capecitabine \(Xeloda\)](#) for HER2-positive breast cancer that cannot be removed with surgery or is metastatic. Tucatinib is able to cross the blood–brain barrier, which makes it especially useful for HER2-positive metastatic breast cancer, which tends to spread to the brain.
- [Lapatinib \(Tykerb\)](#) has been approved for treatment of some patients with HER2-positive advanced or metastatic breast cancer, together with capecitabine or letrozole.
- [Neratinib Maleate \(Nerlynx\)](#) can be used in patients with early-stage HER2-positive breast cancer and can also be used together with [capecitabine \(Xeloda\)](#) in some patients with advanced or metastatic disease.
- [Ado-trastuzumab emtansine \(Kadcyla\)](#) is approved to treat patients with metastatic HER2-positive breast cancer who have previously received trastuzumab and a [taxane](#). It's also used in some patients with early-stage HER2-positive breast cancer who have completed neoadjuvant therapy and have [residual disease](#) at the time of surgery.

HER2-Low Breast Cancer

A newly defined subtype, HER2-low, accounts for more than half of all metastatic breast cancers. HER2-low tumors are defined as those whose cells contain lower levels of the HER2 protein on their surface. Such tumors have traditionally been classified as HER2-negative because they did not respond to drugs that target HER2.

However, in a recent clinical trial, trastuzumab deruxtecan (Enhertu) [improved the survival of patients with HER2-low breast cancer compared with chemotherapy](#), and the drug has recently received an accelerated approval for use in such patients.

Breast Cancer Early Detection and Treatment Research

The [Breast Specialized Programs of Research Excellence \(Breast SPOREs\)](#) are designed to quickly move basic scientific findings into clinical settings. The Breast SPOREs support the development of new therapies and technologies, and studies to better understand tumor resistance, diagnosis, prognosis, screening, prevention, and treatment of breast cancer.

The NCI [Cancer Intervention and Surveillance Modeling Network \(CISNET\)](#) focuses on using modeling to improve our understanding of how prevention, early detection, screening, and treatment affect breast cancer outcomes.

The [Confluence Project](#), from NCI's [Division of Cancer Epidemiology and Genetics \(DCEG\)](#), is developing a research resource that includes data from thousands of breast cancer patients and controls of different races and ethnicities. This resource will be used to identify genes that are associated with breast cancer risk, prognosis, subtypes, response to treatment, and second breast cancers. (DCEG conducts [other breast cancer research](#) as well.)

The goal of the [Breast Cancer Surveillance Consortium \(BCSC\)](#)[Exit Disclaimer](#), an NCI-funded program launched in 1994, is to enhance the understanding of breast cancer screening practices in the United States and their impact on the breast cancer's stage at diagnosis, survival rates, and mortality.

There are ongoing programs at NCI that support prevention and early detection research in different cancers, including breast cancer. Examples include:

- The [Cancer Biomarkers Research Group](#), which promotes research in cancer [biomarkers](#) and manages the [Early Detection Research Network \(EDRN\)](#). EDRN is a

network of NCI-funded institutions that are collaborating to discover and validate early detection biomarkers. Within the EDRN, the Breast and Gynecologic Cancers Collaborative Group conducts research on breast and ovarian cancers.

- NCI's [Division of Cancer Prevention](#) houses the [Breast and Gynecologic Cancer Research Group](#) which conducts and fosters the development of research on the prevention and early detection of breast and gynecologic cancers.

Breast Cancer Clinical Trials

NCI funds and oversees both early- and late-phase clinical trials to develop new treatments and improve patient care. Trials are available for breast cancer [prevention](#), [screening](#), and [treatment](#).

OBJECTIVE ANALYSIS

Objective analysis of breast cancer involves using scientific methods to study the disease, its causes, and its effects on the body. Here are some examples of objective analysis techniques used in breast cancer research:

1. **Epidemiological studies:** Epidemiological studies investigate the distribution and determinants of disease in populations. These studies can help identify risk factors associated with breast cancer, such as age, family history, and exposure to certain environmental factors. Epidemiological studies can also help identify patterns and trends in breast cancer incidence and mortality rates.
2. **Imaging techniques:** Imaging techniques such as mammography, ultrasound, and MRI are used to detect and diagnose breast cancer. These techniques can provide detailed images of the breast tissue, allowing doctors to identify abnormalities and determine the stage of the cancer.
3. **Histopathological analysis:** Histopathological analysis involves examining breast tissue samples under a microscope to identify cancerous cells and determine the type and stage of the cancer. This analysis can help guide treatment decisions and provide information on the prognosis of the disease.

4. Molecular analysis: Molecular analysis techniques, such as DNA sequencing and gene expression profiling, can provide information on the genetic and molecular changes associated with breast cancer. This analysis can help identify potential targets for new therapies and improve our understanding of the disease.
5. Clinical trials: Clinical trials are studies that test new treatments or interventions in human subjects. Clinical trials can provide important data on the safety and effectiveness of new treatments for breast cancer, as well as insights into the underlying mechanisms of the disease.
6. Overall, objective analysis of breast cancer involves a multidisciplinary approach that integrates knowledge from various fields, including epidemiology, imaging, pathology, molecular biology, and clinical research. By using objective analysis techniques, researchers can gain a better understanding of the disease and develop new strategies for prevention, diagnosis, and treatment.

PROPOSED SOLUTION AND RESULT ANALYSIS

In this project we will use Data Mining and Machine Learning Algorithms to detect breast cancer, based off data. Breast Cancer (BC) is a common cancer for women around the world.

Early detection of BC can greatly improve prognosis and survival chances by promoting clinical treatment to patients.

We will use the UCI Machine Learning Repository for breast cancer dataset.

<https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset> [3...].

The dataset used in this story is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells

in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector Attribute Information:

1. ID number 2) Diagnosis (M = malignant, B = benign) 3–32)

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

2. radius (mean of distances from center to points on the perimeter)
3. texture (standard deviation of gray-scale values)
4. perimeter
5. area
6. smoothness (local variation in radius lengths)
7. compactness ($\text{perimeter}^2 / \text{area} - 1.0$)
8. concavity (severity of concave portions of the contour)
9. concave points (number of concave portions of the contour)
10. symmetry
11. fractal dimension (“coastline approximation” — 1)

The mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

In this project we use the following machine learning algorithms:

Decision tree algorithms: Decision tree algorithms are successful machine learning classification techniques. They are the supervised learning methods which use information gained and pruned to improve results. Moreover, decision tree algorithms are commonly used for classification in many research, for example, in the medicine area and health issues. There are many kinds of decision tree algorithms such as ID3 and C4.5. However, J48 is the most popular decision tree algorithm. J48 is the implementation of an improved version of C4.5 and is an extension of ID3.

K-nearest-neighbours (kNN) algorithm: It is a simple supervised learning algorithm in pattern recognition. It is one of the most popular neighborhood classifiers due to its simplicity and efficiency in the field of machine learning. KNN algorithm stores all cases and classifies new cases based on similarity measures; it searches the pattern space for the k training tuples that are closest to the unknown tuples. The performance depends on the optimal number of neighbors (k) chosen, which is different from one data sample to another.

Support Vector Machine (SVM): It is a supervised learning method derived from statistical learning theory for the classification of both linear and nonlinear data. SVM classifies data into two classes over a hyperplane at the same time avoiding over-fitting the data by maximizing the margin of hyperplane separating.

Naïve Bayes (NB) It is a probabilistic classifier: It is one of the most efficient classification algorithms based on applying Bayes' theorem with strong (naïve) independent assumptions. It assumes the value of the feature is independent of the value of any other features.

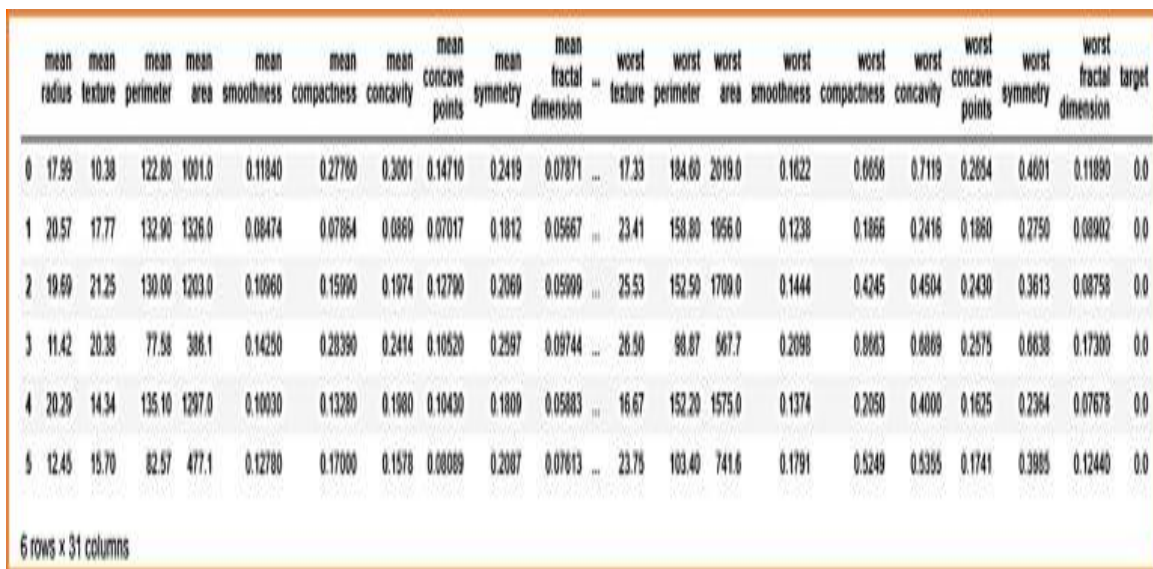
Logistic regression: In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one. Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable (target) is categorical.

SYSTEM DESIGN

We will first go with importing the necessary libraries and import our dataset to colab.research.google.com.

We can examine the data set using the pandas' head() method.

```
df.head(7) {first 7 rows of the data}
```

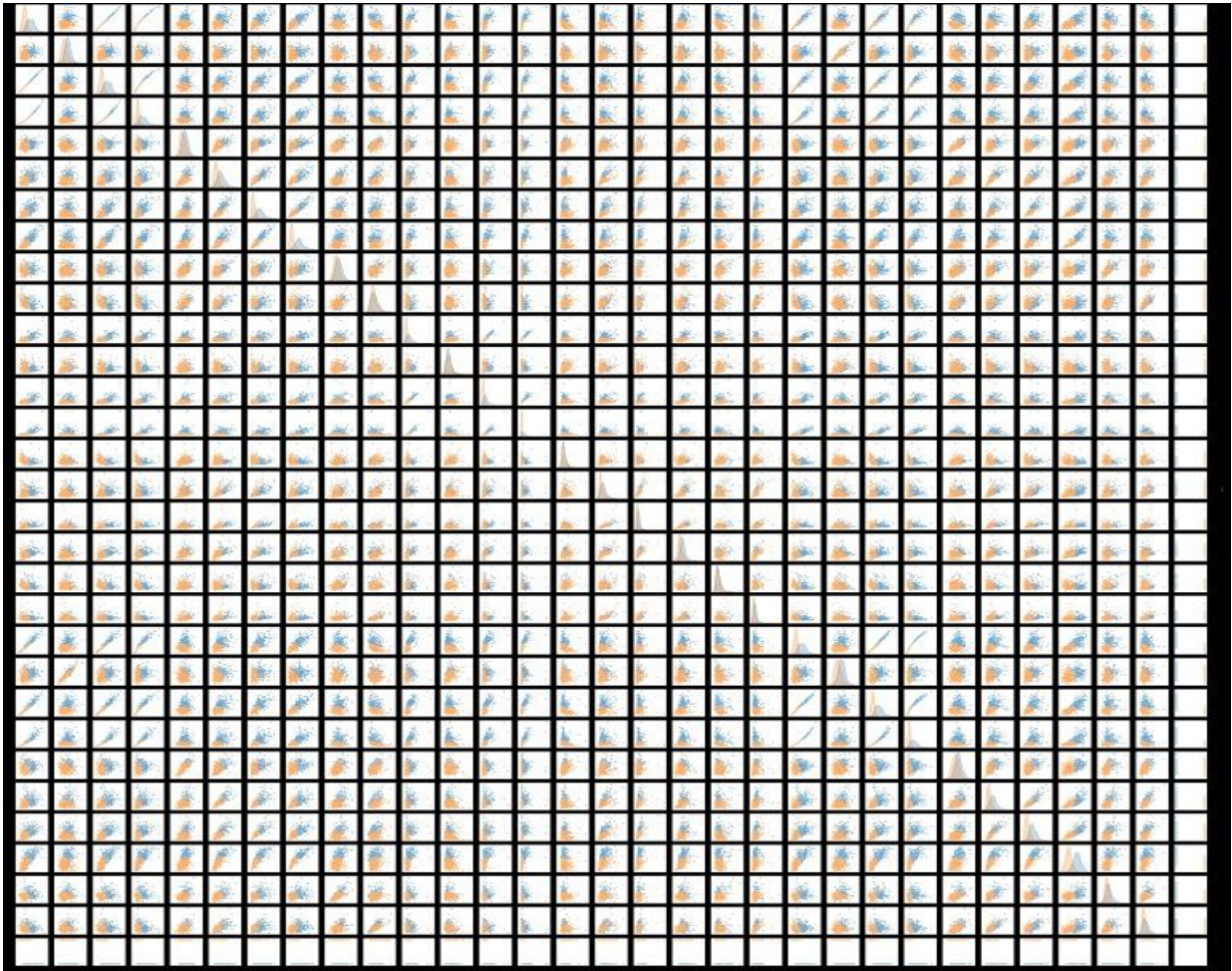


	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	target
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.1622	0.6658	0.7119	0.2654	0.4601	0.11890	0.0
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902	0.0
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05909	...	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758	0.0
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.2098	0.8683	0.6869	0.2575	0.6638	0.17300	0.0
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678	0.0
5	12.45	15.70	82.57	477.1	0.12780	0.17000	0.1578	0.08089	0.2087	0.07813	...	23.75	103.40	741.6	0.1791	0.5249	0.5395	0.1741	0.3985	0.12440	0.0

6 rows x 31 columns

We can observe that the data set contain 569 rows and 33 columns. 'Diagnosis' is the column which we are going to predict , which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labeled as B (benign) and 212 as M (mal)Each row represents a patient and 33 features on the 569 patients.The last column Unnamed: 32 has NaN values so we need to remove that column with empty values. So we count the number of empty columns and drop the columns with empty values.

We can see that id column acts as the identifier of the patient and it is of integer type and it cannot be used as a feature to predict the tumor.Next we encode categorical data values (Transforming categorical data/ Strings to integers).



Pair plot of breast cancer data

Basically, the pair plot is used to show the numeric distribution in the scatter plot.

Paiplot of cancer dataframe

```
sns.pairplot(cancer_df, hue = 'target')
```

Breast Cancer Detection Machine Learning Model Building

We have clean data to build the ML model. But which Machine learning algorithm is best for the data we have to find. The output is a categorical format so we will use supervised classification machine learning algorithms.

To build the best model, we have to train and test the dataset with multiple Machine Learning algorithms then we can find the best ML model.

First, we need to import the required packages.

```
1 from sklearn.metrics import confusion_matrix, classification_report,  
    accuracy_score
```

Support Vector Classifier

```
1 # Support vector classifier  
2 from sklearn.svm import SVC  
3 svc_classifier = SVC()  
4 svc_classifier.fit(X_train, y_train)  
5 y_pred_scv = svc_classifier.predict(X_test)  
6 accuracy_score(y_test, y_pred_scv)
```

Output >>> 0.5789473684210527

```
1 # Train with Standard scaled Data  
2 svc_classifier2 = SVC()  
3 svc_classifier2.fit(X_train_sc, y_train)  
4 y_pred_svc_sc = svc_classifier2.predict(X_test_sc)  
5 accuracy_score(y_test, y_pred_svc_sc)
```

Output >>> 0.9649122807017544

Logistic Regression

```
1 # Logistic Regression  
2 from sklearn.linear_model import LogisticRegression  
3 lr_classifier = LogisticRegression(random_state = 51, penalty = 'l1')  
4 lr_classifier.fit(X_train, y_train)  
5 y_pred_lr = lr_classifier.predict(X_test)  
6 accuracy_score(y_test, y_pred_lr)
```

Output >>> 0.9736842105263158

```

1 # Train with Standard scaled Data
2 lr_classifier2 = LogisticRegression(random_state = 51, penalty = 'l1')
3 lr_classifier2.fit(X_train_sc, y_train)
4 y_pred_lr_sc = lr_classifier.predict(X_test_sc)
5 accuracy_score(y_test, y_pred_lr_sc)

```

Output >>> 0.5526315789473685

K – Nearest Neighbor Classifier

```

1 # K – Nearest Neighbor Classifier
2 from sklearn.neighbors import KNeighborsClassifier
3 knn_classifier = KNeighborsClassifier(n_neighbors = 5, metric =
4 'minkowski', p = 2)
5 knn_classifier.fit(X_train, y_train)
6 y_pred_knn = knn_classifier.predict(X_test)
  accuracy_score(y_test, y_pred_knn)

```

Output >>> 0.9385964912280702

```

1 # Train with Standard scaled Data
2 knn_classifier2 = KNeighborsClassifier(n_neighbors = 5, metric =
3 'minkowski', p = 2)
4 knn_classifier2.fit(X_train_sc, y_train)
5 y_pred_knn_sc = knn_classifier.predict(X_test_sc)
  accuracy_score(y_test, y_pred_knn_sc)

```

Output >>> 0.5789473684210527

Naive Bayes Classifier

```

1 # Naive Bayes Classifier
2 from sklearn.naive_bayes import GaussianNB
3 nb_classifier = GaussianNB()
4 nb_classifier.fit(X_train, y_train)
5 y_pred_nb = nb_classifier.predict(X_test)
6 accuracy_score(y_test, y_pred_nb)

```

Output >>> 0.9473684210526315

```
1 # Train with Standard scaled Data
2 nb_classifier2 = GaussianNB()
3 nb_classifier2.fit(X_train_sc, y_train)
4 y_pred_nb_sc = nb_classifier2.predict(X_test_sc)
5 accuracy_score(y_test, y_pred_nb_sc)
```

Output >>> 0.9385964912280702

Decision Tree Classifier

```
1 # Decision Tree Classifier
2 from sklearn.tree import DecisionTreeClassifier
3 dt_classifier = DecisionTreeClassifier(criterion = 'entropy', random_state
4 = 51)
5 dt_classifier.fit(X_train, y_train)
6 y_pred_dt = dt_classifier.predict(X_test)
accuracy_score(y_test, y_pred_dt)
```

Output >>> 0.9473684210526315

```
1 # Train with Standard scaled Data
2 dt_classifier2 = DecisionTreeClassifier(criterion = 'entropy',
3 random_state = 51)
4 dt_classifier2.fit(X_train_sc, y_train)
5 y_pred_dt_sc = dt_classifier.predict(X_test_sc)
accuracy_score(y_test, y_pred_dt_sc)
```

Output >>> 0.7543859649122807

Random Forest Classifier

```
1 # Random Forest Classifier
2 from sklearn.ensemble import RandomForestClassifier
3 rf_classifier = RandomForestClassifier(n_estimators = 20, criterion =
4 'entropy', random_state = 51)
5 rf_classifier.fit(X_train, y_train)
6 y_pred_rf = rf_classifier.predict(X_test)
accuracy_score(y_test, y_pred_rf)
```

Output >>> 0.9736842105263158

```
1 # Train with Standard scaled Data
2 rf_classifier2 = RandomForestClassifier(n_estimators = 20, criterion =
3 'entropy', random_state = 51)
4 rf_classifier2.fit(X_train_sc, y_train)
5 y_pred_rf_sc = rf_classifier.predict(X_test_sc)
  accuracy_score(y_test, y_pred_rf_sc)
```

Output >>> 0.7543859649122807

Adaboost Classifier

```
1 # Adaboost Classifier
2 from sklearn.ensemble import AdaBoostClassifier
3 adb_classifier = AdaBoostClassifier(DecisionTreeClassifier(criterion =
4 'entropy', random_state = 200),
5                               n_estimators=2000,
6                               learning_rate=0.1,
7                               algorithm='SAMME.R',
8                               random_state=1,)
9 adb_classifier.fit(X_train, y_train)
1 y_pred_adb = adb_classifier.predict(X_test)
0 accuracy_score(y_test, y_pred_adb)
```

Output >>> 0.9473684210526315

```
1 # Train with Standard scaled Data
2 adb_classifier2 = AdaBoostClassifier(DecisionTreeClassifier(criterion =
3 'entropy', random_state = 200),
4                               n_estimators=2000,
5                               learning_rate=0.1,
6                               algorithm='SAMME.R',
7                               random_state=1,)
8 adb_classifier2.fit(X_train_sc, y_train)
9 y_pred_adb_sc = adb_classifier2.predict(X_test_sc)
```



```
accuracy_score(y_test, y_pred_adb_sc)
```

```
Output >>> 0.9473684210526315
```

XGBoost Classifier

```
1 # XGBoost Classifier
2 from xgboost import XGBClassifier
3 xgb_classifier = XGBClassifier()
4 xgb_classifier.fit(X_train, y_train)
5 y_pred_xgb = xgb_classifier.predict(X_test)
6 accuracy_score(y_test, y_pred_xgb)
```

```
Output >>> 0.9824561403508771
```

```
1 # Train with Standard scaled Data
2 xgb_classifier2 = XGBClassifier()
3 xgb_classifier2.fit(X_train_sc, y_train)
4 y_pred_xgb_sc = xgb_classifier2.predict(X_test_sc)
5 accuracy_score(y_test, y_pred_xgb_sc)
```

```
Output >>> 0.9824561403508771
```

XGBoost Parameter Tuning Randomized Search

```
1 # XGBoost classifier most required parameters
2 params={
3 "learning_rate" : [0.05, 0.10, 0.15, 0.20, 0.25, 0.30 ] ,
4 "max_depth" : [ 3, 4, 5, 6, 8, 10, 12, 15],
5 "min_child_weight" : [ 1, 3, 5, 7 ],
6 "gamma" : [ 0.0, 0.1, 0.2 , 0.3, 0.4 ],
7 "colsample_bytree" : [ 0.3, 0.4, 0.5 , 0.7 ]
8 }
1 # Randomized Search
2 from sklearn.model_selection import RandomizedSearchCV
3 random_search = RandomizedSearchCV(xgb_classifier,
4 param_distributions=params, scoring= 'roc_auc', n_jobs= -1, verbose= 3)
random_search.fit(X_train, y_train)
```

Output >>>

```
1 RandomizedSearchCV(cv='warn', error_score='raise-deprecating',
2     estimator=XGBClassifier(base_score=0.5, booster='gbtree',
3     colsample_bylevel=1,
4     colsample_bynode=1, colsample_bytree=1, gamma=0,
5     learning_rate=0.1,
6     max_delta_step=0, max_depth=3, min_child_weight=1,
7     missing=None,
8     n_estimators=100, n_jobs=1, nthread=None,
9     objective='binary:logistic', random_state=0, reg_alpha=0,
1    reg_lambda=1, scale_pos_weight=1, seed=None, silent=None,
0    subsample=1, verbosity=1),
1    fit_params=None, iid='warn', n_iter=10, n_jobs=-1,
1    param_distributions={'learning_rate': [0.05, 0.1, 0.15, 0.2, 0.25,
1 0.3], 'max_depth': [3, 4, 5, 6, 8, 10, 12, 15], 'min_child_weight': [1, 3, 5,
2 7], 'gamma': [0.0, 0.1, 0.2, 0.3, 0.4], 'colsample_bytree': [0.3, 0.4, 0.5,
    0.7]}},
    pre_dispatch='2*n_jobs', random_state=None, refit=True,
    return_train_score='warn', scoring='roc_auc', verbose=3)
1 random_search.best_params_
```

Output >>>

```
1 {'min_child_weight': 1,
2  'max_depth': 3,
3  'learning_rate': 0.3,
4  'gamma': 0.4,
5  'colsample_bytree': 0.3}
1 random_search.best_estimator_
```

Output >>>

```
1 XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
```

```

2     colsample_bynode=1, colsample_bytree=0.3, gamma=0.4,
3     learning_rate=0.3, max_delta_step=0, max_depth=3,
4     min_child_weight=1, missing=None, n_estimators=100, n_jobs=1,
5     nthread=None, objective='binary:logistic', random_state=0,
6     reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
7     silent=None, subsample=1, verbosity=1)
1 # training XGBoost classifier with best parameters
2 xgb_classifier_pt = XGBClassifier(base_score=0.5, booster='gbtree',
3     colsample_bylevel=1,
4     colsample_bynode=1, colsample_bytree=0.4, gamma=0.2,
5     learning_rate=0.1, max_delta_step=0, max_depth=15,
6     min_child_weight=1, missing=None, n_estimators=100, n_jobs=1,
7     nthread=None, objective='binary:logistic', random_state=0,
8     reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
9     silent=None, subsample=1, verbosity=1)
1
0 xgb_classifier_pt.fit(X_train, y_train)
1 y_pred_xgb_pt = xgb_classifier_pt.predict(X_test)
1
1 accuracy_score(y_test, y_pred_xgb_pt)

```

Output >>> 0.9824561403508771

Confusion Matrix

```

1 cm = confusion_matrix(y_test, y_pred_xgb_pt)
2 plt.title('Heatmap of Confusion Matrix', fontsize = 15)
3 sns.heatmap(cm, annot = True)
4 plt.show()

```



The model is giving 0% type II error and it is best.

Classification Report of Model

```
1 print(classification_report(y_test, y_pred_xgb_pt))
```

Output >>>

```
1      precision  recall f1-score  support
2
3      0.0      1.00   0.96   0.98     48
4      1.0      0.97   1.00   0.99     66
5
6  micro avg    0.98   0.98   0.98    114
7  macro avg    0.99   0.98   0.98    114
8  weighted avg  0.98   0.98   0.98    114
```

Cross-validation of the ML model

To find the ML model is overfitted, under fitted or generalize doing cross-validation.

```
1 # Cross validation
```

```
2 from sklearn.model_selection import cross_val_score
```

```
3 cross_validation = cross_val_score(estimator = xgb_model_pt2, X =
```

```

4 X_train_sc, y = y_train, cv = 10)
5 print("Cross validation of XGBoost model = ",cross_validation)
6 print("Cross validation of XGBoost model (in mean) =
7 ",cross_validation.mean())
8 from sklearn.model_selection import cross_val_score
9 cross_validation = cross_val_score(estimator = xgb_classifier_pt, X =
  X_train_sc,y = y_train, cv = 10)
  print("Cross validation accuracy of XGBoost model = ", cross_validation)
  print("\nCross validation mean accuracy of XGBoost model = ",
  cross_validation.mean())

```

Output >>>

```

Cross validation accuracy of XGBoost model = [0.9787234 0.97826087 0.97826087
2 0.97826087 0.93333333 0.91111111
3 1.      1.      0.97777778 0.88888889]

```

4

Cross validation mean accuracy of XGBoost model = 0.9624617124062083

The mean accuracy value of **cross-validation is 96.24%** and **XGBoost model accuracy is 98.24%**. It showing XGBoost is slightly overfitted but when training data will more it will generalized model.

CHALLENGES

Breast cancer detection faces several challenges, including:

5. Dense breast tissue: Dense breast tissue can make it more difficult to detect breast cancer on mammography images. This is because dense tissue appears white on mammograms, which can obscure small lesions or tumors.
6. False positives and false negatives: Mammography and other breast cancer detection tests can produce false positives (detecting abnormalities that are not cancer) and false negatives (failing to detect cancerous lesions). False positives can lead to unnecessary biopsies and anxiety for patients, while false negatives can delay diagnosis and treatment.

7. Interpretation variability: Interpreting mammography images and other breast cancer tests can be subjective, and different radiologists or pathologists may have different opinions on the presence or absence of cancerous lesions.
8. Limited access to screening: Limited access to breast cancer screening, particularly in low-income and rural areas, can lead to delayed diagnosis and poorer outcomes for patients.
9. Data quality and availability: Developing accurate and effective breast cancer detection algorithms requires access to high-quality data, which can be limited in certain settings. Data quality can also be impacted by issues such as data bias or incomplete data.
10. Cost: Breast cancer detection tests and procedures can be expensive, particularly for patients without adequate health insurance coverage.
11. Ethical concerns: The use of artificial intelligence and machine learning algorithms in breast cancer detection raises ethical concerns, such as issues related to patient privacy and the potential for algorithmic bias.

Addressing these challenges requires a multi-faceted approach that involves improving screening technologies, ensuring equitable access to breast cancer screening, developing more accurate and consistent interpretation methods, improving data quality and availability, and addressing ethical concerns associated with the use of AI and machine learning algorithms in breast cancer detection.

SUGGESTION

There are several suggestions for improving breast cancer detection:

12. Improved screening technologies: Developing more sensitive and specific screening technologies, such as digital breast tomosynthesis (DBT) and molecular breast imaging (MBI), can improve early detection and reduce false positives and false negatives.
13. Personalized screening: Tailoring breast cancer screening recommendations based on individual risk factors, such as age, family history, and breast density, can improve detection rates and reduce unnecessary testing.
14. Standardized interpretation: Developing standardized interpretation guidelines for breast cancer screening tests can reduce variability in interpretation and improve accuracy.

15. Access to screening: Ensuring equitable access to breast cancer screening, particularly in low-income and rural areas, can improve detection rates and reduce disparities in breast cancer outcomes.
16. Advances in artificial intelligence: Developing and validating machine learning algorithms that can accurately detect breast cancer from mammography and other imaging data can improve detection rates and reduce the need for unnecessary biopsies.
17. Education and awareness: Increasing education and awareness about breast cancer and the importance of screening can encourage more women to participate in screening programs and seek medical attention if they notice any breast abnormalities.
18. Research and development: Continued research and development in breast cancer detection, including molecular and genetic biomarkers and new imaging technologies, can improve our understanding of the disease and lead to new and more effective screening methods.

CONCLUSION

In conclusion, breast cancer detection is a critical component of breast cancer management, as early detection and diagnosis can improve outcomes and reduce mortality rates. However, breast cancer detection faces several challenges, including dense breast tissue, false positives and false negatives, interpretation variability, limited access to screening, data quality and availability, cost, and ethical concerns. To improve breast cancer detection, there are several suggestions, including the development of improved screening technologies, personalized screening recommendations, standardized interpretation guidelines, equitable access to screening, advances in artificial intelligence, education and awareness, and continued research and development. By addressing these challenges and implementing these suggestions, we can improve breast cancer detection rates, reduce disparities in breast cancer outcomes, and ultimately improve the lives of individuals affected by breast cancer.

REFERENCES

- [1.] <https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>.

[2.] <https://www.coursera.org/articles/machine-learn>

[3.] Original dataset

<https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset>

[4.] <https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3>

[5.] Confusion Matrix:

[https://tatwan.github.io/How-To-Plot-A-Confusion-Matrix- In-Python/](https://tatwan.github.io/How-To-Plot-A-Confusion-Matrix-In-Python/)

[6.] https://seaborn.pydata.org/tutorial/axis_grids.html

[7.] <https://seaborn.pydata.org/generated/seaborn.pairplot.html>

[8.] <https://seaborn.pydata.org/generated/seaborn.heatmap.html>

